Investigating the Effect of Coarse-Graining on Chemical Compound Space

Dissertation

zur Erlangung des Grades
„Doktor rerum naturalium (Dr. rer. nat.)“
der Fachbereiche:
08 - Physik, Mathematik und Informatik
09 - Chemie, Pharmazie, Geographie und Geowissenschaften
10 - Biologie, Universitätsmedizin
der Johannes
Gutenberg-Universität
in Mainz

Kiran H. Kanekal

Datum der mündlichen Prüfung:
15.12.2020

Contents

List of Figures

List of Tables

Included Publications

While this is not a cumulative dissertation (thesis by publication), several published works are included in this thesis. These results were all obtained during my time at the Max Planck Institute for Polymer Research in Mainz since I began working there in July 2016, and since joining the Max Planck Graduate Center in July 2017. "I hereby declare that I wrote the dissertation submitted without any unauthorized external assistance and used only sources acknowledged in the work. All textual passages which are appropriated verbatim or paraphrased from published and unpublished texts as well as all information obtained from oral sources are duly indicated and listed in accordance with bibliographical rules. In carrying out this research, I complied with the rules of standard scientific practice as formulated in the statutes of Johannes Gutenberg University Mainz to insure standard scientific practice. Significant portions of three publications are included in the first half of Chapter 2, in which I was not the first author. Chapter 3 is made up of a single publication, in which I was the first author. Chapter 4 is currently in preparation to be submitted as an article in which I will be the first author. The remaining work is original to this thesis. The contributions of each author are detailed in the *Contributions* section of this dissertation.

Chapter 2:
Roberto Menichetti, Kiran H. Kanekal, Kurt Kremer, and Tristan Bereau
In silico screening of drug-membrane thermodynamics reveals linear relations between bulk partitioning and the potential of mean force
The Journal of Chemical Physics 147(12):125101, 2017.
DOI: 10.1063/1.4987012

Roberto Menichetti, Kiran H. Kanekal, and Tristan Bereau
Drug–membrane permeability across chemical space
ACS Central Science 5(2):290, 2019.
DOI: 10.1021/acscentsci.8b00718

Christian Hoffmann, Roberto Menichetti, Kiran H. Kanekal, and Tristan Bereau
Controlled exploration of chemical space by machine learning of coarse-grained representations
Physical Review E 100(3):033302, 2019.
DOI: 10.1103/PhysRevE.100.033302

Chapter 3:
Kiran H. Kanekal and Tristan Bereau
Resolution limit of data-driven coarse-grained models spanning chemical space
The Journal of Chemical Physics 151:164106, 2019.
DOI: 10.1063/1.5119101

Abstract

Chemical structure-property relationships are essential for the development of new materials used in all facets of life. Practically, this process amounts to projecting regions of the chemical compound space (CCS) onto certain descriptors related to the property of interest, allowing the structure-property relationship to be inferred. The challenge in constructing these relationships usually stems from a lack of data, as their accuracy and transferability will depend on how well-sampled CCS is with respect to the chosen descriptors. High-throughput screening, in which the properties of compounds are determined in an automated fashion, is one strategy used to overcome this problem. However, for many properties of soft-matter systems, this approach is difficult to implement computationally. This difficulty arises due to the large costs associated with adequately sampling complex free energy landscapes at atomistic resolutions using established tools such as molecular dynamics (MD) simulations. Coarse-grained (CG) models, parameterized at lower resolutions compared to their atomistic counterparts, provide a means to circumvent these costs. However, many of these models are constructed in order to specifically reproduce the properties of a small number of compounds, making it difficult to generalize across CCS. In this work, we demonstrate that the coarse-grained Martini model reduces the size of CCS, and can be used in computational high-throughput screening methods to efficiently construct chemical structure-property relationships over wide ranges of CCS. We find that this reduction of CCS is due to a limited number of Martini interaction types, with multiple atomistic chemical fragments mapping to the same CG interaction type. We then investigate the relationship between unsupervised machine learning and coarse-graining, yielding strategies for parameterizing chemically transferable CG models from both a top-down and bottom-up perspective. We employ these data-driven techniques to parameterize new top-down CG models and quantify their transferability and accuracy as a function of the number of CG interaction types for each model. Finally, we develop a method that uses unsupervised machine learning in combination with the bottom-up multiscale coarse-graining technique to generate chemically-transferable CG models with high structural accuracy. We examine the limitations of both top-down and bottom-up approaches and make recommendations for the future development of these methodologies. Overall, our work demonstrates the means by which chemically-transferable CG models can be both constructed and utilized to efficiently infer chemical structure-property relationships for materials discovery.

Zusammenfassung

Das Verständnis der Beziehung zwischen Struktur und Eigenschaft chemischer Verbindungen ist essenziell um neue Materialien zu entwickeln, die in allen Facetten unseres Lebens benutzt werden. In der Praxis wird dies erreicht, indem Regionen des Raums der chemischen Verbindungen (CCS) auf bestimmte Deskriptoren projiziert werden, die mit den gewünschten Eigenschaften zusammenhängen und auf die Struktur-Eigenschaft Beziehung folgern lassen. Die Herausforderung beim Konstruieren dieser Beziehungen entsteht häufig durch ein Mangel an Daten, da die Korrektheit und Generalisierbarkeit davon abhängig ist, wie gut der Raum der chemischen Verbindungen abgebildet wurde bezüglich der gewählten Deskriptoren. Hochdurchsatz-Screening, bei dem die Eigenschaften der Verbindungen automatisiert bestimmt werden, ist eine Strategie, um dieses Problem zu lösen. Leider ist dieser Ansatz für viele Eigenschaften von weicher Materie schwierig zu implementieren aufgrund des hohen Rechenaufwands, der betrieben werden muss, um die Freie-Energie-Landschaft mithilfe herkömmlicher Methoden, wie Moleculardynamik-Simulationen (MD), adäquat mit atomistischer Auflösung abzutasten. Coarse-grained (CG) Modelle, die auf geringer Auflösung als ihre atomistischen Gegenstücke parametrisiert wurden, sind nützlich, um diese hohen Kosten zu verringern. Leider sind viele dieser Modelle konstruiert, um spezielle Eigenschaften von einer kleinen Anzahl an Verbindungen zu reproduzieren, was es schwierig macht diese im CCS zu Generalisieren. In dieser Arbeit demonstrieren wir, dass das coarse-grained Martini-Modell die Größe des CCS reduziert und dass es benutzt werden kann, um hochdurchsatz Methoden durchzuführen, damit chemische Struktur-Eigenschaft Beziehungen für große Bereiche des CCS effizient konstruiert werden können. Wir zeigen, dass der CCS durch eine beschränkte Anzahl an Martini-Wechselwirkungstypen reduziert wird, wobei mehrere atomistische Fragmente auf denselben CG-Wechselwirkungstyp projiziert werden. Des Weiteren untersuchen wir die Beziehung zwischen unüberwachtem Maschinellem Lernen (ML) und CG, was zu Strategien führt, um Parametrisierungen für chemisch generalisierbare CG-Modelle für top-down sowie auch bottom-up Methoden zu finden. Wir wenden diese datengesteuerte Technik an, um neue top-down CG Modelle zu parametrisieren und quantifizieren ihre Generalisierbarkeit und Korrektheit für jedes Modell als eine Funktion der Anzahl an CG Wechselwirkungstypen. Zuletzt entwickeln wir eine Methode, die unüberwachtes Maschinelles Lernen mit bottom-up Multiskalen-CG kombiniert, um chemisch generalisierbare CG-Modelle mit hoher Korrektheit zu generieren. Wir untersuchen die Grenzen von top-down sowie bottom-up Ansätzen und machen Empfehlungen für die weitere Entwicklung dieser

Methoden. Zusammengefasst demonstriert unsere Arbeit unter welchen Bedingungen chemisch generalisierbare CG Modelle konstruiert und verwendet werden können, um chemische Struktur-Eigenschaft Beziehungen für Materialien effizient zu untersuchen.

1 Introduction and Theory

The design and production of novel materials has been an intrinsic aspect of the human experience for many millennia. While ancient humans may not have realized it, they were discovering relationships between different chemical compounds and the properties of these compounds that interested them. Some examples include tuning the mechanical properties of bronze, which would depend on the specific composition of the copper alloy, or the use of different plant-based substances in the creation of dyes with varying colors. The modern scientific method, coupled with a fundamental understanding of chemistry and physics, has led to explicit mappings of specific chemical compounds to desired properties of interest. Thus, the creation of these so-called "structure–property relationships" has continued into the present day, leading to the development of new materials that aid humanity in all facets of life.

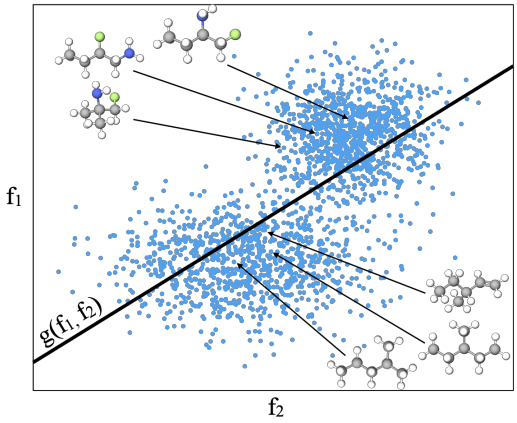

Figure 1.1: Schematic of a typical structure–property relationship. Each point denotes a molecule projected onto descriptors f_1 and f_2. In this cartoon example, data is roughly separated into two clusters, one corresponding to small branched hydrocarbons with F and N substitutions and another cluster without any heteroatom substitutions. The data is fit to a line, relating the projected chemical structures to the property of interest, g.

A good structure–property relationship not only provides a holistic and intuitive sense for the physical phenomena that give rise to properties of interest, but also enables quantitative predictions for new compounds. A cartoon example of a structure–property relationship is shown in Fig. 1.1. Here, the points on the plot represent a subset of the space of all stable chemical compounds, known as chemical compound space (CCS), projected onto two descriptors f_1 and f_2 [1–3]. In this example, the structure–property relationship is obtained by fitting the data to some physically-informed function, g, which corresponds to the property of interest.

How large is CCS? For small, drug-like molecules the number of stable compounds was estimated to be about 10^{60} [4]. For comparison, the estimated number of stars in the entire universe is only about 10^{11} (lending some credence to chemical space, rather than outer space, being the *true* final frontier). Despite this daunting size, one of the main goals of materials science is to obtain structure–property relationships that span CCS, enabling the design of novel, high-performing materials. Ideally, these structure–property relationships can be used for both direct and inverse molecular design [5, 6]. Direct molecular design allows for the prediction of a property value given a new chemical structure. Inverse molecular design, however, yields a chemical structure or set of structures given a desired property value. Regardless of the approach, accurate predictions can only be obtained if CCS is sufficiently sampled. The overarching, long-term goal of this work is to investigate methods for quickly sampling broad regions of CCS in order to facilitate the construction of structure–property relationships for materials design [7–10].

One commonly used strategy used to explore CCS is high-throughput screening. High-throughput screening is a method in which a large number of chemical compounds are systematically tested in an automated fashion to obtain their chemical properties of interest. Experimentally, there have been notable successes in using this approach for direct molecular design. Zhang et al. has developed a high-throughput cell imaging-based screening assay which has led to the discovery of new chemotherapeutic agents [11]. Shevlin et al. has demonstrated a high-throughput method for discovering new and efficient catalysts for chemical synthesis of chiral drug-like compounds [12]. Wambach et al. developed a high throughput approach for obtaining the thermoelectric properties of the entire Ti-Ni-Sn ternary system, in which several thin films could be characterized at once [13]. However, the experimental approaches require a large monetary and time investment for the synthesis of new compounds that are not commercially available.

On the other hand, the rapid growth of computer technology over the last thirty years has made running high-throughput computational simulations a feasible alternative. While this approach also requires a large monetary and time investment in terms of CPU cores and CPU hours, unlike the experimental high-throughput

methods, it has no restrictions stemming from compound synthesis, as even highly unstable compounds can be probed, for example transition states in chemical reactions [14]. Furthermore, this approach benefits from the recent development of robust, numerical approaches that can infer underlying relationships given sufficient data, known as data-driven methods, or machine learning (ML) [15]. ML has been used by several groups to infer structure–property relationships for the electronic properties of small organic compounds [16–18]. Recently, Faber et al. have developed a new ML molecular representation that enables the prediction of many electronic ground state properties of organic molecules with accuracy comparable to ab initio methods [19]. He et al. have used a high throughput approach in combination with ML to discover new, conductive metal-organic complexes [20]. Körbel et al. has used high-throughput ab initio calculations to find and characterize new inorganic perovskites for numerous electronic applications [21]. The high-throughput step for these schemes requires ab initio calculations, in which the electronic probability distribution is obtained by numerically solving Schrödinger's equation, to be run in vacuo for each compound screened. Relatively few high-throughput computational methods have been proposed that build structure–property relationships for thermodynamic properties in the condensed phase, for which thermal fluctuations play an important role [22, 23]. For these methods, the corresponding computational method used for screening is usually classical molecular dynamics (MD) simulations. This simulation approach approximates the quantum-mechanical interactions between atoms as classical force-fields, enabling the use of Newton's equations of motion to evolve condensed-phase systems in time. Yang et al. used a combination of high-throughput MD simulations and ML to predict the stiffness of silicate glasses [24]. Xu et al. used a combination of high-throughput docking calculations and MD to discover new drug molecules that target human bromodomains, which are critical to many different cellular processes [25]. The main bottleneck to implementing high-throughput screening that incorporates MD simulations is the time required to produce trajectories from which thermodynamic properties can be obtained. Specifically, the equations of motion must be applied to large systems usually containing thousands of particles for hundreds of nanoseconds, corresponding to approximately 10^5 CPU hours [26, 27]. Consequently, many studies that rely on MD simulations sample an extremely narrow region of CCS, usually containing O(10) compounds [28–30].

One method for overcoming the computational bottlenecks caused by large system sizes and long sampling times required when running MD simulations is the use of coarse-grained (CG) models [31–34]. A CG model represents chemical compounds as particles in a similar fashion to all-atomistic (AA) MD simulations. However, each particle in a CG model represents groups of atoms rather than a single atom. The model can be constructed by projecting information from

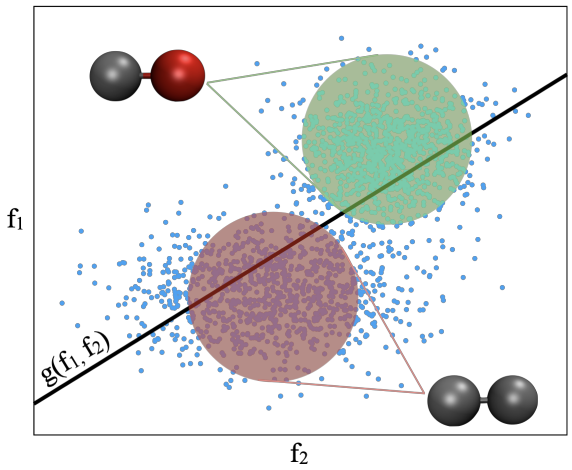

Figure 1.2: The effect of coarse-graining on structure–property relationships. The majority of the compounds map to only two coarse-grained representations, as shown by the red and green circles. This demonstrates the reduction of CCS and enables a broader variety of compounds to be sampled.

a high-resolution simulation (e.g. AA), by inferring microscopic behavior using macroscopic experimental results, or some combination of both. In all cases, the goal of the CG model is to reproduce certain properties of interest by projecting the information pertaining to the property onto a minimal set of parameters. This approach has interesting parallels with the construction of chemical structure–property relationships, as both involve relating chemical structure to a desired property using a reduced model. Furthermore, CG MD simulations require fewer particles compared to AA (usually by some multiplicative factor between two and ten) and do not require as much sampling time due to the removal of degrees of freedom that are irrelevant to the studied property due to a scale separation. For example, the bond vibrations of the hydrogen atoms in a methyl group may be fully decorrelated with the secondary structure formation of the protein to which it belongs. These two effects can reduce the number of CPU hours needed by orders of magnitude [31, 35]. Therefore, CG modeling may provide a means to accelerate the computational high-throughput screening process for condensed phase thermodynamic properties.

The difficulty in utilizing CG models for high-throughput screening stems from the fact that many CG models are chemically specific, meaning that they are constructed for a single chemical compound or small set of chemical compounds, usually at a single thermodynamic state point [31, 36–39]. Because they require an

AA MD simulation or other, equally expensive experimental data for their parameterization, the transferability of these models is usually limited to the chemistry used in their construction. This means that for each compound, the high-resolution data would have to first be obtained, making the actual construction of the CG model unnecessary for a high-throughput approach. Several instances of extending the transferability of CG models have been demonstrated, but these are applied to the state point variables, allowing for CG models to be run at various temperatures, pressures, and concentrations given the same set of chemical compounds [38, 40, 41]. On the other hand, relatively little work has been done that investigates the chemical transferability of CG models [42, 43]. A chemically-transferable CG model would be highly beneficial in a high-throughput screening process because a single CG molecule would be representative of many different chemical compounds. In the context of structure–property relationships, this would correspond to Fig. 1.2, in which two CG molecules represent the majority of the individual chemical compounds shown. Fig. 1.2 indicates that chemically-transferable CG models essentially reduce the size of the CCS, enabling the construction of structure–property relationships that are more robust to chemical variety.

The central theme of this work is to investigate the different ways in which CG modelling can be used to augment computational high-throughput screening methods for condensed phase thermodynamic properties. The rest of the thesis is organized as follows. The remainder of this introductory chapter provides a theoretical foundation for all of the methods used in the subsequent chapters, written so as to be comprehensible to anyone with at least an undergraduate degree in Chemistry, Physics, or Chemical Engineering. In Chapter 2, we demonstrate the chemical transferability of a specific top-down CG model called Martini and use it to construct low-dimensional structure property relationships pertaining to the permeability of lipid-bilayer membranes. In particular, we focus on how inverse molecular design is enabled via this approach, and explore the regions of CCS that map to single CG representations using unsupervised ML techniques. We also explore the relationship between unsupervised machine learning and coarse-graining by examining how different molecular representations and unsupervised learning techniques coarsen CCS in different ways. In Chapter 3, we quantify the chemical transferability of Martini and compare it to three other CG models in the Martini framework with varying number of bead types. This provides a method for optimizing chemically-transferable top-down CG models so that they more efficiently represent CCS. In Chapter 4, we use a combination of unsupervised machine learning techniques and extended-ensemble coarse-graining methods to construct a new bottom-up CG model that is both chemically-transferable and structurally accurate. Finally, we conclude with Chapter 5, emphasizing the importance of the work as well as highlighting many new questions to be answered in the future.

1.1 Theory

1.1.1 Statistical Mechanics and Thermodynamics

An important goal of thermodynamics is to explain, using statistical mechanics, how certain properties of matter are derived from the collective behavior of that matter on the single particle level. While statistical mechanics is useful for describing both quantum and classical phenomena, here we will focus on the latter. Central to the thermodynamic view of statistical mechanics is the idea of microstates and microstate ensembles. In this context, a microstate is defined as a configuration of particles with all of their degrees of freedom specified [44]. An ensemble of microstates refers to the collection of microstates that contribute to a particular thermodynamic system at equilibrium, as well as the corresponding weights attached to each microstate. In the following subsections, we further explain some of the results of statistical mechanics which are highly relevant for understanding the work as a whole. For a more thorough understanding of many of the concepts outlined here, the interested reader is recommended the textbook by M. Scott Shell [44].

Partition Free Energies

For a closed system at constant temperature and volume, known as a canonical ensemble, the Helmholtz free energy, A, can be defined in terms of potential energy, E, temperature, T, and entropy, S, via the following relation:

$$A = E - TS. \tag{1.1}$$

This essentially states that the free energy determines whether the system is driven by the internal energy or entropic fluctuations at a given temperature [44]. The corresponding free energy for a closed system at constant temperature and pressure is known as the Gibbs free energy, G, and also takes into account a pressure-volume potential-energy term. In the canonical ensemble, the system is at thermodynamic equilibrium when the Helmholtz free energy of the system is minimized [44]. From a microscopic view the probability P of a microstate with energy E existing in this ensemble is

$$P = \frac{1}{Z} e^{(-\beta E)}, \tag{1.2}$$

where β is the inverse temperature scaled by Boltzmann's constant, $\beta = 1/(k_{\mathrm{B}}T)$, and Z is the canonical partition function, defined as

$$Z = e^{(-\beta A)}. \tag{1.3}$$

The potential energy is, by definition, a relative energy, and the free energy by itself has little meaning. Consequently, the free energy is expressed in terms of

free energy differences between two thermodynamic state points. A negative free energy difference indicates that the system will be driven towards the second state point over the first in order to reach equilibrium, whereas a positive value denotes the opposite case. These thermodynamic state points can be defined for a large variety of different systems. Some common examples include the solvation free energy, which is the difference in free energy of a molecule in vacuum versus the same molecule in a solvent, binding free energy, for which the state points are the molecule close to/far from a surface, or alchemical free energy differences, in which atoms of the molecule are changed into other atom types while the system remains fixed [45, 46]. A free energy change going from state 1 to 2 in the canonical ensemble at temperature T can be determined using the following expression

$$\Delta A = A_2 - A_1 = -\frac{1}{\beta} \ln \left(\frac{\int \mathrm{d}\boldsymbol{r}^N e^{-\beta U_2(\boldsymbol{r}^N)}}{\int \mathrm{d}\boldsymbol{r}^N e^{-\beta U_1(\boldsymbol{r}^N)}} \right). \tag{1.4}$$

In this equation, U_2 and U_1 are the potential energies of the system at states 2 and 1, respectively [45, 46]. These energies are functions of the positions of all N particles in the system, $\boldsymbol{r}^N$, and the integrals in the equation are carried out over the entire configurational space of the system. Note that the free energy is a path-independent quantity (also known as a state function), meaning that the free energy difference between two state points is the same regardless of the path taken. This path-independence allows for the creation of thermodynamic cycles, which are a series of state points in which thermodynamic variables are changed with the first state point being equal to the last. This means that the net change in free energy over the entire cycle is zero.

An important type of free energy difference is the water/octanol partition free energy, $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$, which gives the free energy change when changing the solvent surrounding a solvated molecule from water to octanol. It is defined using the Gibbs free energy as follows:

$$\Delta G_{\mathrm{W}\to\mathrm{Ol}} = G_{\mathrm{W}} - G_{\mathrm{Ol}}, \tag{1.5}$$

where G_{W} and G_{Ol} are the free energies of the water-solvated and octanol-solvated systems, respectively [47]. Since water is a highly polar solvent whereas octanol is relatively apolar, the water/octanol partition free energy is highly correlated with the polarity of the solute molecule. Many factors can influence the polarity of a molecule, including its size, flexibility, and the existence of specific chemical functional groups within the molecule [48]. A hydrophobic molecule will have a clear preference for octanol over water, resulting in a negative $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ value, whereas the opposite will be true for a hydrophilic molecule. The hydrophobic/hydrophilic character, and by extension the water/octanol partition free energy, play an im-

portant role in the physics of soft-matter systems as a major driving force in phenomena such as self-assembly, protein-ligand binding, and membrane permeability [49–51].

The Potential of Mean Force

It is often useful to determine how the free energy of the system varies as a function of a specific variable or reaction coordinate of interest. Computing the free energy along one of these reaction coordinates yields a potential of mean force (PMF). The PMF, $F(\xi)$, is formally defined by partially integrating the configurational partition function, Z, expressed in terms of the reaction coordinate, ξ [45].

$$F(\xi) = -\frac{1}{\beta} \ln Z(N, V, T, \xi) = -\frac{1}{\beta} \ln \int \mathrm{d}\boldsymbol{r}^N e^{-\beta U(\boldsymbol{r}^N)} \delta[\xi - \hat{\xi}(\boldsymbol{r}^N)] \tag{1.6}$$

$\hat{\xi}$ is a function that outputs the value of the reaction coordinate as a function of the system configuration, $\boldsymbol{r}^N$. The name "potential of mean force" stems from the fact that taking the derivative of the potential with respect to the reaction coordinate over the entire ensemble yields the average force projected onto that reaction coordinate.

1.1.2 Lipid-bilayer Membranes

A lipid bilayer membrane is a self-assembled structure that serves as the basis for the cell membranes in the human body, making the study of lipid bilayer membranes an active field in science [52]. The membrane is made up of lipid molecules, which are organic molecules consisting of a polar functional group, known as the lipid head, and a long hydrocarbon chain, known as the lipid tail. From a macroscopic, elastic-physics perspective, the Canham-Helfrich equation dictates that a collection of lipid molecules minimize exposure of their hydrophobic tails as well as their total surface free energy [52]. The "bilayer" aspect means that the structure itself consists of lipid molecules arranged into two oppositely facing sheets of lipids sandwiched together such that the exposure of the tails to water is limited. From a microscopic perspective, this hydrophobic effect stems from a combination of two factors. The first contributor is an entropic force that aims to maximize the number of hydrogen bonds that can be formed by water molecules and prevent any water molecules from being "locked" in position as they would be around a hydrophobic surface. The other effect is the maximization of the attractive dispersion interaction formed as a result of weak, spontaneous dipole moments in the hydrophobic tails. Fig. 1.3 shows an example lipid molecule as well as the structure of the lipid-bilayer membrane. As is the case for most soft-matter systems, there are no chemical bonds formed between lipid molecules in the bilayer, and the

energy scale of their self-assembly is comparable to the ambient thermal energy at room temperature. The lipid molecules are free to diffuse within the plane of the bilayer membrane, which is why lipid bilayer membranes are commonly referred to as 2-D liquids. There are many different types of lipid molecules, varying based on their degree of saturation (the number of double bonds in the tail groups), the length and number of tail groups, and the chemistry of the head group [53]. These differences lead to different membrane properties such as the surface density of the lipids, the rigidity of the membrane, and the diffusivity of the lipid molecules within the plane of the membrane. Other types of biomolecules, such as membrane proteins, and cholesterol, are also found in a typical cell membrane, and often responsible for significant deviations to the properties of pure lipid bilayers. As an initial approximation, however, lipid bilayer membranes provide many useful insights into the biological mechanisms that occur in and around the cell membrane.

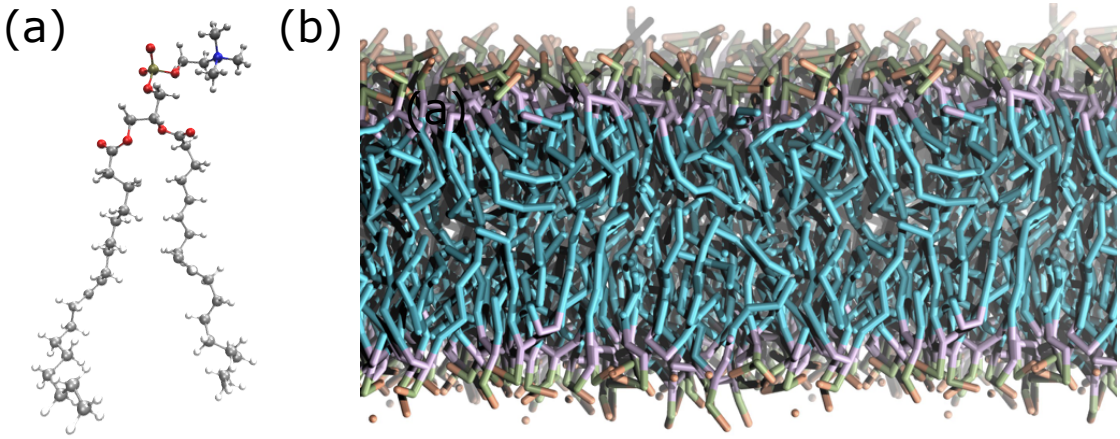

Figure 1.3: (a) The lipid molecule 1,2-Dioleoyl-sn-glycero-3-phosphocholine (DOPC) is shown, with Carbon atoms colored gray, Oxygen atoms red, and Phosphorous atoms orange. (b) A snapshot from an MD simulation of a DOPC lipid-bilayer membrane. The individual atoms are not shown; rather, the molecules are divided into polar head groups and apolar tails. Adapted with permission from Menichetti et al. [54].

Lipid Membrane Permeability

Since lipid membranes behave as two-dimensional liquids, passive permeation of small molecules through the membrane can occur. Lipid membrane permeability is a measure of how quickly molecules can travel across a membrane through either

an active or passive mechanism [51]. The active mechanism refers to the active transport of the permeating molecule, either through some sort of directed motion of the molecule itself or vesicular uptake, meaning that a lipid vesicle fuses with the membrane as in order to facilitate transport of the molecule contained within the vesicle. On the other hand, passive permeation of the lipid membrane occurs as a result of Brownian diffusion only. Therefore, the passive permeability of small molecules is an important property for designing drug molecules, as it tells us whether the drug molecule prefers to pass through the membrane regardless of any external influences. For example, passive permeation of the membrane is known to be a significant mechanism for the uptake of many local anaesthetics [55].

The passive permeability is usually quantified in terms of the permeability coefficient, P. The derivation for this permeability coefficient was first performed by Marrink and Berendsen, and is reproduced below [56].

Given a solute particle of species i, in the low Sherwood number limit (meaning the motion of i is dominated solely by diffusion), its average velocity,u_i, is proportional to the gradient of the potential energy, μ_i, as shown below:

$$u_i = -\frac{1}{\xi_i}\nabla\mu_i. \tag{1.7}$$

Here, ξ_i is the friction coefficient of the ith solute particle. The flux,J_i for the system is given as

$$J_i = c_i u_i = -\frac{c_i}{\xi_i}\nabla\mu_i, \tag{1.8}$$

where c_i is the concentration of species i. We then carry out two substitutions, the first of which uses the Einstein relation,

$$D_i = \frac{k_\mathrm{B} N_\mathrm{A} T}{\xi_i}, \tag{1.9}$$

where D_i is the diffusivity of species i, N_A is the ideal gas constant, and T is the temperature. The second expression to be substituted is the chemical potential in an ideal solution, which is obtained by applying Raoult's law to the expression for the chemical potential in an ideal gas.

$$\mu_i = \mu_i^0 + k_\mathrm{B} N_\mathrm{A} T \ln c_i \tag{1.10}$$

Applying both of these substitutions to Eq. 1.8, we get Fick's first law of diffusion,

$$J_i = -D_i \nabla c_i. \tag{1.11}$$

The only dimension of importance for this case is the z dimension, which is defined as normal to the plane of the lipid membrane. Note the conservation law

$$\frac{\partial J_i(z)}{\partial z} + \frac{\partial c_i(z)}{\partial t} = 0, \tag{1.12}$$

which states that the change in the flux of solute particles must be equal to the change in the concentration profile of those solute particles with respect to time. In the steady state case, this means that the flux is a constant and $\partial J/\partial z$ equals to 0, allowing us to solve Eq. 1.11 for changes in chemical potential as a function of z, $\Delta\mu_i$.

$$\Delta\mu_i = -J_i k_\mathrm{B} N_\mathrm{A} T \int_{z1}^{z2} \mathrm{d}z \frac{1}{c_i(z) D_i(z)} \tag{1.13}$$

Marrink and Berendsen noted that the integral expression on the right hand side is similar to the continuous form of the electrical resistance equation, allowing them to define a corresponding resistance R_i^P as

$$R_i^P = c_i^* \int_{z1}^{z2} \mathrm{d}z \frac{1}{c_i^\mathrm{eq}(z) D_i(z)}, \tag{1.14}$$

where c_i^* is the bulk concentration difference across the membrane for species i, and c_i^eq is the localized equilibrium concentration of i [56]. The inverse of this resistance is known as the passive permeability coefficient. Many standardized experimental techniques have been developed to measure this quantity. For example, the parallel artificial membrane permeability assay (PAMPA) involves measuring the concentration of a drug molecule after being injected onto one side of a small volumetric container containing water and divided by a lipid bilayer membrane [57]. A similar method involves using a type of human epithelial cell called Caco-2 cells, which are considered to provide more biologically accurate results due to the presence of membrane proteins, channels, and microvilli [58]. The experimental setup is essentially identical to that of PAMPA, but the dividing layer consists of a monolayer of caco-2 cells grown over a perforated substrate. Furthermore, these methods often result in permeability coefficients with errors spanning an entire order of magnitude. For this reason, the permeability coefficient is usually expressed as its base-10 logarithm, $\log P$. In the following sections, the methods by which each of the parameters affecting the passive permeability are obtained from computer simulations is discussed.

The Membrane Potential of Mean Force

Eq. 1.14 expresses the permeability coefficient as dependant on the localized equilibrium concentration of the solute particle, c_i^eq. This concentration can also be expressed in terms of the number of microstates in which the solute particle is found in the interval $(z, z + \mathrm{d}z)$ [56]. The equilibrium concentration as a function of z can therefore be written as a "constrained" partition function,

$$c_i^\mathrm{eq}(z) \sim Z(z) = a \int \mathrm{d}\boldsymbol{r}^N \delta(z_0 - z) \exp\left(-\beta \boldsymbol{r}^N\right). \tag{1.15}$$

where a is a proportionality constant. The concentration profile can then be expressed in terms of free energy differences along z by taking the ratio of the localized equilibrium concentration,$c_i^{\mathrm{eq}}(z)$, to the solute concentration in the bulk phase, c_i^*:

$$G(z) = -k_{\mathrm{B}} N_{\mathrm{A}} T \ln\left(\frac{Z(z)}{Z(z^*)}\right) = -k_{\mathrm{B}} N_{\mathrm{A}} T \ln\left(\frac{c_i^{\mathrm{eq}}(z)}{c_i^*}\right), \tag{1.16}$$

where z^* is any z in the bulk phase relative to the membrane. $G(z)$ is the potential of mean force acting in the direction normal to the lipid bilayer midplane, z. Substituting this expression back into Eq. 1.14 yields

$$R_i^P = \int_{z_1}^{z_2} \mathrm{d}z \frac{\exp(G(z)/k_{\mathrm{B}} N_{\mathrm{A}} T)}{D_i(z)} = \frac{1}{P}. \tag{1.17}$$

This equation is commonly used to compute permeability coefficient, P, from computer simulations of drug-membrane permeation [26, 59]. The minimum of the membrane PMF denotes the most thermodynamically favorable distance for a drug molecule to be placed relative to the bilayer midplane. The PMF is also used to determine the barrier heights in the free energy for a molecule to permeate the lipid membrane. Fig. 1.4 shows the structure of a typical PMF calculated for the membrane permeability of a small organic molecule. Note the three key free energy differences that are contained within the PMF, the water/membrane free energy, the membrane/surface free energy, and the water/surface free energy. These three free energy differences form a thermodynamic cycle, meaning that their sum equals to zero.

The Diffusivity

As seen in Eq. 1.11, the diffusivity relates the concentration gradient driving diffusion of particles with the flux of those particles through some area. For a spherical solute particle in a uniform environment, the diffusivity, D, takes the following constant value, commonly known as the Stokes-Einstein equation [61]:

$$D = \frac{k_{\mathrm{B}} T}{6\pi\eta r}. \tag{1.18}$$

Here, k_{B} is the Boltzmann constant, T is the temperature, η is the viscosity and r is the radius of the spherical particle. While this holds for dilute concentrations of particles solvated in a single medium, for particles diffusing through a lipid bilayer membrane, the surrounding environment changes based on the distance away from the bilayer midplane, z. Therefore, a localized diffusivity that is a function of z is used in Eq. 1.17.

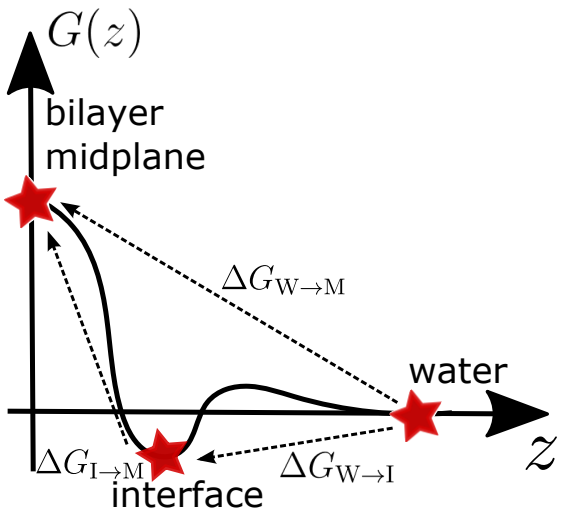

Figure 1.4: A schematic of the potential of mean force for the insertion of a small molecule into a lipid-bilayer membrane. The three key environments are denoted with red stars, with arrows between them indicating the change in free energy between them. Adapted with permission from Menichetti et al. [60].

The Effect of Acidity on Permeability

The acidity or basicity of drug-like molecules will also affect their passive permeability. The acidity/basicity of a molecule refers to its tendency to lose a hydrogen nucleus (deprotonate) or gain a hydrogen nucleus (protonate) resulting in a net negative/positive charge in the molecule [59]. The probability of this type of protonation event occurring depends on the chemical structure of the molecule as well as the concentration of protons in the surrounding water. This is quantified via a dissociation constant which gives the ratio of concentrations of the molecule itself, HA, its dissociated form, A^-, and the dissociated protons, H^+, as dictated by the following chemical reaction

$$HA \rightleftharpoons A^- + H^+. \tag{1.19}$$

The acid dissociation constant, K_a, is then expressed as

$$K_\mathrm{a} = \frac{[A^-][H^+]}{[HA]}, \tag{1.20}$$

where the brackets denote the molar concentration of each species [62]. These concentrations are the equilibrium concentrations corresponding to the acid dissociation reaction shown above. A similar reaction can be written for the molecules

which protonate to gain a net positive charge, known as bases.

$$AH^+ \rightleftharpoons A + H^+ \tag{1.21}$$

The dissociation constants are commonly expressed as pK_as, in which the negative log is applied to the dissociation constant [62].

$$\mathrm{p}K_\mathrm{a} = -\log_{10}(K_\mathrm{a}) \tag{1.22}$$

For each of the chemical reactions shown above, this results in an acidic and a basic pK_a, the apK_a and bpK_a. Note that these are not the standard definitions used to quantify acidity and basicity (pK_a and pK_b). The apK_a is equivalent to the standard definition of the pK_a. The bpK_a, however, is a rearrangement of the chemical reaction commonly used to define basicity, expressed as the deprotonation of a conjugate acid instead of protonation of the base. These nonstandard definitions, which are used by the CHEMAXON software to predict these properties, are introduced here and revisited in Chapter 2 of this work, where we utilized this software.

In the context of drug membrane permeability, at equilibrium, molecules which have multiple protonation states may permeate the membrane in both their protonated and deprotonated form. As seen from Eq. 1.17, the permeability coefficient is the inverse of the membrane resistivity. For the case where multiple permeating species exist, a total resistivity, $R_\mathrm{T}(z)$ can be defined as

$$R_\mathrm{T}^{-1} = R_\mathrm{n}^{-1} + R_\mathrm{c}^{-1}, \tag{1.23}$$

where R_n and R_c correspond to the resistivities of the neutral and charged species, respectively [54, 59]. The individual resistivites shown in this equation have the same form as those seen in Eq. 1.17, with one significant difference. The PMFs must account for the free energy difference between the neutral and charged forms of the molecule. This can be derived from a thermodynamic cycle in which the compound is first neutral in water, neutral in the membrane, charged in the membrane, and charged in water [59]. Solving for the free energy difference between the acidic and basic forms yields the following equation:

$$G_\mathrm{base} = G_\mathrm{acid} + k_\mathrm{B}T(\mathrm{p}K_\mathrm{a} - \mathrm{pH})\ln 10, \tag{1.24}$$

indicating that the changes to the PMF come in the form of a vertical shift, the height of which is determined by the difference between the molecule's apK_a (or bpK_a) and the neutral pH of water, pH = 7.4.

It is also possible for certain molecules to be zwitterionic. These molecules contain separate chemical functional groups with basic and acidic character, retaining

a net neutral charge despite the occurrence of multiple protonation/deprotonation events. In this case, Eq. 1.24 takes the form

$$G_{\text{neut}} = G_{\text{zwit}} + k_{\text{B}}T(\text{bp}K_{\text{a}} - \text{ap}K_{\text{a}}) \ln 10, \tag{1.25}$$

meaning that the vertical shift depends on the relative pK_{a}s (and therefore relative dissociation strength) of each functional group. Eq. 1.25 assumes that the dissociation reactions are completely independent from each other.

1.1.3 Molecular Dynamics Simulations

MD simulations are a computational tool used to model many different types of systems, including soft-matter systems. In an MD simulation, atoms are represented as particles whose interactions are governed by classical (as opposed to quantum) mechanics, although different flavors of MD also exist that account for electronic degrees of freedom [63, 64]. Newton's equations of motion (specifically Newton's second law) are then solved in order to propagate the system in time. Assuming the simulation has been properly initialized and has propagated for a sufficient amount of time, many thermodynamic quantities from the simulation can be computed and compared to experimental results. MD can also be used to measure kinetic properties and observe specific transitions in the system of interest [65, 66]. In general, the steps in an MD simulation are as follows: for each time step, the forces acting on each particle in the system are calculated, and the equations of motion are integrated in order to determine the changes in particle position and velocity for the following time step. In this section, we detail the specific numerical machinery used to implement these steps. Most of these explanations are adapted directly from the textbook written by Frenkel and Smit, which is recommended for a thorough understanding of MD simulations [46]. Additionally, the lecture notes of M. Scott Shell are frequently cited as they provide a cogent and intuitive summary of many of the same concepts [45].

The Force Field

After initializing the system, the first step in the MD algorithm is the calculation of forces. The forces are obtained by taking the derivative of the potential energy function, $U(\boldsymbol{r}^N)$, with respect to the relative positions of the particles themselves. This potential energy function, commonly referred to as the force field, is usually expressed as a sum of intramolecular and intermolecular interactions between particles, U_{intra} and U_{inter}.

$$U(\boldsymbol{r}^N) = U_{\text{intra}} + U_{\text{inter}} \tag{1.26}$$

$\boldsymbol{r}^N$ refers to the positions of all particles in the system, otherwise known as a single configuration of the system [45]. The intramolecular part of the energy function is further expressed as a sum over all bonded, angle, and dihedral interactions within the molecules of the system. Each of these interaction types approximate the energy corresponding to the electronic probability density obtained by solving Schrödinger's equation for the atoms in the ground state with fixed nuclear positions. Many atomistic resolution force fields have been obtained by fitting energies calculated via ab initio methods to simple analytical functions. In these cases, bonded and angle interactions are usually written as harmonic oscillators that deviate from the equilibrium bond distances, d_0, and angles, θ_0, whereas dihedral interactions are described using a cosine series [45]:

$$U_{\text{intra}} = \sum_{\text{bonds}} a(d - d_0)^2 + \sum_{\text{angles}} b(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} \left(\sum_{n=0}^{N} c_n \cos(\omega)^n \right). \quad (1.27)$$

In this equation, a, b, are constants that determine the strength of the harmonic potentials. Similarly, c_n coefficients are defined for the cosine series used to approximate the ab initio dihedral potential for all possible dihedral angles, ω.

The intermolecular part of the potential energy function, U_{inter}, is separated into two parts, corresponding to the pairwise neutral and charged non-bonded interactions between particles belonging to different molecules in the system, as given by the following equation [45]:

$$U_{\text{inter}} = \sum_{\text{pairs}} \left(4\epsilon \left[\left(\frac{r_{ij}}{\sigma} \right)^{-12} - \left(\frac{r_{ij}}{\sigma} \right)^{-6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right). \quad (1.28)$$

For many systems, the interactions between two neutral non-bonded particles is expressed as a Lennard-Jones potential (shown as the first term in equation 1.28), which accounts for Pauli repulsion and van der Waals attraction as a function of the pairwise distance, r_{ij} [46]. Here, ϵ and σ are Lennard-Jones parameters that depend on the atoms i and j that make up the pair. The contribution to the Lennard-Jones potential becomes minimal after $r_c \approx 2.5\sigma$ [45]. Therefore, in practice, the potential is usually truncated (by cutting and shifting the potential to avoid discontinuities) after this cutoff distance, r_c, to avoid iterating over all pairs in the system, thereby reducing computational overhead. However, this contribution becomes significant when computing the total pressure and potential energy of the system. Assuming that the system is isotropic and homogeneous in composition beyond the cutoff distance (i.e., $g(r > r_c) = 1$), a correction can be added to the net potential energy and pressure by integrating the Lennard-Jones interaction from $0 \to \infty$ for the energy and $r_c > r > \infty$ for the pressure, both of which can be expressed analytically [45, 46]. The force due to electrostatic interactions is, in theory, perfectly expressed via Coulomb's law, shown in the second

term in Eq. 1.28. q_i and q_j are the charges assigned to atoms i and j, and ϵ_0 is the electric permittivity in a vacuum. In practice, the Ewald summation method is commonly used to evaluate this contribution to the intermolecular interactions [46]. In this approach, the electrostatic interaction is split into a short-range and long-range contribution, with the electrostatic force being calculated in real space for the short-range contribution and in Fourier space for the long-range contribution. This method drastically reduces the time needed for the force calculation to converge, making it advantageous when compared to performing a summation over all charged particles in the system. The Particle Mesh Ewald summation further reduces this computational cost by employing the Fast Fourier Transform algorithm to calculate the long-range term, which requires a projection of the charge densities onto a discretized grid [46].

Evolving the System in Time

When running MD simulations, several methods have been developed to numerically integrate Newton's equations of motion and accurately compute positions and velocities for each particle at each time step. In this section we will derive only two, the Verlet algorithm and the Leap-Frog algorithm, the latter of which is used in this work.

The following derivation of the Verlet algorithm is lifted from the work of Frenkel and Smit [46]. A particle with mass m and position r after some time t can be approximated via Taylor expansion to the fourth order:

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 + \frac{\Delta t^3}{3!}r'''(t) + \mathcal{O}(\Delta t^4), \tag{1.29}$$

where $v(t)$ and $f(t)$ are the velocity of the particle and the force acting on the particle, respectively, and $r'''(t)$ refers to the third derivative of r with time. A similar expression can be derived for the particles previous position, making the algorithm time-reversible.

$$r(t-\Delta t) = r(t) - v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 - \frac{\Delta t^3}{3!}r'''(t) + \mathcal{O}(\Delta t^4) \tag{1.30}$$

By summing the previous two equations and moving $r(t-\Delta t)$ to the right hand side, we obtain

$$r(t+\Delta t) \approx 2r(t) - r(t-\Delta t) + \frac{f(t)}{m}\Delta t^2 + \mathcal{O}(\Delta t^4), \tag{1.31}$$

which is the equation used in the Verlet algorithm. The equation is evaluated up to the second-order term and has an error that is of the order Δt to the

fourth power. However, calculating the second-order term will cause numerical imprecision to arise. The forces $f(t)$ are obtained by taking the derivative of the potential U at time t with respect to the particles position, r.

$$f(t) = -\frac{dU(r(t))}{dr}. \tag{1.32}$$

The Leap-Frog equation is derived in a similar fashion, but begins with expressions for the velocities of the particles at time $t - \Delta t/2$ and $t + \Delta t/2$,

$$v(t - \Delta t/2) = \frac{r(t) - r(t - \Delta t)}{\Delta t} \tag{1.33}$$

and

$$v(t + \Delta t/2) = \frac{r(t + \Delta t) - r(t)}{\Delta t} \tag{1.34}$$

The second expression can be rearranged to obtain the positions at the next time step:

$$r(t + \Delta t) = r(t) + v(t + \Delta t/2)\Delta t, \tag{1.35}$$

and the sum of the two equations is used to obtain the velocities

$$v(t + \Delta t/2) = v(t - \Delta t/2) + \frac{f(t)}{m}\Delta t. \tag{1.36}$$

Because the leap-frog equation comes from the Verlet equation, the trajectories produced by each should be identical [46]. However, the numerical imprecision that stemmed from the second-order term in Eq. 1.31 is no longer present. Therefore, using this algorithm enables our system to evolve in time in a deterministic fashion while maximizing precision.

Calculation of Average Properties

In general, to compute the average, $\bar{A}$, of some observable A from an MD simulation, the following integral is evaluated over some length of simulation time τ [46].

$$\bar{A} = \frac{1}{\tau}\int_0^\tau \mathrm{d}t A(t) \tag{1.37}$$

This is value is assumed to be equivalent to the true statistical mechanical average at equilibrium if the simulation has first been properly equilibrated, and then run for long enough time such that A is sampled over multiple correlation times [45, 46].

Controlling Temperature and Pressure

In order to run simulations in the canonical ensemble, it is necessary for the system to maintain a fixed temperature. This is equivalent to ensuring that the system is in contact with a large heat bath which is at the desired temperature [46]. Note that, for a system at constant temperature T, the probability density function, P for finding a specific particle with mass m_i and momentum p_i corresponding to T is known as the Maxwell-Boltzmann distribution [44].

$$P(p_i) = \left(\frac{\beta}{2\pi m_i}\right)^{3/2} \exp\left[-\beta p_i^2/(2m_i)\right] \tag{1.38}$$

The Andersen thermostat is used to maintain a constant temperature by randomly selecting particles in the system to undergo a simulated collision with the heat bath at fixed time intervals [45]. The simulated collision in this case is just a reassignment of the particle velocities to velocities taken from the Maxwell-Boltzmann distribution. Note that the velocity-rescaling thermostat used in Chapters 2 and 3 of this work smoothly incorporates this canonical equilibrium distribution for the kinetic energy directly into the integrator used to evolve the system in time [67].

The Nosé thermostat enables a deterministic approach for maintaining constant temperatures in MD simulations [46]. For a system with N particles, this is achieved by using an extended Lagrangian formulation which explicitly includes an imaginary heat bath with corresponding position and momentum,

$$\mathcal{L}_{\text{Nosé}} = \sum_i^N \left(\frac{m_i}{2} s^2 v_i^2\right) - U(\boldsymbol{r}^N) + \frac{Q}{2} v_s^2 - \frac{3N+1}{\beta} \ln(s). \tag{1.39}$$

The variable s can be considered as the position of the imaginary heat reservoir coupled to the system, with velocity v_s and mass Q. Performing a Legendre transform on this Lagrangian yields the corresponding Hamiltonian

$$H_{\text{Nosé}} = \sum_i^N \left(\frac{p_i^2}{2m_i s^2}\right) + U(\boldsymbol{r}^N) + \frac{p_s^2}{2Q} + (3N+1)\frac{\ln(s)}{\beta}. \tag{1.40}$$

The equations of motion can then be derived and implemented using this Hamiltonian, requiring the mass of the heat bath, Q, to be specified by the user. This Hamiltonian yields the following partition function in the cannonical ensemble:

$$Z_{\text{Nosé}} = \frac{C}{N!} \int \mathrm{d}p_i' \mathrm{d}\boldsymbol{r}^N \exp\left[-\beta \sum_i^N \left(\frac{p_i'^2}{2m_i}\right) + U(\boldsymbol{r}^N)\right]. \tag{1.41}$$

Here, $p_i^{'}$ is the momentum scaled by the variable s, and C is the scaled kinetic energy contribution of the heat bath, which is constant at fixed T. We can now compute average properties using Eq. 1.37, but with the scaled momentum.

$$\bar{A} = \frac{1}{\tau}\int_0^\tau \mathrm{d}t A(p'(t), r(t)) \tag{1.42}$$

By evolving the system using this approach, both the coordinates and momenta for the particles as well as the imaginary heat bath are obtained deterministically [46]. The relationship between s and p_i is key to maintaining a constant temperature. Because the scaled quantities are ultimately used to progress the simulation, they are denoted here with the $'$ symbol as being the "real" variable whereas the lack of this symbol denotes a "virtual" variable. While r remains unscaled, the momentum p_i and time step Δt are both scaled by s.

$$\Delta t' = \Delta t / s. \tag{1.43}$$

However, because the position of the imaginary heat bath, s, will vary with time, the previous equation implies that the time step is not a constant during the simulation. A modification of the previously defined Hamiltonian was proposed by Hoover [46].

$$H_{\text{Nosé–Hoover}} = \sum_i^N \left(\frac{p_i^2}{2m_i}\right) + U(\boldsymbol{r}^N) + \frac{\xi^2 Q}{2} + 3N\frac{\ln(s)}{\beta} \tag{1.44}$$

Where the friction coefficient, ξ, is defined as

$$\xi = \frac{\mathrm{d}\ln s}{\mathrm{d}t}. \tag{1.45}$$

This reformulation results in equations of motion that no longer imply a variable time step. Rather, the value of the friction term changes based on fluctuations in the instantaneous kinetic energy. The mass of the heat bath remains a parameter to be specified by the user.

In order to simulate isobaric ensembles, a barostat is used to maintain constant pressure. Here, we focus on the Parinello-Rahman Barostat [68]. This barostat is constructed in a similar fashion to the Nosé-Hoover thermostat, but with a pressure-bath, rather than a heat bath, being coupled to the system in an extended Lagrangian formulation. The vector corresponding to the simulation box, b is coupled to the pressure bath via the following equation:

$$\frac{\mathrm{d}b^2}{\mathrm{d}t^2} = V\boldsymbol{W}^{-1}b'^{-1}(\boldsymbol{P} - \boldsymbol{P}_{\text{ref}}). \tag{1.46}$$

where V is the volume of the box, and $\boldsymbol{P}_{\text{ref}}$ and $\boldsymbol{P}$ are the reference (desired) and instantaneous pressure tensors, respectively [68]. $\boldsymbol{W}^{-1}$ is known as the inverse mass parameter matrix, and it determines the strength of the coupling. The Hamiltonian, when including this coupling, is as follows:

$$H_{\text{PR}} = \sum_{i}^{N} \left(\frac{p_i^2}{2m_i} \right) + U(\boldsymbol{r}^N) + \sum_{j} \boldsymbol{P}_{jj} V + \sum_{j,k} \frac{1}{2} \boldsymbol{W}_{jk} \left(\frac{\mathrm{d}b_{jk}}{\mathrm{d}t} \right)^2 . \tag{1.47}$$

Both the Nosé-Hoover thermostat and the Parinello-Rahman barostat are usually used in tandem, with a joint Hamiltonian constructed from a Lagrangian that contains both heat and pressure-bath couplings. Deriving the equations of motion from this Hamiltonian enables the simulation of systems in the NPT ensemble.

Constraints

The time step used when running atomistic MD simulations is usually limited by the frequency of harmonic oscillations used to model bonds between atoms [46]. Therefore, one strategy that enables the use of a larger time step is to replace these harmonic oscillations with constraints that are holonomic. The term holonomic is used to describe a type of constraint which is only dependant on the positions of the particles and the time, t. In this work, an algorithm known as the LINear Constraint Solver, or LINCS, is used to constrain the bonds containing Hydrogen atoms in atomistic systems [69]. This enables the use of a 2 fs timestep when running atomistic MD simulations of bulk organic liquids, further discussed in chapter 4. The method and its implementation were developed by Hess et al, and a detailed derivation and implementation can be found in their work [69]. We now provide a short summary of that derivation. For a system with N particles, we define a position vector, r, for each particle. Assuming K number of bonds to be constrained, the holonomic constraint expression is as follows:

$$h_i(r) = |r_{i_1} - r_{i_2}| - d_i = 0 \qquad i = 1, \ldots, K, \tag{1.48}$$

where $|r_{i_1} - r_{i_2}|$ is the distance between constrained particles 1 and 2, and d_i is the user-specified length of the ith constraint [69]. The implementation of these constraints takes place in three steps for each time step. The first step consists of performing an unconstrained update, allowing for changes in the bond length and orientation. Next, the projection of the forces in the bond direction from the previous time step is removed. In the last step, corrections are applied to account for lengthening of the bond due to rotation (i.e., to account for the centripetal forces remaining after removing the forces in the previous step).

The Radial Distribution Function

The radial distribution function (RDF) is commonly used to quantify the structure of a simulated bulk, condensed-phase system [45]. For a system in the canonical ensemble with only a single particle type, the RDF, g, between two particles with positions r_1 and r_2 is defined as the following:

$$g(r_1, r_2) = \frac{V^2(N-1)}{N} \frac{\int \mathrm{d}r_3 \mathrm{d}r_4 ... \mathrm{d}r_N \; e^{-\beta U(\boldsymbol{r}^N)}}{Z(N,V,T)}, \tag{1.49}$$

where V is the volume of the system and $Z(N,V,T)$ is the partition function [45]. This can be extended for different particle types A and B as shown below:

$$g_{AB}(r_1, r_2) = V^2 \frac{\int \mathrm{d}r_3 \mathrm{d}r_4 ... \mathrm{d}r_N \; e^{-\beta U(\boldsymbol{r}^N)}}{Z(N_A, N_B, V, T)}. \tag{1.50}$$

An example RDF is shown in Fig. 1.5. Conceptually, the RDF is the probability of finding another particle of the desired type some distance away from the reference particle. For soft-matter systems in the isotropic bulk-liquid phase, this value approaches 1 at long distances. The Fourier transform of the RDF, the static structure factor, can also be obtained by performing small-angle and wide-angle scattering, making the RDF a useful tool for validating MD simulations using experimental scattering data.

Thermodynamic Integration

Previously, we defined a general equation for the free energy difference between two states using Eq. 1.4. One method to compute this property using MD trajectories is called thermodynamic integration. This is done by first defining the thermodynamic path between the two states and breaking this path into intermediate states [70]. Typically, these intermediate states are specified via a coupling parameter, λ, that ranges from 0 to 1, where $\lambda = 0$ is equivalent to state 1 and $\lambda = 1$ is equivalent to state 2. In order to compute the free energy difference between states 1 and 2 we first take the derivative of the free energy A with respect to λ:

$$\frac{\mathrm{d}A}{\mathrm{d}\lambda} = -\frac{1}{\beta Z} \frac{\mathrm{d}}{\mathrm{d}\lambda} \int \mathrm{d}\boldsymbol{r}^N e^{-\beta U(\lambda, \boldsymbol{r}^N)}. \tag{1.51}$$

which can then be expressed as the ensemble averaged derivative of the potential with respect to λ:

$$\frac{\mathrm{d}A}{\mathrm{d}\lambda} = \left\langle \frac{\mathrm{d}U(\lambda, \boldsymbol{r}^N)}{\mathrm{d}\lambda} \right\rangle_\lambda. \tag{1.52}$$

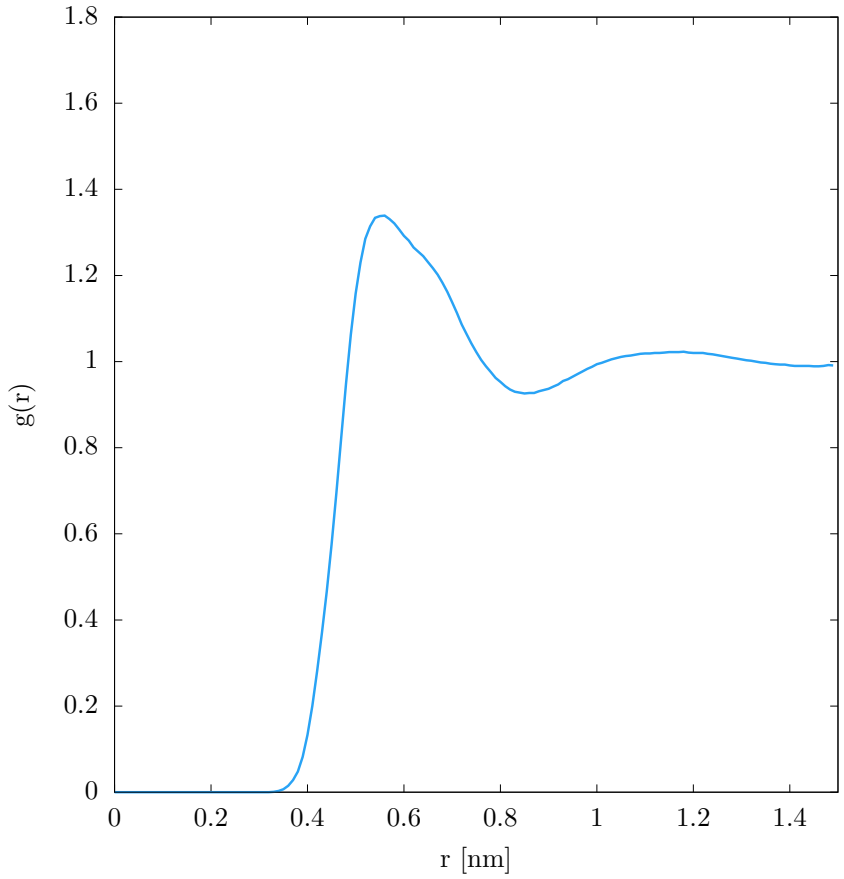

Figure 1.5: An example intermolecular radial distribution function for a homogeneous, bulk-phase organic liquid.

Integrating this expression over all lambda then yields the change in free energy:

$$\Delta A = \int_1^2 \mathrm{d}\lambda \left\langle \frac{\mathrm{d}U(\lambda, \boldsymbol{r}^N)}{\mathrm{d}\lambda} \right\rangle_\lambda . \tag{1.53}$$

In practice, since a finite number of trajectories are run at different lambda values, a numerical integration is performed [70]:

$$\Delta A \approx \sum_{k=1}^{K} w_k \left\langle \frac{\mathrm{d}U(\lambda, \boldsymbol{r}^N)}{\mathrm{d}\lambda} \right\rangle_k . \tag{1.54}$$

Here, w_k refers to the weights that correspond to the free-energy histogram with total number of bins K used when performing the numerical integration. A simple use-case for thermodynamic integration is to calculate the free energy difference of changing a Hydrogen atom to a Fluorine atom. In this example, λ is chosen as a switching parameter between the potential applied to the Hydrogen atom and the Fluorine atom, respectively. To ensure accurate results, enough λ states must be simulated such that $\mathrm{d}U/\mathrm{d}\lambda$ is smooth and continuous over λ.

The Bennett Acceptance Ratio

In many cases, running a large number of simulations at different lambda values is computationally unfeasible, especially for large system sizes. Another method to estimate free energy differences was derived by Bennett, requiring only two trajectories at the initial and final thermodynamic state points [46]. The Bennett Acceptance Ratio method is derived from Eq. 1.4, which we restate below:

$$\Delta A = -\frac{1}{\beta} \ln \left(\frac{\int \mathrm{d}\boldsymbol{r}^N e^{-\beta U_2(\boldsymbol{r}^N)}}{\int \mathrm{d}\boldsymbol{r}^N e^{-\beta U_1(\boldsymbol{r}^N)}} \right) . \tag{1.55}$$

Bennett modified this expression by multiplying and dividing an expression similar to that used in free-energy perturbation methods, in which this expression would be the partition function of only one of the states. However, unlike the free-energy perturbation approach, the expression used here results in a free energy difference that is based on averages over both trajectories [46].

$$\begin{aligned} \Delta A = -\frac{1}{\beta} \ln \left(\frac{\int \mathrm{d}\boldsymbol{r}^N e^{-\beta U_2(\boldsymbol{r}^N)}}{\int \mathrm{d}\boldsymbol{r}^N e^{-\beta U_1(\boldsymbol{r}^N)}} \frac{\int \mathrm{d}\boldsymbol{r}^N w(\boldsymbol{r}^N) e^{-\beta U_1(\boldsymbol{r}^N) - \beta U_2(\boldsymbol{r}^N)}}{\int \mathrm{d}\boldsymbol{r}^N w(\boldsymbol{r}^N) e^{-\beta U_1(\boldsymbol{r}^N) - \beta U_2(\boldsymbol{r}^N)}} \right) \\ = -\frac{1}{\beta} \ln \left(\frac{\langle w e^{-\beta U_2} \rangle_1}{\langle w e^{-\beta U_1} \rangle_2} \right) \end{aligned} \tag{1.56}$$

Here, $w(\boldsymbol{r}^N)$ is a weighting function that is obtained by minimizing the variance of the free energy difference to be:

$$w(\boldsymbol{r}^N) = \frac{C}{(Z_1/n_1)\exp(-\beta U_2) + (Z_2/n_2)\exp(-\beta U_1)}, \tag{1.57}$$

where C is a constant to be determined self-consistently. If we plug this equation into Eq. 1.56, we get

$$\Delta A = -\frac{1}{\beta}\ln\left(\frac{\langle f(U_2 - U_1 + C)\rangle_1}{\langle f(U_1 - U_2 - C)\rangle_2}\right). \tag{1.58}$$

where $f(x)$ is the fermi-dirac function,$f(x) = 1/(1+\exp(\beta x))$. A caveat when using the Bennett Acceptance Ratio is that there must be significant overlap in phase space between the trajectories at the different state points. In the case where there is no phase space overlap between the state points, a λ coupling parameter can be used in a similar fashion to thermodynamic integration, to bridge the gap in phase space. The overall free energy difference is then the sum of each of the free energy differences between consecutive lambda states.

Umbrella Sampling and the Weighted Histogram Analysis Method

For many systems, large free energy barriers exist that prevent the sampling of certain energetically favorable microstates in a computationally tractable amount of time when sampling using conventional MD simulations. For example, the free energy barrier for a highly polar molecule to enter a lipid bilayer membrane is roughly 20 times the ambient thermal energy, k_BT, at room temperature [26]. Computational methods used to quickly overcome these barriers are known as enhanced sampling techniques. One of these methods, known as umbrella sampling, is specifically used in this work to obtain potentials of mean force in the aforementioned lipid bilayer membrane example [71]. Generally, this involves applying a bias potential that is a function of the reaction coordinates (also referred to as collective variables) upon which the free energy landscape is projected. When calculating the lipid membrane potential of mean force for a small molecule, we use the distance along the normal to the bilayer midplane, z, as the reaction coordinate. The full range of z values is first divided into simulation windows. For each window, i, a harmonic biasing potential, $w(z)$, is included in addition to the unbiased potential, U^{u}, yielding the biased potential, U^{b} as shown in the following equation:

$$U^{\mathrm{b}} = U^{\mathrm{u}} + w_i(z) = U^{\mathrm{u}} + k_i(z - z_{0,i})^2. \tag{1.59}$$

The choice of force constant,k_i, used for these harmonic potentials is important, as small values will prevent barrier crossing, whereas large values result in limited

sampling along the reaction coordinate. After sufficiently sampling each window via an MD simulation, the probability distribution corresponding to the biased potential is obtained, $P^{\mathrm{b}}(z)$. This is related to the unbiased probability, $P^{\mathrm{u}}(z)$ via the following expression:

$$P^{\mathrm{u}}(z) = P^{\mathrm{b}}(z)e^{\beta w(z)}\langle e^{-\beta w(z)}\rangle. \tag{1.60}$$

By taking the logarithm of equation 1.60, we obtain the potential of mean force for each window, $G_i(z)$:

$$G_i(z) = \frac{-1}{\beta}\ln P_i^{\mathrm{b}}(z) + \frac{1}{\beta}w_i(z) + C_i, \tag{1.61}$$

where C_i is a constant. This constant must be computed in order to combine each of these PMFs into a $G(z)$ which spans all of the previously defined windows. The method used in this work to obtain this constant is known as the weighted histogram analysis method [72]. Using this method, the global unbiased probability distribution is expressed as a weighted average of the distributions from each window:

$$P^{\mathrm{u}}(z) = \sum_{i}^{W} p_i(z)P_i^{\mathrm{u}}(z), \tag{1.62}$$

where W is the total number of windows and $p_i(z)$ are the weights. By minimizing the error of $P_i^{\mathrm{u}}(z)$, and applying a normalization condition on the weights, the following equations are obtained:

$$p_i(z) = \frac{a_i z}{\sum_j^W a_j} \tag{1.63}$$

$$a_i(z) = N_i e^{-\beta w_i(z)+\beta C_i}. \tag{1.64}$$

Here, a_i is another constant which, along with C_i, must be solved self-consistently. Solving equations 1.63 and 1.64 enables the unbiased probability distribution to be calculated, which can then be substituted into equation 1.6 to obtain the unbiased potential of mean force.

1.1.4 The Generated Database of Compounds

The Generated DataBase (GDB) is a list of organic compounds automatically generated using a computer algorithm developed by Fink et al. [73]. The algorithm proceeds in the following steps. First, a large number of graphs are systematically generated with a maximum of four edges per node. The nodes correspond to the Carbon atoms in saturated Carbon chains, and the edges represent carbon-carbon

single bonds. Next, a series of filters is applied to remove graphs representing non-physical or highly unstable structures. For example, some of these filters remove graphs from the data set if they contain single carbon atoms shared across multiple 3- or 4-membered rings, or if the energy-minimized ring strain is above a certain threshold value. In the next step, double and triple bonds are combinatorially introduced to the remaining structures, followed by hetero-atom substitutions. This results in several highly unstable combinations of hetero-atoms being covalently bonded to each other, and another filter is applied to remove these instances. The final step is to remove compounds that are tautomers of each other, meaning that they can spontaneously interconvert between the two structures under ambient conditions. By setting the maximum number of heavy atoms to be 11 and only allowing N, O, and F substitutions, the GDB-11 was created, containing 26.4 million unique chemical compounds [73]. Note that the number of compounds increases exponentially with respect to the number of heavy atoms per compound. Therefore, in this work, only the molecules containing up to 10 heavy atoms (as well as subsets of this database) are used as a proxy for CCS. This data set contains approximately 3.5 million molecules. The GDB-11 can be found on the Reymond group's website and is stored as a series of text files, with each text file containing a list of compounds for a given number of heavy atoms. The compounds themselves are represented as a simplified molecular-input line-entry system (SMILES) string, which is an intuitive notation used to represent chemical compounds [74].

Chemical Functional Groups

In this work, we define a chemical functional group as a single perturbation or a localized series of perturbations of a saturated carbon scaffold. A perturbation in this instance can be either a replacement of a single bond with a double or triple bond, or the replacement of carbon atoms with another atom type. The former case is referred to as a bond substitution and the latter case is called a hetero-atom substitution. We used the program CHECKMOL, developed by Norbert Haider, to automatically identify the functional groups present in chemical compounds from the GDB [75]. The program requires a 3D structure file containing the coordinates and topology of a compound as an input. It then checks for the presence of specific functional groups by calling a series of subroutines that exhaustively search for different types and combinations of bond substitutions and hetero-atom substitutions. In this way, over 200 different functional groups can be identified by CHECKMOL.

1.1.5 Coarse-Graining

As previously discussed in the introduction of this chapter, coarse-grained (CG) models provide a means to drastically reduce the computational cost of simulating soft matter systems while retaining their underlying physics. In this section, we describe the general steps involved in building CG models and highlight many of the different approaches that have been developed in this regard. For more thorough explanations of the fundamental concepts discussed here, the reader is referred to this review by W.G. Noid [31].

Bottom-up vs. Top-down Coarse Graining

There are two broad categories into which coarse-grained modeling can be classified: bottom-up and top-down approaches. Bottom-up coarse-grained models are constructed using data from higher-resolution models of the same system. In this work, the higher-resolution models refer to AA MD simulations, although this is itself a coarsened approximation of an ab initio model. Rigorous methods have been developed that guarantee the preservation of the underlying physics at the higher resolution in the resulting CG model [76–78]. Therefore, the accuracy of bottom-up CG models depends on the quality of the higher-resolution data which is used in its construction. Consequently, bottom-up CG models tend to be highly chemically-specific, meaning that they are only meant to model certain molecules under specific conditions, because of the expensive requirement for high-quality higher-resolution data. However, rigorous methods for extending the transferability of bottom-up CG models have also been developed [40, 41].

On the other hand, top-down methods do not rely on higher resolution models, but instead are constructed to match specific macroscopic properties, usually obtained from experiments. This approach can provide insight as to what physical concepts must be included in a CG model in order to explain phenomena observed at the macroscopic scale. Furthermore, since many top-down CG models aim to reproduce these phenomena using as few parameters as possible, the degree of chemical specificity in top-down models is often significantly reduced compared to bottom-up models [79]. However, it is by no means evident that these top-down CG models can be easily related to the physical phenomena of the same system at the microscopic level. For example, certain thermodynamic properties, like the bulk density or partitioning free energy of a compound, may be accurately reproduced by a top-down model while failing to reproduce the specific conformations or packing behavior at the atomistic resolution.

The Mapping Function

The first step in the coarse graining process is to define a mapping function $M(\boldsymbol{r})$ [31]. This function assigns atoms in the high resolution MD trajectory with position $\boldsymbol{r}$ to pseudo-atoms called beads, and sets their configuration, $\boldsymbol{R}$, according to some assignment rule. For example, the coordinates $\boldsymbol{R}_I$ for a single coarse grained bead I, can be expressed as the weighted center of mass of all of the atomic positions i that correspond to I at the all-atomistic (AA) resolution.

$$\boldsymbol{R}_I = M_I(\boldsymbol{r}) = \sum_i c_{Ii}\boldsymbol{r}_i. \tag{1.65}$$

Since the total degrees of freedom are reduced in the CG system, it is impossible to preserve all of the features of the high resolution system, regardless of the accuracy of the potentials assigned to the CG beads. This makes the choice of which atomistic fragments should be mapped to a single bead an important one, although, in practice, this is often based on chemical intuition alone. Recently, however, more systematic methods have been developed to diagnose the quality of CG mappings [80, 81].

The Many-Body Potential of Mean Force

The next step in the coarse-graining process is to determine the coarse-grained potential. When taking a bottom-up approach, the goal is to ensure that the probabilities of obtaining coarse-grained configurations, $P_{\mathrm{CG}}(\boldsymbol{R})$, are the same as the corresponding atomistic configurational probabilities, $p_{\mathrm{AA}}(\boldsymbol{r})$ [45].

$$P_{\mathrm{CG}}(\boldsymbol{R}) = p_{\mathrm{AA}}(\boldsymbol{r}) = \frac{e^{-\beta U_{\mathrm{CG}}(\boldsymbol{R})}}{Z_{\mathrm{CG}}} = \frac{e^{-\beta u_{\mathrm{AA}}(\boldsymbol{r})}}{z_{\mathrm{AA}}}, \tag{1.66}$$

where Z_{CG} and z_{AA} are the partition functions corresponding to the CG and AA systems. We then express $p_{\mathrm{AA}}(\boldsymbol{R})$ in terms of only the atomistic configurations and the mapping function, $M(\boldsymbol{r})$.

$$p_{\mathrm{AA}}(\boldsymbol{R}) = \int \mathrm{d}\boldsymbol{r} p_{\mathrm{AA}}(\boldsymbol{r})\delta[M(\boldsymbol{r}) - \boldsymbol{R}] \tag{1.67}$$

Note that multiple atomistic configurations will map to the same coarse-grained configuration. Plugging this equation into Eq. 1.66 yields the following expression:

$$\frac{e^{-\beta U_{\mathrm{CG}}(\boldsymbol{R})}}{Z_{\mathrm{CG}}} = \int \mathrm{d}\boldsymbol{r} \frac{e^{-\beta u_{\mathrm{AA}}(\boldsymbol{r})}}{z_{\mathrm{AA}}} \delta[M(\boldsymbol{r}) - \boldsymbol{R}]. \tag{1.68}$$

Next, we take the log of both sides of the equation and group the expressions containing partition functions together as a constant, C:

$$U_{\mathrm{CG}}(\boldsymbol{R}) = -\frac{1}{\beta} \ln \int \mathrm{d}\boldsymbol{r} e^{-\beta u_{\mathrm{AA}}(\boldsymbol{r})} \delta[M(\boldsymbol{r}) - \boldsymbol{R}] + C. \tag{1.69}$$

The first term on the right hand side is known as the many-body potential of mean force (MBPMF) [31, 45]. Like the PMF defined in Eq. 1.6, it is also a projection of the free energy. In this case, however, the MBPMF is a projection of the atomistic free energy surface onto the coarse-grained degrees of freedom, as specified by the mapping function. Note that there is no restriction on the functional form of the MBPMF. The "many-body" aspect of the MBPMF refers to the fact that integrating over the residual atomistic degrees of freedom results in the generation of many-body interactions between CG beads. Therefore, conventional CG potentials, which are expressed as a sum of pair-wise contributions, will never fully approximate the MBPMF [31].

Direct Boltzmann Inversion

Because it is computationally unfeasible to calculate the MBPMF for large complex systems, an approximate CG potential is often developed in order to reproduce certain structural distributions of the atomistic systems. The simplest approach to doing so is known as direct Boltzmann inversion [76]. As a bottom-up approach, it initially requires the generation of an atomistic MD trajectory of the system. After determining the CG mapping scheme, an atomistic structural distribution is projected onto the corresponding CG degrees of freedom. From this distribution, labeled $p_\zeta(x)$, the CG potentials U_ζ for each intramolecular and pairwise interaction ζ are obtained via the following equation:

$$U_\zeta = -\frac{1}{\beta} \ln \left(\frac{p_\zeta(x)}{J_\zeta(x)} \right), \tag{1.70}$$

where x refers to the interaction-specific variable (i.e., distances or angles), and $J_\zeta(x)$ is the Jacobian factor for the specific interaction type. For example, bonded distributions should be scaled by the square of the bond length, whereas angle distributions should be scaled by the sine of the angle [76]. Direct Boltzmann inversion is mainly successful when the interactions modelled are highly isolated from the other interactions occurring in the system. However, if there are any interactions that are strongly coupled, applying this method will ignore the cross-correlations between them, and the subsequent CG structural distributions will not match those of the AA reference.

Iterative Boltzmann Inversion

While direct Boltzmann inversion will fail to reproduce atomistic distributions in systems with coupled interactions, it is possible that iteratively tuning these potentials and measuring the change to the resulting CG structural distribution will lead to an improved CG potential. The Iterative Boltzmann inversion approach proceeds in the following steps [45, 82]. First, compute coarse-grained potentials via direct Boltzmann inversion. Next, run a CG MD simulation and calculate the corresponding structural distributions. The third step is to modify the CG potential via the following expression:

$$U_{\zeta,\mathrm{new}}(x) = U_{\zeta,\mathrm{old}}(x) - \frac{1}{\beta} \ln \left(\frac{p_{\zeta,\mathrm{AA}}(x)}{p_{\zeta,\mathrm{old}}(x)} \right). \tag{1.71}$$

The previous two steps are repeated until the coarse-grained potential converges. Simultaneously applying this method to the full CG potential during each iteration implicitly accounts for cross-correlations between interactions. However, for systems containing highly coupled cross-correlations compounded over many different interactions, convergence of the CG potential may not be possible.

The Multiscale Coarse Graining Method

Unlike the previous two methods, which specifically aim to reproduce structural distributions found in the reference atomistic system, certain bottom-up CG methods aim to produce a CG potential that best approximates the MBPMF via a variational approach. One such method, known as Multiscale Coarse Graining (MSCG), finds the CG potential that minimizes the following functional:

$$\chi_1^2[U] = \left\langle \frac{1}{3N} \sum_I \left| \boldsymbol{f}_I(\boldsymbol{r}) - \boldsymbol{F}_I(M(\boldsymbol{r})) \right|^2 \right\rangle, \tag{1.72}$$

where χ_1^2 is the ensemble-averaged sum of square errors between the force acting on a group of atoms $\boldsymbol{f}_I(\boldsymbol{r})$ that map to the coarse interaction site I, and the corresponding coarse-grained force acting on I, $\boldsymbol{F}_I(M(\boldsymbol{r}))$. The specific notation used in this section is taken from the work of Dunn et al. [83]. Eq. 1.72 essentially states that the optimal CG potential will be the one that best reproduces the average net force acting on mapped CG sites from the atomistic trajectory. For this reason, the MSCG approach is also commonly referred to as the force-matching method for bottom-up coarse-graining. If a potential is found that sets the right hand side of the equation to zero, this potential must be the MBPMF. Therefore, minimizing this expression yields the closest possible approximation to the MBPMF. A major

advantage of this variational approach is that it does not restrict the functional forms used to express the CG potential. Rather, we express the CG potential as:

$$U_R(\boldsymbol{R}) = \sum_{\zeta} \sum_{\lambda} U_\zeta(\psi_{\zeta\lambda}(\boldsymbol{R})). \tag{1.73}$$

where the first sum is over all interaction types, ζ, the second sum is over all groups of particles, λ, and $\psi_{\zeta\lambda}(\boldsymbol{R})$ is the corresponding scalar function, with CG coordinates $\boldsymbol{R}$. The force on site I is obtained by taking the gradient of the potential:

$$\boldsymbol{F}_I(\boldsymbol{R}) = \sum_{\zeta} \sum_{\lambda} \boldsymbol{F}_\zeta(\psi_{\zeta\lambda}(\boldsymbol{R}))\nabla_I \psi_{\zeta\lambda}(\boldsymbol{R}) \tag{1.74}$$

where $\boldsymbol{F}_\zeta = -dU_\zeta/dx$ and $\nabla_I = \partial/\partial \boldsymbol{R}_I$. $\boldsymbol{F}_\zeta$ can be expressed as a linear combination of basis functions $\boldsymbol{f}_{\zeta d}(x)$ with weights $\phi_{\zeta d}$. This allows the force on a site I to be rewritten as:

$$\boldsymbol{F}_\zeta(x) = \sum_{d} \phi_{\zeta d} \boldsymbol{f}_{\zeta d}(x). \tag{1.75}$$

We next define the corresponding force-field basis vectors, $\boldsymbol{G}_{I;\zeta d}$, as:

$$\boldsymbol{G}_{I;\zeta d} = \sum_{\lambda} \boldsymbol{f}_{\zeta d}(\psi_{\zeta\lambda}(\boldsymbol{R}))\nabla_I \psi_{\zeta\lambda}(\boldsymbol{R}). \tag{1.76}$$

For convenience, we replace the subscript combination ζd with D, which denotes a single ζd pair. We now rewrite Eq. 1.75 as:

$$\boldsymbol{F}_I(\boldsymbol{R}) = \sum_{D} \phi_D \boldsymbol{G}_{I;D}(\boldsymbol{R}). \tag{1.77}$$

By plugging this expression into Eq. 1.72 and then minimizing the resulting expression, the following system of linear equations is obtained:

$$\sum_{D'} \boldsymbol{G}_{DD'} \phi_{D'} = \boldsymbol{b}_D. \tag{1.78}$$

where

$$\boldsymbol{b}_D = \left\langle \frac{1}{3N} \sum_{I} \boldsymbol{f}_I(\boldsymbol{r}) \cdot \boldsymbol{G}_{I;D}(M(\boldsymbol{r})) \right\rangle, \tag{1.79}$$

and

$$\boldsymbol{G}_{DD'} = \left\langle \frac{1}{3N} \sum_{I} \boldsymbol{G}_{I;D}(M(\boldsymbol{r})) \cdot \boldsymbol{G}_{I;D'}(M(\boldsymbol{r})) \right\rangle. \tag{1.80}$$

In this equation, $\boldsymbol{G}_{DD'}$ is a symmetric matrix called the metric tensor that measures the cross-correlations between all atomistic interactions when projected onto

the force-field basis vectors defined in Eq. 1.76. $\boldsymbol{b}_D$ is a vector obtained by projecting the MBPMF of the atomistic reference onto these force field basis vectors. Solving Eq. 1.78 yields the weights $\phi_{D'}$ corresponding to the optimal CG potential that minimizes χ_1^2 given the force field basis vectors $\boldsymbol{G}_{I;D}$.

Extended Ensemble Coarse-Graining

Because a variational approach is used to find the potential that best approximates the MBPMF in the MS-CG method, it is a simple matter to extend the variational principle over multiple thermodynamic state points [40]. This collection of multiple state points is called an extended ensemble. An average over the extended ensemble is defined in the following manner

$$\langle a_\gamma(\boldsymbol{r}_\gamma)\rangle = \sum_{\gamma}^{\Gamma} p_\gamma \langle a_\gamma(\boldsymbol{r}_\gamma)\rangle_\gamma, \tag{1.81}$$

where γ denotes the specific state point, and p_γ gives the probability of being in that ensemble, set to $1/\Gamma$, where Γ is the total number of state points in the extended ensemble. Each ensemble will have its own mapping, and a corresponding MBPMF. In this case, solving the force-matching functional yields the potential that best approximates all of the MBPMFs over the extended ensemble, averaged as shown in the above equation. The derivation proceeds just as described in the previous section, while additionally taking the sums over all Γ corresponding to the different state points [40, 83].

Coarse-Graining and Pressure

The averaging over unnecessary atomistic degrees of freedom results in a smoother CG free energy landscape, which can lead to dramatic changes in the resulting thermodynamic properties. For example, it is usually the case that bottom-up methods result in CG systems with immensely large internal pressures compared to the atomistic reference. Furthermore, these CG potentials are constructed by mainly accounting for short-range repulsive interactions, while the long-range Van der Waals attractive forces are neglected, facilitating the drastic increase in pressure [84]. Das and Anderson proposed that the following functional be minimized in order to modify an existing CG potential such that the correct internal pressure, $p_{int}(\boldsymbol{r}, \boldsymbol{p}, v)$ was recovered [85].

$$\chi_2^2[U] = \langle |p_{\text{int}}(\boldsymbol{r}, \boldsymbol{p}, v) - P_{\text{int}}(M(\boldsymbol{r}), M(\boldsymbol{p}), v)|^2 \rangle \tag{1.82}$$

The CG potential now has the following form:

$$U(\boldsymbol{R}, V) = U_R(\boldsymbol{R}) + U_V(V), \tag{1.83}$$

where $U_R(\boldsymbol{R})$ is the CG interaction potential and $U_V(V)$ is the volume-dependent potential. The corresponding pressure can be written as

$$P_{\text{int}}(\boldsymbol{R}, \boldsymbol{P}, V) = P^0_{\text{int}}(\boldsymbol{R}, \boldsymbol{P}, V) + \boldsymbol{F}_V(V). \tag{1.84}$$

where

$$P^0_{\text{int}}(\boldsymbol{R}, \boldsymbol{P}, V) = \frac{2}{3V} K(\boldsymbol{P}) - \left(\frac{\partial U_R(\boldsymbol{R})}{\partial V} \right). \tag{1.85}$$

The two terms on the right hand side of Eq. 1.84 correspond to pressure contributions from the interaction potential, $P^0_{\text{int}}(\boldsymbol{R}, \boldsymbol{P}, V)$, broken into their kinetic and virial contributions (shown in Eq. 1.85), and a pressure correction term $F_V(V)$ [83, 84]. Similar to the MSCG approach, $U_V(V)$ is expressed as a sum of basis functions, $u_{Vd}(V)$:

$$U_V(V) = \sum_d \psi_d u_{Vd}(V) \tag{1.86}$$

weighted by coefficients ψ. The basis functions take the following form:

$$u_{Vd}(V) = \begin{cases} N(V/\bar{v}) & \text{for } d = 1 \\ N(V/\bar{v} - 1)^d & \text{for } d = 2 \end{cases} \tag{1.87}$$

Here, $\bar{v}$ is the average volume computed from the atomistic reference simulation. Only two basis functions are required, and their corresponding weights are related to the pressure and compressibility corrections, respectively [83]. Similar to the approach taken to solve the force-matching functional, by plugging Eq. 1.86 into the pressure matching functional and subsequently minimizing this expression yields a set of linear equations that can be solved for the coefficients ψ_1 and ψ_2. While this results in a CG system that qualitatively matches the density fluctuations of the atomistic reference, it does not ensure quantitative agreement. Dunn and Noid developed an iterative scheme to ensure quantitative agreement between CG and atomistic pressures. The steps are highly analogous to the Iterative Boltzmann approach [83]. In the first step, a CG NPT simulation is run with a trial U_V, with corresponding force $\boldsymbol{F}_V$. The difference between the CG and atomistic pressures is then used to correct $\boldsymbol{F}_V$.

$$\delta \boldsymbol{F}_V(V) = p_{\text{int}}(v) - P_{\text{int}}(V) \tag{1.88}$$

These steps are repeated until F_V converges, which, in practice, only requires two to three iterations. Successfully applying these pressure corrections guarantees that the coarse-grained system will exhibit the same thermodynamic properties as the atomistic reference.

The Martini Force Field

The Martini force field is a popular top-down coarse-grained model used for simulating a large variety of soft matter systems, with an emphasis on biomolecules [79, 86–88]. It was originally developed to model lipid membranes but has since been extended for use in protein, nucleic acid, and even non-biological soft molecules. This top-down model was optimized using many different experimental partition free energies between multiple organic liquids and water, as well as certain structural properties of lipid bilayers. The philosophical goal of the force field is to be broadly applicable to many different systems without requiring drastic reparameterization for specific state points.

The Martini force field consists of four sets of different bead types–each corresponding to different levels of chemical polarity–designated as polar, nonpolar, apolar, and charged bead types. The polar bead types represent molecules or functional groups that are greatly stabilized in the aqueous phase, while the apolar bead types correspond to chemistries highly stabilized in the organic phase. The nonpolar bead types represent molecules containing both polar and apolar character, and the charged bead types correspond to any ions or molecules with a net nonzero charge. Each bead interacts with other beads via bonded and non-bonded interactions, following the same template as discussed in Section 1.1.3. The intramolecular potentials are usually obtained via bottom-up methods such as direct Boltzmann inversion [89]. The non-bonded potentials consist of electrostatic interactions for the charged bead types as well as Lennard-Jones interactions. The Lennard-Jones σ values vary depending on the size and certain characteristics of the atomistic reference structure. Normal-sized neutral beads have a σ value of 0.47 nm, while charged beads have σ values of 0.62 nm. These beads are prescribed for mappings of 4-5 heavy atoms per bead, although there is no strict upper limit to this criterion [88]. The small-sized beads, used to represent molecular rings, have $\sigma = 0.43$ nm, and the tiny-sized beads, which were specifically developed for nucleotide modelling, have $\sigma = 0.32$ nm. The ϵ values are fitted so as to match the partitioning free energies of several alkane-water mixtures, as well as certain structural features of lipid bilayers, such as the surface area per lipid and bending modulus [88]. These ϵ values form the Martini interaction matrix and, by construction, are well correlated with the organic-aqueous partition free energies of the beads themselves. The assignment of molecules or molecular fragments to specific bead types is done by matching its hydrophobicity, quantified by the water/organic partition free energy, with that of the closest corresponding bead type. Furthermore, if the molecule/fragment contains a hydrogen bond donor, acceptor, or donor/acceptor, special nonpolar and charged bead types have been included with modified Lennard-Jones well depths to account for these interactions. Example molecules and fragments are provided in the so-called Martini Bible [79].

Automatic Martini Parameterization

Bereau and Kremer developed an automated Martini parameterization algorithm for organic small molecules, called AUTO-MARTINI [90]. This program systematically determines a Martini representation given the structure of a small organic molecule or a SMILES representation of that molecule. The algorithm proceeds in four steps. First, beads are systematically placed on the heavy atom positions of the atomistic molecule to define the set of possible mappings. In this systematic placement, beads must contain more than a single atom, beads cannot be placed on atoms bonded to each other, and all atoms in a bead must be connected to at least one other atom in the bead. In the second step, the best mappings are found by optimizing the following function:

$$\begin{aligned} E(M_N) = w_{\mathrm{nr}} N_{\mathrm{nr}} + w_{\mathrm{r}} N_{\mathrm{r}} + w_{\mathrm{BB}} \sum_{I \neq J} \exp\left(-\frac{\boldsymbol{r}_{IJ}^2}{4\sigma_{IJ}^2}\right) \\ -w_{\mathrm{aB}} \sum_i \sum_J m_i \exp\left(-\frac{\boldsymbol{r}_{iJ}^2}{2\sigma_J^2}\right) + w_{\mathrm{a}} \sum_i m_i \prod_J \theta(\boldsymbol{r}_{iJ} - \sigma_J) \end{aligned} \tag{1.89}$$

Each of the terms on the right side of the equation denote different contributions to the mapping energy, with a corresponding weight used to scale that contribution. The first term is an energy penalty for introducing a new bead. The second term is a repulsion between beads, which prevents beads from being placed too close to each other. The third term is an attraction between each bead and the atoms of the molecule that are close to the bead. The last term adds an penalty for atoms that are far from any bead. Once this optimal mapping is found, the third step is to assign Martini bead types to each bead. This requires that the atomistic molecule be broken up into fragments corresponding to each bead, done via a simple Voronoi partitioning scheme. The algorithm then assigns a bead type to each fragment based on the water/octanol partition free energy of the fragment. This is obtained by using ALOGPS, a neural network that predicts the water/octanol partition coefficient of small organic molecules with an absolute error of 0.36 kcal/mol [91, 92]. If the fragment contains hydrogen bond donor/acceptor groups and is within a certain threshold value (1.0 kcal/mol) of the Nd/Na/Nda water/octanol partition free energy, the fragment is assigned to the corresponding bead type. If the fragment contains a charged group, it is automatically assigned to the corresponding Q-type bead. The final step is to compare the overall water/octanol partition free energy of the atomistic molecule with that of the coarse-grained molecule. Atomistically, this partition free energy is obtained via ALOGPS, whereas an additivity assumption is used to determine the CG partition free energy, meaning that the CG partition free energy is assumed to be the sum of the partition free energies of each bead. If the CG partition free energy is within 50% of the atomistic partition

free energy, the CG representation is accepted. Otherwise, steps 3-4 are repeated with the next optimal mapping until the additivity check is passed.

1.1.6 Machine Learning

The term "Machine Learning" (ML) can be applied to any computational method that uses data to construct statistical models [15]. In general, a successful ML model will show improved performance as more data is used in its construction. This is where the learning aspect of ML comes into play; with increasing data, the model better "learns" the underlying structure of the data. ML methods can be broadly classified into two categories: supervised and unsupervised learning. The goal of supervised learning techniques is to take labeled data points and either learn the function that best fits the data (regression) or learn the function that best separates the data into categories (classification). Unsupervised learning, on the other hand, aims to find inherent structure in unlabeled data. For this reason, common applications for unsupervised learning techniques include the identification of both clusters within the data set and manifolds upon which the data might lay.

In the following sections we begin with a brief review of Bayes theorem. We then describe all of the clustering techniques used in the work. Next, we describe several methods for dimensionality reduction, another form of unsupervised machine learning. Because much of the data obtained in this work has high dimension, it is often necessary to visualize the data in 2-D so as to obtain a sense for the structure of the data. This is especially important for validating clustering techniques, as it allows for a visual confirmation of the clustering. We next introduce the single supervised ML technique which is used in this work, Kernel Ridge Regression. We conclude with the BASIN-HOPPING algorithm and the relative entropy, both of which are useful for understanding and navigating large, high-dimensional data sets.

Bayes Theorem

Bayes theorem is a powerful equation that is ubiquitous in the field of artificial intelligence and machine learning. This theorem predicts the conditional probability of event A occuring given event B, utilizing external knowledge about the probabilities of A and B. An intuitive derivation follows. Assume that, in the space of all possible events, Ω, A and B are overlapping subsets of Ω, and that their individual probabilities, $P(A)$ and $P(B)$ are greater than zero. It is clear that the intersection of A and B can be written in terms of both the probability of A given B, $P(A|B)$ and the probability of B given A, $P(B|A)$, as well as their

individual probabilities.

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \tag{1.90}$$

Solving this for $P(A|B)$ yields Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{1.91}$$

The $P(A|B)$ is commonly referred to as the posterior probability, the $P(B|A)$ is known as the likelihood, $P(A)$ is known as the prior, and $P(B)$ is a normalization factor [93]. In a practical context, the likelihood is usually the output of some experiment with inputs A and outputs B. Bayes theorem states that including information about the independent probabilities of A and B will augment the information gained from sampling the likelihood alone.

K-means Clustering

In the next three subsections, we introduce different unsupervised learning techniques used later on in the work. Note that, unlike the previously discussed Bayes Theorem, these are not strictly probabilistic methods. The goal of unsupervised learning is to identify and separate groups of data points that share common features based on their similarity to each other. One of the simplest and most commonly used methods in unsupervised learning is K-means clustering [94]. The K-means clustering algorithm proceeds in the following steps. The algorithm takes as input the data itself as well as the number of clusters to which the data should be assigned. Given N points in the data set, $x_0, x_1, x_2, ..., x_N$, the first K points are randomly chosen to be cluster centroids. The rest of the data is then assigned to be in clusters with the closest centroids to each point.

$$w_{ik} = \begin{cases} 1 & \text{if } k = \arg\min_j \|x_i - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases} \tag{1.92}$$

Here, w_{ik} indicates whether or not the ith data point, x_i, is included in cluster k by checking to see if the L_2 norm between x_i and the nearest cluster centroid, μ_j is minimized when $j = k$. After all of the data has been divided into clusters, new centroids are selected by averaging over all points in each cluster.

$$\mu_k = \frac{\sum_i^N w_{ik} x_i}{\sum_i^N w_{ik}} \tag{1.93}$$

The previous two steps of first assigning points to centroids and then recalculating new centroids are repeated for each new data point until the centroid positions

converge. Doing so minimizes the following cost function:

$$C = \sum_{i}^{N} \sum_{k}^{K} w_{ik} \|x_i - \mu_k\|^2, \tag{1.94}$$

which is the overall cost function of the K-means algorithm. This approach is stochastic because different initializations of centroids can lead to differences in the converged clusters. Therefore, several iterations of K-means are usually performed with different random initializations in each iteration. The best of the converged set of clusters is chosen based on how well it minimizes C, the sum of the variances within each cluster.

There are several assumptions that the K-means algorithm makes that may lead to incorrect cluster assignment. First, the number of clusters must be specified beforehand, requiring some intuitive sense of how the data should cluster. Secondly, the algorithm assumes that clusters are spherical in shape, and will not accurately assign ovular clusters or clusters with more complex geometries. Finally, even if the correct number of clusters are chosen, because K-means aims to minimize the total variance per cluster, a greater importance is given to larger clusters. Therefore, small clusters located close to larger clusters are often mistakenly identified as part of the larger cluster.

Spectral Clustering

Spectral clustering is a popular alternative to K-means because it does not require specifying the number of clusters beforehand [95]. Instead, this approach treats each point in the data set as a node in a graph, and identifies groups of nodes based on how their edges are connected. Along with the data itself, the input for this method is either the k-nearest neighbors that specify the number of edges that each node must have, or a cut-off distance within which all nodes are connected. It is also possible to take the fully-connected graph, meaning that each of the N nodes will have $N-1$ edges to every other node in the graph. We define the graph $G = (V, E)$ with a set of vertices $V = V_1, V_2, ..., V_N$ and the edges E connecting them. Each edge can be left either unweighted or weighted by some similarity measure, usually chosen as a Gaussian function applied to the distance between points. Once a network has been constructed out of the data set, the first step is to build the adjacency matrix $\boldsymbol{W}$.

$$\boldsymbol{W} = (w_{ij})_{i,j=1,2,...,N} \tag{1.95}$$

This is simply an $N \times N$ matrix with elements of 0 or w_{ij} depending on whether or not the ith and jth node are connected. The next step is to construct the degree

matrix $\boldsymbol{D}$, which is an $N \times N$ matrix with the sum of the edge weights per node on the diagonal.

$$\boldsymbol{D} = \begin{cases} d_i & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{1.96}$$

where

$$d_i = \sum_{j=1}^{N} w_{ij}. \tag{1.97}$$

Next we obtain the graph Laplacian, $\boldsymbol{L}$ by subtracting the adjacency matrix from the degree matrix, normalizing it with respect to the degree matrix.

$$\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2} \tag{1.98}$$

Performing an eigendecomposition of $\boldsymbol{L}$ reveals some interesting properties of our input graph. The number of eigenvectors with eigenvalues equal to zero is the number of connected components in the graph. In the case where the fully-connected graph is used, only the first eigenvalue will have a value of zero. Furthermore, the magnitude of the nonzero ith eigenvalue provides a measure of how densely connected the ith set of components are. The closer this magnitude is to zero, the fewer connections exist between components, and the more likely it is that these components can be separated. If the data-set is well-behaved, there will be a jump in the magnitudes of the eigenvalues that specifies the total number of clusters to be included. The eigendecomposition of the graph Laplacian is also referred to as a graph Fourier transform, with the eigenvectors forming a Fourier basis with frequencies denoted by the eigenvalues. This is one of the main advantages of graph-Laplacian based methods: they allow discrete representations to be translated into continuous representations and vice versa.

Once the number of clusters c has been chosen, the first c eigenvectors are computed and concatenated into a matrix $\boldsymbol{V}$, with each column corresponding to an eigenvector. Next, we construct the matrix $\boldsymbol{U}$, which is the same as $\boldsymbol{V}$ but with the row sums normalized to 1:

$$u_{ij} = \frac{v_{ij}}{(\sum_c v_{ic}^2)^{1/2}}. \tag{1.99}$$

Here, u_{ij} and v_{ij} denote each element of the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Each row of matrix $\boldsymbol{U}$ is now a transformed coordinate for each data point i. Furthermore, this transformation automatically separates the data such that the K-means algorithm can be used to easily identify the c clusters, without running into the issues mentioned in the previous section. However, for this approach to be considered successful, there must be a clear jump in the eigenvalue magnitudes such that the total number of clusters is easily identifiable. When a fully-connected

graph is used, performing the eigendecomposition can be highly computationally demanding based on the size of the data set. Even in this case, there may not be a significant jump in the eigenvalue magnitudes that delineates the total number of clusters. In cases where a fully-connected graph is not used as the input, the change in the eigenvalue magnitude will be heavily influenced by the choice of k-nearest neighbours or the cutoff distance. This is because, in many cases, the local density of the graph can vary drastically such that a single global parameter will be insufficient to properly identify all clusters in the data [96].

Hierarchical Density-Based Clustering

Recently, Campello et al. developed a Hierarchical Density-Based clustering algorithm called HDBSCAN [97, 98]. This algorithm overcomes the limitations of using a single length-scale or cutoff parameter by how "long-lived" they are. The main inputs that HDBSCAN takes in, aside from the input data, is the smallest number of points, s, that can be considered a cluster, and the number of nearest neighbors, k, used to define the "core distance", d_{core}, of each point. This is the minimum distance required from each data point such that the k-nearest neighbors are included within the resulting hypersphere. Given core distances for each point, the "mutual reachability distance" (MRD) between two data points x_i and x_j, separated by distance $d(x_i, x_j)$, is defined as follows:

$$\text{MRD}(x_i, x_j) = \max\{d_{\text{core}}(x_i), d_{\text{core}}(x_j), d(x_i, x_j)\}. \tag{1.100}$$

This ensures that the distance between points in dense clusters will be preserved, whereas sparse regions will be further separated from the rest of the data. Similar to the spectral clustering approach, we now construct an adjacency matrix. However, in this case, the off-diagonal elements are weighted by the MRD. This means that the weighting is different for each edge depending on the local density surrounding each point. This adjacency matrix is used to construct the minimum spanning tree, which is the smallest set of edges required to fully connect all components. Connected components are defined by a cut-off MRD. Points with MRDs above this value are discarded, and as the MRD cut-off decreases, the minimum spanning tree goes from being fully connected to fully disconnected. Plotting this progression as a function of the cut-off MRD results in a dendrogram, a useful tool for visualizing how many connected components are in the data and how the clusters subdivide as the cut-off MRD decreases. One could then choose a single cut-off MRD value as the characteristic length-scale used to separate clusters, as is done in the spectral clustering approach. However, this results in the inability to identify variable-density clusters. Instead, we use the minimum cluster size, s, as a starting point, discarding any clusters with fewer than s points. As we decrease the MRD cut-off, we see splits in the dendrogram and evaluate them based

on the size of each sub-cluster. If one of the sub-cluster has fewer than s points, than this sub-cluster is labeled the child of the other, parent cluster. However, if both sub-clusters have greater than s points, this is considered a true cluster split. For any parent-child splits, the child sub-cluster is removed from the dendrogram, whereas true cluster splits are preserved in the dendrogram. Finally we can define the stability of each cluster, c, as the following sum:

$$\text{stability} = \sum_{i \in c} \lambda_i - \lambda_{\text{birth}}. \tag{1.101}$$

In this equation, $\lambda = 1/\text{MRD}$, and i refers to each data point within a cluster. λ_{birth} is the inverse MRD at which the cluster was first formed and λ_i is the inverse MRD at which the data point i was separated from the cluster. As we decrease the MRD, at each point where a cluster breaks into sub-clusters, we calculate the stability of the initial cluster and compare to the sum of the stabilities of the sub-clusters. If the initial cluster has a higher stability, we have found the most stable cluster in the branch. However, if the sub-clusters have a higher stability, we discard the initial cluster and repeat the process with each of the sub-clusters. By continuing to evaluate the cluster stabilities at each break, we eventually find the clusters that maintain stability over the widest range of lambda with respect to their sub-clusters or super-clusters in the data set. By doing so, we identify stable clusters without the use of a single length-scale or cut-off value. Any data points not included in the stable clusters are classified as noise.

Principle Component Analysis

In the next four subsections, we introduce different dimensionality reduction techniques used later on in the work. Principle Component Analysis (PCA) is one of the most commonly used methods in dimensionality reduction. It projects the data onto the linear combinations of its dimensions that show the highest variance in the data set [99]. Given a series of datapoints $X = \{X_1, X_2, ..., X_N\}$ with y_M dimensions per datapoint, we calculate the covariance matrix, defined as:

$$K = (\text{cov}(y_m, y_n))_{m,n=1,2,...,M}, \tag{1.102}$$

where

$$\text{cov}(y_m, y_n) = \frac{\sum_i^N (y_{m,i} - \bar{y_m})(y_{n,i} - \bar{y_n})}{N-1}, \tag{1.103}$$

and $\bar{y_m}$ and $\bar{y_n}$ are the mean values over the entire data set for each dimension. By performing an eigenvalue decomposition on the covariance matrix, one obtains the eigenvectors along which the data shows the highest variance. These are known as the principle components of the data set, and their corresponding eigenvalues

denote the amount of the total variance captured along their axis. In the case where the first two principle components, meaning those with the two highest eigenvalues, account for a significant percentage of the total variance, the data may be projected onto these principle components. This generates a 2-D plot of the data that accurately reflects its structure in the high-dimensional space.

Multidimensional Scaling

Another popular dimensionality reduction technique is Multidimensional Scaling (MDS) [100]. Unlike PCA, which seeks to maximize the projected variance of the data, the goal of MDS is to obtain a low-dimensional mapping that preserves the distances between data points in the high dimension. Given high-dimensional data points $X = \{X_1, X_2, ..., X_N\}$, we wish to find the optimal low-dimensional points, $x = \{x_1, x_2, ..., x_N\}$, via the following cost function

$$C(x) = \sum_{i}^{N} \sum_{j}^{N} (R_{ij} - r_{ij})^2, \tag{1.104}$$

where R_{ij} and r_{ij} correspond to the distances between high-dimensional points X_i and X_j and low-dimensional points x_i and x_j. The procedure is similar to PCA, except it requires the eigenvalue decomposition of the Gram matrix built using the distances between data points instead of the Covariance matrix built from the data points directly. We start by centering the data matrix by subtracting the mean, as was done in Eq. 1.103.

$$X = \tilde{X} - \bar{X} = \tilde{X} \left(\boldsymbol{I} - \frac{1}{N} \mathbb{1} \right) \tag{1.105}$$

Here, $\tilde{X}$ is a matrix of concatenated column vectors with each vector corresponding to a high-dimensional data point, $\boldsymbol{I}$ is the identity matrix, and $\mathbb{1}$ is the matrix of ones. The expression in parentheses in the right-most expression above is also called the centering matrix. Next, we define the distance matrix as the adjacency matrix, $\boldsymbol{W}$, for a fully connected graph constructed from the data set, as previously defined in equation1.95, with the weights set to be the distance between points. The Gram matrix, $\boldsymbol{G}$, is obtained by taking the pairwise dot products for each row/column in the centered distance matrix with respect to every other row/column.

$$\boldsymbol{G} = \left(\boldsymbol{I} - \frac{1}{N} \mathbb{1} \right) \boldsymbol{W} \left(\boldsymbol{I} - \frac{1}{N} \mathbb{1} \right) = X^T X \tag{1.106}$$

We then express the Gram matrix both in terms of its eigendecomposition as well as in terms of the singular value decomposition of the original data.

$$\boldsymbol{G} = V_G \boldsymbol{\Lambda}_G V_G^T = V_X \boldsymbol{\Lambda}_X^2 V_X^T \tag{1.107}$$

This shows that the eigenvectors V_G and V_X are equal and the eigenvalue diagonal matrix $\mathbf{\Lambda}_G$ is equal to the square root of the eigenvalues from the spectral decomposition of the data matrix. We can now project the high dimensional points into a lower-dimensional space by using the eigenvectors corresponding to the first k significant eigenvalues in the following manner

$$x_i \approx (\sqrt{\lambda_1} v_1^i, ..., \sqrt{\lambda_k} v_k^i), \tag{1.108}$$

where λ and v correspond to the diagonal entries of $\mathbf{\Lambda}_G$ and the rows of V_G, respectively. Note that using a euclidean distance metric to calculate the Gram matrix yields a solution that is equivalent to PCA. Furthermore, if the high-dimensional data points are assumed to lie on a locally connected manifold, and the geodesic distance is used to construct the distance matrix, the ISOMAP algorithm is recovered [101]. The ability to choose the distance metric used in constructing the distance matrix makes MDS a highly robust method for representing many different types of high-dimensional data.

The Sketch-Map Algorithm

In many cases, it has been shown that a linear dimensionality reduction technique is insufficient for accurately representing the structure of high-dimensional data, especially if it lies on some nonlinear topological manifold [102, 103]. Ceriotti et al. showed that, when reducing the dimensionality of data collected from MD trajectories, linear methods like PCA would be insufficient to characterize its global structure [104, 105]. This is because the majority of the accessible phase space when running MD simulations lies in free energy basins. Therefore, one can imagine that the free energy landscape can be likened to a network of basins connected via specific transition pathways in the high-dimensional space. Ceriotti et al. have proposed a nonlinear dimensionality reduction method called SKETCH-MAP, which is meant to preserve this type of structure even when reduced to two dimensions [104]. The method uses a modified version of Eq. 1.104,

$$C(x) = \sum_{i}^{N} \sum_{j}^{N} (F(R_{ij}) - f(r_{ij}))^2, \tag{1.109}$$

Where R_{ij} and r_{ij} are the distances between points i and j in the high-dimensional and low-dimensional spaces, respectively. F and f are sigmoid functions that take the following forms:

$$F(R) = 1 - (1 + (2^{A/B} - 1)(R/\sigma)^A)^{-B/A} \tag{1.110}$$

$$f(r) = 1 - (1 + (2^{a/b} - 1)(r/\sigma)^a)^{-b/a}, \tag{1.111}$$

where σ, A, B, a, and b, are all fitting parameters used to specify the sigmoid functions. Eq. 1.110 transforms the distances of the high-dimensional space by applying the switching function, F, and searches for an embedding that preserves these nonlinearly transformed distances in the lower dimension. The sigmoid function effectively selects an "interesting" subset of the data to be highlighted based on the histogram of pairwise distances calculated in the high-dimensional space. Distances that are either very close or very far with respect to σ are assigned a distance close to zero or one after the sigmoid function is applied. However, distances in the vicinity of σ will be preserved and will primarily dictate the structure of the lower-dimensional map. Ceriotti et al. posited that, when examining MD trajectory data, the small pairwise distances correspond to thermal fluctuations within a free energy basin, while the large pairwise distances are uniformly distributed in the high-dimensional space, corresponding to poor sampling of transitions between far basins [104]. Therefore, by selecting a sigma value that highlights the intermediate distances found in MD trajectory data, the SKETCH-MAP approach ensures that the essential structure of the high-dimensional landscape is preserved for well-sampled transitions between neighboring free energy basins. However, in many cases, multiple key length-scales may be present that are well represented in the data, and projecting this data using a single sigmoid function will be insufficient to fully capture the high-dimensional structure.

Uniform Manifold Approximation and Projection

Recently, McInnes et al. has developed a nonlinear dimensionality reduction technique that assumes the input data lies on a high-dimensional topological manifold and aims to project the data onto a similarly structured low-dimensional manifold [106]. The algorithm is called Uniform Manifold Approximation and Projection (UMAP), and as implied by the name, it assumes that the data is uniformly distributed on a Riemannian manifold which has local connectivity and a localized distance metric. The main input parameter required for UMAP, other than the input data, is the number of k-nearest neighbors to be taken for each data point. The procedure is divided into two steps. The first step is to express the data as a graph and construct an adjacency matrix similar to the one used in spectral clustering. In order to determine the weights that should be assigned to each edge in the graph, the UMAP algorithm assumes that each point can be expressed as part of a localized fuzzy simplicial set with its nearest neighbor. For each data point, we first determine the nearest neighbor distance, p_i, and then use the k-nearest neighbor distances to find the localized length-scale σ_i used to normalize p_i. σ_i is

determined by iteratively solving the following expression:

$$\sum_{j=1}^{k} \exp\left(\frac{-\max\left(0, d(X_i, X_j) - p_i\right)}{\sigma_i}\right) = \log_2(k), \tag{1.112}$$

where $d(X_i, X_j)$ is the distance between high-dimensional points X_i and X_j by any metric of choice. Next, a weight-directed graph $\bar{G}$ can be constructed with vertices V, edges E, and weight function w. The vertices correspond to the data points, the edges denote the connectivity for each vertex up to the k-nearest neighbors, and the weight function is defined as follows:

$$w((X_i, X_j)) = \exp\left(\frac{-\max\left(0, d(X_i, X_j) - p_i\right)}{\sigma_i}\right). \tag{1.113}$$

Given the directed adjacency matrix corresponding to $\bar{G}$, $\boldsymbol{A}$, we construct an undirected adjacency matrix $\boldsymbol{B}$ via the following expression:

$$\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{A}^T - \boldsymbol{A} \circ \boldsymbol{A}^T. \tag{1.114}$$

This undirected adjacency matrix now corresponds to an undirected graph G that represents a manifold of locally connected fuzzy simplicial sets in the high dimensional space. The next step is to find a low-dimensional approximation for this graph. In practice, this is done by using a force-directed graph layout approach. The low-dimensional positions are initialized by performing a spectral embedding (i.e., transforming the data using the graph Laplacian shown in 1.98). Next, attractive and repulsive forces are defined that act on edges and vertices, respectively. The attractive force is defined as

$$F_{\text{attr}} = \frac{-2ab\|x_i - x_j\|_2^{2(b-1)}}{1 + \|x_i - x_j\|_2^2} w((X_i, X_j))(x_i - x_j). \tag{1.115}$$

where x_i and x_j are the low-dimensional positions of X_i and X_j. a and b are hyperparameters that must be optimized. The repulsive force is given by

$$F_{\text{rep}} = \frac{b}{(\epsilon + \|x_i - x_j\|_2^2)(1 + \|x_i - x_j\|_2^2)} w((X_i, X_j))(x_i - x_j). \tag{1.116}$$

where ϵ is a negligible constant that ensures that division by zero does not occur. The optimization of the low-dimensional positions occurs iteratively via stochastic gradient descent, with a nonlinear least-squares fitting applied at each iteration to solve for the hyperparameters a and b which are then used in the force-directed graph layout approach. It has been shown that minimizing these force functions and thus optimizing the low-dimensional positions is equivalent to minimizing

the cross-entropy between the high-dimensional and low-dimensional manifolds. Unlike the methods described in the previous sections, UMAP applies a series of nonlinear transformations to the high-dimensional distances in a localized manner, meaning that the relative densities of the points in the high-dimensional space is not quantitatively preserved. However, by taking a manifold learning approach, UMAP better preserves the global structure of the high-dimensional data set.

Kernel Ridge Regression

Kernel Ridge Regression (KRR) is a supervised ML method that is often favored over deep learning methods because it requires comparatively less data to achieve convergence of the model [107, 108]. Conceptually, KRR can be understood as linear ridge regression with the inclusion of the kernel trick, which allows nonlinear problems to be approximated as high-dimensional linear problems by mapping the input to a high-dimensional implicit feature space [109]. We first discuss both of these aspects separately before combining them to formally define KRR.

Linear ridge regression can be thought of as applying Tikhonov regularization to linear regression. Linear regression assumes that the data can be fit to a line by minimizing the least-squares error between the data and the line. Given a line of the form

$$Y = \beta^T X + \epsilon, \tag{1.117}$$

where Y is the predicted value given inputs $X = X_1, X_2, ..., X_N$, with coefficients β and irreducible error values ϵ. When performing linear least squares regression, the next step is to minimize the following expression

$$C = \sum_i^N (Y_i - \beta^T X_i)^2. \tag{1.118}$$

However, if the number of dimensions used to represent X is large, there is a significant danger of over-fitting the data. This over-fitting can also be expressed in terms of the variance in β, which should be restricted to prevent this issue. The solution, known as Tikhonov regularization, is to add a term to the cost function that penalizes the L_2 norm of β. This has the added benefit of providing numerical stability when inverting the matrix equation to solve for β

$$C = \sum_i^N (Y_i - \beta^T X_i)^2 + \lambda \|\beta\|^2. \tag{1.119}$$

This effectively restricts the domain of β to fall within the hypersphere whose radius is set to λ, causing a significant decrease in the variance of the model

weights while limiting the degree to which the model is biased by outliers in the data.

This works well for data that has a highly linear structure, but this may not be true for many data sets. If we generalize Eq. 1.119 such that $\beta^T X$ is replaced by some arbitrary function $f(X)$ and substitute this expression in to the cost function we get

$$C = \sum_i^N (Y_i - f(X_i))^2 + \lambda \|f\|^2. \tag{1.120}$$

We can then express this as a constrained optimization problem to be solved using Lagrange multipliers. We do this by introducing the constraint $\xi = Y - \langle f, X_i \rangle$ (the angular brackets denote the inner product) to get

$$\mathcal{L} = \sum_i \xi^2 \tag{1.121}$$

such that

$$\|f\| \leq B. \tag{1.122}$$

The Lagrangian can then be written as

$$\mathcal{L} = \sum_i \xi^2 + \sum_i \beta_i [Y_i - < f, X_i > -\xi_i] + \lambda(\|f\|^2 - B^2) \tag{1.123}$$

Setting the partial derivatives with respect to f and ξ equal to zero gives the following relations

$$2\xi_i = \beta_i, 2\lambda f = \sum_i \beta_i X_i \tag{1.124}$$

and

$$2\lambda f = \sum_i \beta_i X_i. \tag{1.125}$$

Substituting this back into the Lagrangian and simplifying gives the dual Lagrangian,

$$\mathcal{L}_D = \sum_i (-\frac{1}{4}\beta_i^2 + \beta_i Y_i) - \frac{1}{4\lambda} \sum_{ij} (\beta_i \beta_j \boldsymbol{K}_{ij}) - \lambda B^2 \tag{1.126}$$

where $\boldsymbol{K}_{ij}$ is the Kernel defined as

$$\boldsymbol{K}_{ij} = \langle X_i, X_j \rangle. \tag{1.127}$$

Note that the kernel $\boldsymbol{K}$ can be any function of the similarity between two inputs. Common choices are Laplace or Gaussian kernels, which will introduce parameters σ or σ^2, respectively. The similarity between inputs depends heavily on the choice

of distance metric (i.e., L_1 vs. L_2 norm) as well as the mathematical object used to represent the input data, known as the representation or fingerprint. The kernel matrix maps the input into a high-dimensional (potentially infinitely so), implicit feature space. This property, known as the Kernel Trick, explains how KRR solves nonlinear problems by approximating them as a high-dimensional linear problem. We next define $\alpha = \beta_i/2\lambda$ and rewrite this expression as

$$\mathcal{L}_D = -\lambda^2 \sum_i \alpha_i^2 + 2\lambda \sum_i \alpha_i X_i - \lambda \sum_{ij} \alpha_i \alpha_j \boldsymbol{K}_{ij} - \lambda B^2 \tag{1.128}$$

Optimizing this with respect to α gives the solution for the dual Lagrangian,

$$\alpha_i = (\boldsymbol{K}(X_i, X_j) + \lambda \boldsymbol{I})^{-1} X_i, \tag{1.129}$$

and the final solution is then:

$$f(X) = Y(\boldsymbol{K}(X_i, X_j) + \lambda \boldsymbol{I})^{-1} \boldsymbol{K}'(X_i', X_j). \tag{1.130}$$

In the above equations, we explicitly show the dependence of $\boldsymbol{K}$ on the known data points X, also referred to as the training data set. Given this set of input data, with corresponding output Y, the coefficients α are obtained by using 1.129. $\boldsymbol{K'}$ is a separate kernel matrix constructed using the similarity between any new inputs, X', and the training data. Outputs, $f(X)$, for this new data (referred to as the test data set) can be predicted by solving Eq. 1.130. Usually, in order to check the robustness of the model, the data is separated into training and test categories, of which the latter is used to check the accuracy of the predictions made after training the model using the training data set. In order to remove the possibility of bias when choosing the training and test set, a procedure known as cross-validation is used. This involves averaging over a statistically significant number of training/testing iterations with different training sets in each iteration.

Y can also be reframed as a Bayesian posterior distribution with a prior expression that accounts for the covariance matrix $\boldsymbol{K}$ between all training inputs as well as the covariance $\boldsymbol{K'}$ between the training set and the test input [109]. If one assumes that the $f(X)$ in Eq. 1.130 is Gaussian in each dimension, solving for the mean of this posterior distribution, $\bar{f}(X'|Y)$ yields the following expression:

$$\bar{f}(X'|Y) = Y(\boldsymbol{K}(X_i, X_j) + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{K}'(X_i', X_j). \tag{1.131}$$

This approach, known as Gaussian Process Regression, results in a solution exactly equal to that of KRR if the kernel matrices are equal and $\lambda = \sigma^2$. The variance of the posterior distribution is known as the predictive variance, and it relates the variance of the error, which is also assumed to be Gaussian, to the covariance matrices $\boldsymbol{K'}$ and $\boldsymbol{K}$ between test-training and training-training data

points, respectively. The predictive variance is a useful metric for calculating confidence intervals on test set predictions. Rigorously, the hyperparameters σ_K and σ_ϵ can be obtained by maximizing the marginal likelihood functions $p(\sigma_K|X,Y)$ and $p(\sigma_\epsilon|X,Y)$, but this is computationally expensive. In practice, a grid search optimization is usually performed for both hyperparameters, with the accuracy of the KRR model minimized at each iteration.

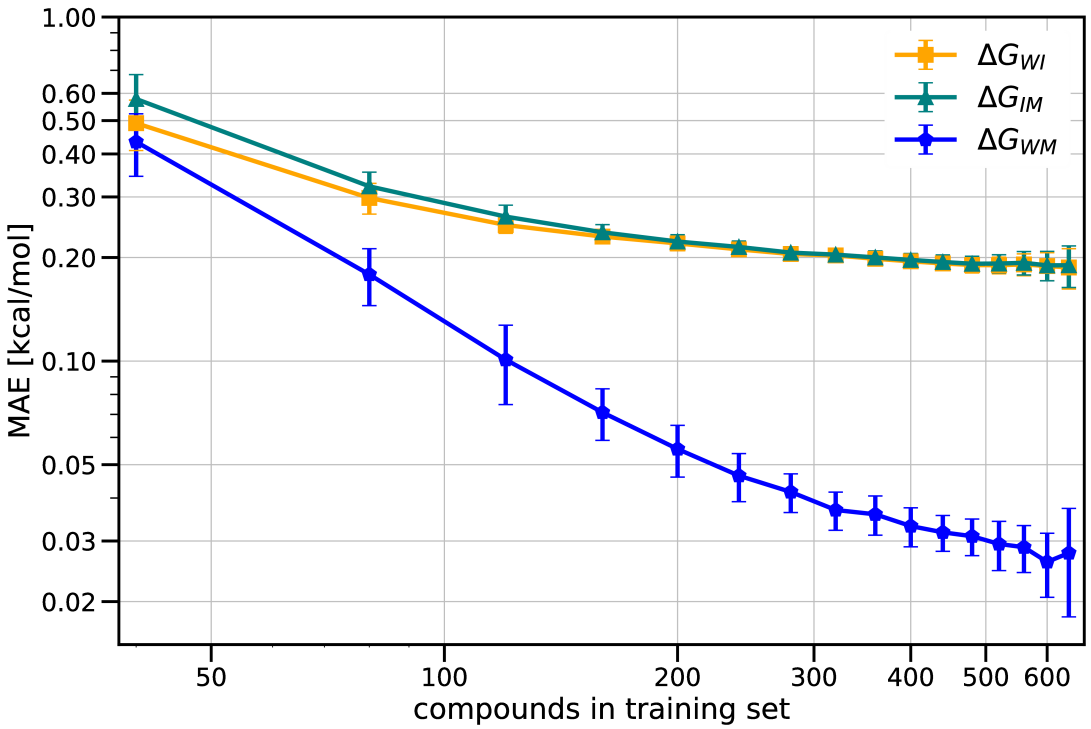

Figure 1.6: Three learning curves for the prediction of the three transfer free energies shown in Fig. 1.4 for coarse-grained Martini compounds consisting of three beads with a linear topology. The error bars correspond to the standard deviation of the mean absolute error values obtained during cross validation. Adapted with permission from Hoffmann et al. [110].

Two major questions remain: (1) how does one choose a representation, and (2) how does one evaluate the quality of the KRR model? In general, a good representation is one that encodes and highlights the information most relevant to the problem from the input data. The necessary attributes of representations used in machine learning of chemical properties, as well as examples of these representations, are given in Section 1.1.7. We focus now on the second question. A common strategy for evaluating the quality of KRR models is to plot the cross-

validated error as a function of the size of the training set. This type of plot is known as a "learning curve," and an example is shown in Fig. 1.6 [108, 111]. According to statistical learning theory, any ML model will have a corresponding learning curve that maintains a power law decay as the size and uniqueness of the training set increases in the input space [112]. However, if the input space is poorly defined, or the data sampled is highly degenerate, the model will reach a data saturation point past which the error no longer decreases. Therefore, it is often the case that there is a hard limit to the amount of learning that can be accomplished with a given data set, leading to learning curves that saturate, as seen for two of the curves shown in Fig. 1.6. ML models which show the steepest decay when their corresponding learning curve is plotted are desirable, as they enable a high level of accuracy with fewer data points. As such, in addition to providing an informative picture of the quality of a KRR model, these learning curves are useful for optimizing hyperparameters, comparing different representations, and evaluating model transferability across data sets.

The basin-hopping Algorithm

BASIN-HOPPING is a numerical approach to finding global minima in landscapes that have many dimensions [113]. As the name suggests, the algorithm follows the cycle of randomly jumping to a new point in the landscape, performing a local minimization, and either accepting or rejecting the new minimum. The key is to ensure that the random jumps have the correct magnitude: if they are too short, only the local minimum will be explored. The algorithm was first developed by Wales et al. in order to investigate the energy landscape of Lennard Jones clusters, but can be adapted to many different optimization problems [114]. Our implementation of this algorithm is discussed in Chapter 3.

The Relative Entropy

Similar to the thermodynamic entropy, which is related to the number of microstates accessible to a molecular system, the information entropy, H of any probability distribution for a random variable, X with possible outcomes, $(x_1, x_2, ..., x_N)$, can be defined as follows:

$$H(X) = -\sum_{i}^{N} P(x_i) \log_2 P(x_i), \tag{1.132}$$

where $P(x_i)$ is the probability of obtaining outcome x_i [115]. This suggests a useful method to compare two different distributions based on their information content. This relative entropy can be expressed as a Kullback-Leibler divergence (KLD)

between probability distributions $P(X)$ and $Q(X)$ in the following expression:

$$D_{KL}(P||Q) = \sum_{i=1}^{N} P(x_i) \ln\left(\frac{P(x_i)}{Q(x_i)}\right). \tag{1.133}$$

The KLD is equal to 0 when $P(X)$ and $Q(X)$ are identical, corresponding to zero information loss [115]. There is no upper bound for the KLD, as one can imagine scaling $P(X)$ infinitely while keeping the same $Q(X)$. However, problems arise when comparing P and Q if P has a finite value at some x_i while Q is 0 at that same x_i, making $\ln(P(x_i)/Q(x_i))$ undefined. A convenient work-around is to instead use the Jensen-Shannon Divergence (JSD), which is defined as the sum of two KLDs of $P(X)$ and $Q(X)$, each taken with respect to the average of the two distributions $M(x)$:

$$D_{JS} = \frac{1}{2} D_{KL}\left(P(X)||M(X)\right) + \frac{1}{2} D_{KL}\left(Q(X)||M(X)\right), \tag{1.134}$$

where

$$M(X) = \frac{1}{2}(P(X) + Q(X)). \tag{1.135}$$

The JSD is symmetric with respect to both distributions P and Q, and now has an upper bound, which is 1 if the base-2 log is used and ln(2) if the natural log is used [116]. The relative entropy in the form of a JSD is also highly useful as a means to quickly test the validity of results when generating massive amounts of data, as will be discussed further in Chapter 4.

1.1.7 Molecular Representations

The means by which molecules are mathematically represented is a critically important factor when exploring CCS. A good molecular representation encodes chemical information such that projecting CCS onto that representation naturally leads to a correlation with a target property. Therefore, a molecular representation should ideally encode properties of the molecule which pertain to the physics of the problem under consideration. Examples may include the number and type of atoms as well as their positions, the topology of the molecule, the number of hydrogen-bonding sites, etc. In cases where the geometry of the compound is relevant, invariances with respect to geometric translations and rotations should be preserved, unless a specific reference frame can be universally applied (i.e., distance to the binding pocket of an enzyme). Similarly, the representation should be invariant to permutations in the ordering of the atoms that make up a given molecule. Finally, a good representation should be bijective, smooth, and

continuous, allowing for straight-forward empirical fitting of structure–property relationships that span CCS [117, 118].

In the following subsections, we introduce the three molecular representations that will be referenced in the rest of this work: the Coulomb Matrix, the Spectrum of London Axilrod-Teller-Muto potential, and the Smooth Overlap of Atomic Positions Kernel [118–120]. These representations were recently developed for the prediction of quantum-mechanical properties using machine-learning techniques, and all take the atomic numbers and internal geometries of molecules as an input. A cartoon schematic of each representation is shown in Fig. 1.7. For each representation, the molecules are first converted from an input SMILE string to a 3-D structure that is then energy minimized using the UFF force field in a molecular mechanics based optimization scheme with the RDKIT package. The aforementioned atomic numbers and internal geometries are then taken from this 3-D structure and used to create the representation.

The Coulomb Matrix

Developed by Rupp et al., the Coulomb Matrix molecular representation consists of a matrix whose off-diagonal elements are the pairwise coulombic interactions which are calculated using the nuclear charge of each element, with diagonal elements obtained by fitting a polynomial to the atomization energies of individual atoms as a function of their nuclear charge [119]. Given an input molecule with atomic coordinates R_i and corresponding nuclear charges Z_i, the Coulomb Matrix, $\mathbf{C}$, is a symmetric matrix with the number of rows/columns equal to the number of atoms in the molecule, and whose elements $\mathbf{C}_{ij}$ are defined as

$$\mathbf{C}_{ij} = \begin{cases} 0.5 Z_i^{2.4} & \text{for } i = j \\ \frac{Z_i Z_j}{|R_i - R_j|} & \text{for } i \neq j \end{cases} \tag{1.136}$$

The coulomb matrix encodes all of the atom types and their relative distances to each other, ensuring a bijective mapping of chemical compounds. As it was originally designed for prediction of quantum-mechanical properties, the rationale behind this descriptor is to include the same input information as required by the ab initio simulations which are normally used to compute these target properties. Because the pairwise distances are taken irrespective of any reference frame, the representation is both translationally and rotationally invariant. Since the input features are continuous variables, the representation is also continuous. However, it is not permutationally invariant, as changing the order in which atoms are input will result in different Coulomb Matrices. There are many suggested schemes for adding permutational invariance to the Coulomb Matrix, which include sorting, diagonalizing, or using an ensemble of randomly sorted matrices. In this work, we

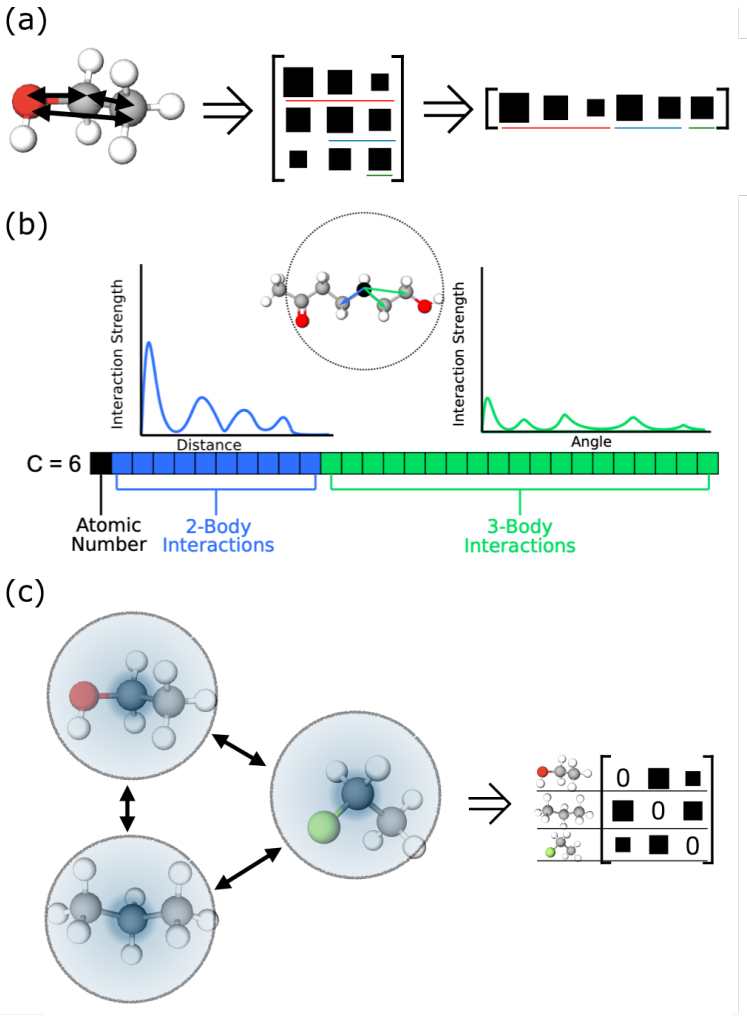

Figure 1.7: Cartoon schematics of the molecular representations used in this work. (a) For each heavy atom in the molecule, the Coulomb Matrix encodes all of the pairwise distances in terms of their Coulombic interaction, with the diagonal term determined based solely on the heavy atom type. The upper triangular of this matrix is taken as the final output [119]. (b) The atomic version of the SLATM vector for the carbon atom (colored black) consists of a one-body term concatenated together with two-body and three-body spectra [121]. (c) The SOAP kernel finds the overlap between local environments for different molecules by modelling the local environment as a sum of Gaussian functions placed at each atom's position. For clarity, only a single Gaussian is shown for each molecule. The output is a similarity matrix with each row/column of the matrix set as the representation for each molecule [118].

sort the Coulomb Matrix by the distance of each atom to the center of mass of the whole molecule, with the atom closest to the center of mass always representing the first row/column of the matrix. Ensuring permutational invariance also results in making the representation smooth with respect to nuclear charge and position. Additionally, in this work, we only include the heavy (non-Hydrogen) atoms of the molecule when constructing the representation. Furthermore, since the matrix is symmetric, it is only necessary to store the upper triangular of the matrix as a sequenced vector in practice. Fig. 1.7a shows the Coulomb Matrix representation of an ethanol molecule as it would be implemented in this work.

The Spectrum of London Axilrod-Teller-Muto vector

The Spectrum of London Axilrod-Teller-Muto (SLATM) vector describes a molecule as a sum of atomic environments that encode the 1-body, 2-body, and 3-body interactions within a cut-off distance [120, 121]. Fig. 1.7b shows a cartoon schematic of an atomic SLATM (aSLATM) vector. For each atom, its corresponding SLATM vector consists of Z_i, the elemental atomic number (1-body), a spectrum of 2-body London interactions convoluted with a gaussian (2-body), and a spectrum of 3-body Axilrod-Teller-Muto interactions also convoluted with a gaussian (3-body). The two-body spectrum is computed over the distance, r, which ranges from zero to a cut-off value with a specified step-size, using the following expression:

$$\text{aSLATM}_{i,2-\text{body}} = \frac{1}{2} Z_i \sum_{i \neq j} Z_j \delta(r - R_{ij}) g(r) \tag{1.137}$$

where R_{ij} is the distance between atoms i and j, $\delta(x)$ is a normalized Gaussian function,

$$\delta(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2}, \tag{1.138}$$

and $g(r)$ is used to scale the distance, in this instance corresponding to the London interaction,

$$g(r) = \frac{1}{r^6}. \tag{1.139}$$

Similarly, the three-body spectrum is computed over the angle, θ, via the following expression:

$$\text{aSLATM}_{i,3-\text{body}} = \frac{1}{3} Z_i \sum_{i \neq j \neq k} Z_j Z_k \delta(\theta - \theta_{ijk}) h(\theta, R_{ij}, R_{ik}). \tag{1.140}$$

Here, θ_{ijk} is the angle between pairwise-distance vectors R_{ij} and R_{ik}. $h(\theta, R_{ij}, R_{ik})$ is the 3-body Axilrod-Teller-Muto potential, defined as

$$h(\theta, R_{ij}, R_{ik}) = \frac{1 + \cos\theta \cos\theta_{jki} \cos\theta_{kij}}{(R_{ij} R_{ik} R_{kj})^3}. \tag{1.141}$$

As seen in Fig. 1.7b, the aSLATM vector is a concatenation of the 2-body and 3-body spectra with the atomic number of the atom. The molecular SLATM vector is constructed by first taking the sum of all aSLATM vectors of the same atom type. Each of the summed SLATM vectors is then concatenated by interaction type. For example, the 1-body component of the SLATM vector corresponding to ethanol would be $[8.0, 12.0]$, corresponding to one oxygen atom and two carbon atoms. When comparing multiple molecules with many different atom types, a reference SLATM vector is first constructed by determining all possible unique many-body interaction types. If a molecule does not have some of these interaction types (for example, in ethanol, there will be no 2-body interaction between Carbon and Nitrogen), the vector is filled with zeros. Similar to the Coulomb Matrix, this representation uses only the internal geometry of the molecule, making it translationally and rotationally invariant. Unlike the Coulomb Matrix, however, making the SLATM vector the sum of the constituent aSLATM vectors ensures permutational invariance when comparing molecules. Convoluting the interactions with Gaussian functions results in a continuous and differentiable metric. Furthermore, the inclusion of the 3-body spectrum via the SLATM vector was shown to vastly improve the performance of statistical models to predict quantum-mechanical properties as opposed to restricting the representation to pairwise interactions only, as is the case for the Coulomb Matrix [120].

The Smooth Overlap of Atomic Positions Kernel

The final representation used in this work is the Smooth Overlap of Atomic Positions (SOAP) Kernel developed by Bartók et al. [118]. Similarly to the SLATM vector, each atom is represented by its local environment. While the SLATM vector explicitly decomposes this local environment into 2-body and 3-body interactions acting on the atom, this representation assumes that an atomic environment is represented as a local density surrounding each atom. The atomic neighbor density function is defined as follows:

$$\rho_N(r) = \sum_{i \in N} \exp\left(-\frac{|r - r_i|^2}{2\sigma^2}\right). \tag{1.142}$$

The atomic density ρ_N of each local environment, N, is written as a sum of Gaussian functions placed on all atomic positions within a defined cutoff distance. The SOAP kernel is obtained by integrating the overlap between these local density functions over the space of all possible 3-D rotations, $\hat{\mathbf{R}}$ [118, 122].

$$\boldsymbol{k}(N, N') = \int \mathrm{d}\hat{\mathbf{R}} \left| \int \mathrm{d}r \rho_N(r) \rho_{N'}(\hat{\mathbf{R}}r) \right|^2 \tag{1.143}$$

The kernel is usually normalized such that the similarity of any environment with itself is set to 1.

$$\boldsymbol{K}(N, N') = \frac{\boldsymbol{k}(N, N')}{\sqrt{\boldsymbol{k}(N, N)\boldsymbol{k}(N', N')}} \tag{1.144}$$

The integral shown in Eq. 1.143 can be evaluated analytically by first expanding ρ_N in a basis of spherical harmonics and orthogonal radial basis functions. Doing so enables the right hand side of Eq. 1.143 to be written as the dot product of unit-length vectors $\hat{\boldsymbol{\rho}_N}$ and $\hat{\boldsymbol{\rho}_{N'}}$, which consist of elements of the rotationally invariant power spectra corresponding to the expanded basis.

$$\boldsymbol{K}(N, N') = \hat{\boldsymbol{\rho}_N} \cdot \hat{\boldsymbol{\rho}_{N'}} \tag{1.145}$$

This essentially means that the dot product between two local environments yields the degree of overlap between them. The corresponding distance between two local environments, $d(N, N')$ can then be defined as:

$$d(N, N') = \sqrt{2 - 2\hat{\boldsymbol{\rho}_N} \cdot \hat{\boldsymbol{\rho}_{N'}}} \tag{1.146}$$

We now have a means of calculating the similarity between specific atomic neighbor density functions, and in order to extend the representation to calculate molecular similarity, we construct a covariance matrix $\boldsymbol{C}_{ij}(A, B)$ between all neighbor densities i and j making up each molecule A and B.

$$\boldsymbol{C}_{ij}(A, B) = \boldsymbol{K}(N_i^A, N_j^B) \tag{1.147}$$

Similar to the Coulomb Matrix representation, we obtain a matrix comparing atomic properties (the local environment around an atom for SOAP versus the Coulombic interaction between an atom and all of its neighbors for the Coulomb Matrix), but the representation is still not permutationally invariant. Following the work of De et al., we apply a global kernel, $\hat{K}(A, B)$, that identifies the best overall match between the two molecules by maximizing their covariance.

$$\hat{K}(A, B) = \frac{1}{N} \max_{\pi} \sum_i \boldsymbol{C}_{i\pi_i}(A, B) \tag{1.148}$$

Furthermore, we use a reference structure that is constructed using the minimum number of atoms and atom types needed for all of the molecules in the data set to be reproduced. In addition to ensuring permutational invariance, this also results in a smooth landscape in the representation space.

So far, we have defined the SOAP kernel assuming only a single atom type. In addition to the previously defined SOAP kernel, we again follow the work of De et al. and apply an alchemical kernel $\boldsymbol{\kappa}_{\alpha\beta}$ between differing atom types α and β [122], yielding the alchemical SOAP (ASOAP) kernel:

$$\begin{aligned}\tilde{K}(N, N') &= \int \mathrm{d}\hat{\mathbf{R}} \left| \int \mathrm{d}r \sum_{\alpha\alpha'} \rho_N^{\alpha}(r) \rho_{N'}^{\alpha'}(\hat{\mathbf{R}}r) \right|^2 \\ &= \sum_{\alpha\beta\alpha'\beta'} \boldsymbol{\rho}_{\alpha\beta}(N) \cdot \boldsymbol{\rho}_{\alpha'\beta'}(N') \boldsymbol{\kappa}_{\alpha\alpha'} \boldsymbol{\kappa}_{\beta\beta'}.\end{aligned} \tag{1.149}$$

This allows us to define a measure of similarity between different atom types, essentially scaling the degree to which the presence of other atom types is seen in the neighbor density function. In this case, we scale the atomic similarity by calculating the difference between the atoms' electronegativity, E_α [122]:

$$\boldsymbol{\kappa}_{\alpha\beta} = \exp\left(\frac{-(E_\alpha - E_\beta)^2}{2}\right). \tag{1.150}$$

Unlike the previous two representations discussed above, the SOAP kernel does not take only a single molecule's geometry and atom types as input. Rather it computes the similarity between molecules across the full set of molecules. The dimensionality of a molecule represented via the SOAP kernel is therefore equal to the number of compounds being compared to each other. Molecule A is considered to be the row/column of the SOAP similarity matrix that gives the dot-product distance defined in Eq. 1.146 between A and every other molecule in the data set.

Having laid the theoretical groundwork for all of the methods used in this work, the next chapters describe how these methods are used to investigate the effect of coarse-graining on CCS.

2 The High-Throughput Coarse-Grained Simulation Method

Several figures and sections of this chapter have been published as sections in three separate research articles, which are listed below. These sections are reproduced here with kind permission from the other authors and the journals which published these articles.

Roberto Menichetti, Kiran H. Kanekal, Kurt Kremer, and Tristan Bereau
In silico screening of drug-membrane thermodynamics reveals linear relations between bulk partitioning and the potential of mean force
The Journal of Chemical Physics 147(12):125101, 2017.
DOI: 10.1063/1.4987012

Roberto Menichetti, Kiran H. Kanekal, and Tristan Bereau
Drug–membrane permeability across chemical space
ACS Central Science 5(2):290, 2019.
DOI: 10.1021/acscentsci.8b00718

Christian Hoffmann, Roberto Menichetti, Kiran H. Kanekal, and Tristan Bereau
Controlled exploration of chemical space by machine learning of coarse-grained representations
Physical Review E 100(3):033302, 2019.
DOI: 10.1103/PhysRevE.100.033302

Additionally, for each of the sections taken from one of the above articles, I include a single * symbol if I was the primary contributor for that specific section and a double ** symbol if I was not the primary contributor for that specific section. If no symbol is added, the section was written by me specifically for this thesis.

2.1 Introduction

Pharmacokinetic profiles are used in the pharmaceutical industry to identify the extent to which different drug-uptake mechanisms are activated in the human body when administering a drug [123]. One critical feature of these profiles is the passive membrane permeability, which denotes the flux for drug molecules passing through a lipid-bilayer membrane without relying on the active transport mechanisms of a cell [124]. Especially when considering small organic molecules as drugs, this passive permeability may compete significantly with active methods for cellular uptake [125]. Overlooking this property can result in unpredictable side effects, since the drug molecule could be acting in unintended regions of the cell at unknown concentrations. Consequently, the passive membrane permeability is a required component in any pharmacokinetic profile.

It then stands to reason that a structure–property relationship linking chemical structure to passive permeability would be extremely helpful as a means to quickly filter results in a high-throughput search for new drug molecules. In this context, a structure–property relationship refers to a small set of molecular descriptors onto which the chemical compound space (CCS) is projected such that an empirical relationship can be determined that maps the chemical structure to the desired property[126]. However, since the size of CCS has been estimated to be on the order of 10^{60} for small drug-like molecules, ensuring that the desired structure–property relationship is flexible enough to enable prediction for a large variety of compounds is a daunting task [4]. In order to prevent model bias towards specific chemistries, the construction of a sufficiently large and varied database of chemical compounds and their corresponding permeabilities is required. Furthermore, a good structure–property relationship enables both direct and inverse molecular design, such that a property can be predicted given a specific input chemistry and a chemical structure(s) can be predicted given a desired property value [5, 6].

Several methods have been developed to calculate the passive permeability of small molecules passing through a lipid-bilayer membrane, usually expressed as the log of the permeability coefficient, or $\log P$. Experimental methods, like the Caco-2 or PAMPA assays, involve measuring concentration gradients of the drug molecules using *in vitro* test systems analogous to the *in vivo* case [127, 128]. Computational approaches to calculating $\log P$ usually require simulating a drug molecule in the

lipid-bilayer membrane environment using molecular dynamics (MD) techniques [129, 130]. From the resulting simulation trajectory, the permeability coefficient can be calculated via the inhomogeneous solubility-diffusion model [56, 131, 132]. This model expresses the permeability coefficient, P, in terms of the potential of mean force (PMF), $G(z)$, and the local diffusivity, $D(z)$, where z is taken as the direction normal to the lipid-bilayer midplane

$$\frac{1}{P} = \int_{z_1}^{z_2} \mathrm{d}z \frac{\exp(\beta G(z))}{D(z)}, \tag{2.1}$$

where $\beta = 1/k_\mathrm{B}T$ [56]. The PMF provides a measure of how statistically likely a molecule will partition to a specific region of the membrane environment, and is heavily influenced by the hydrophobicity of the drug molecule. Note that the PMF can not be obtained using experimental methods and can only be resolved via simulation. The diffusivity quantifies the degree to which the concentration gradient of the drug molecule impacts its flux across the membrane. Both of these quantities can be obtained from MD simulation trajectories, usually computed with the help of enhanced-sampling techniques such as umbrella sampling. Several studies have been conducted that utilize this approach to calculate $\log P$ for small sets of amino acids or drug compounds [26, 133–135].

Both computational and experimental approaches to calculating $\log P$ mentioned previously are too inefficient to enable the construction of structure–property relationships that span CCS. In the experimental case, the limiting factor is the cost of synthesizing new molecules for testing. While the computational approach does not suffer the same costs of synthesizing new molecules as its experimental counterpart, the high computational cost (10^5 CPU hours needed per compound) prevents this route to quickly generating the desired structure–property relationship [26].

One tool commonly used as an alternative to fully atomistic simulations are coarse-grained (CG) molecular dynamics simulations using models which retain the essential physics of the simulated system [34, 136]. A CG representation is formed by assigning groups of atoms in a molecule to a single CG particle, also known as a bead. The interaction potentials of the CG beads are parameterized so as to reproduce the desired phenomena observed either from a higher-resolution simulation (bottom-up approach) or from experimental observations at resolutions even lower than that of the CG model (top-down approach). Many studies involving CG molecular dynamics simulations focus on modeling single molecules or small groups of molecules, limiting the transferability of the resulting potentials to other compounds so as to render a high-throughput screening approach unfeasible. Therefore, a highly transferable CG model with an easily-implemented mapping protocol is required. In this work,we use the CG Martini force field, which meets

the aforementioned criteria, as it has been applied to study many different types of biomolecules [79, 86, 88, 137]. Martini is a top-down CG model consisting of a set of bead types that are parameterized to reproduce the partitioning behavior of molecules of varying hydrophobicity, from fully polar to fully apolar phases. Atomistic molecules are assigned a combination of Martini bead types by matching the corresponding partitioning free energies of the chemical fragments which make up the molecule.

From a materials design perspective, using this type of transferable CG model is akin to reducing the dimensionality of the problem before running an experiment: the CCS is now first projected onto the CG force field, and the results of the CG simulation can be applied to all of CCS that mapped to a particular CG molecule. Since single CG molecules are representative of specific regions of CCS, running simulations over an entire range of different CG compounds results in a data set that spans CCS to the same extent. The reduced data set can then be projected onto molecular descriptors that are independent of the resolution used when modeling the property of interest. This allows the same structure–property relationship derived from the CG model to be applied to real molecules for which the molecular descriptors are available. Essentially, this high-throughput coarse-grained (HTCG) method has two major advantages over other approaches for computational screening of CCS. The CG mapping results in (*i*) the removal of extraneous atomistic degrees of freedom, which leads to (*ii*) a smoother free energy landscape, as well as (*iii*) a smaller number of particles to simulate in the system, requiring fewer computational resources to sufficiently sample the CG conformational space when compared to its atomistic counterpart. Secondly, because multiple atomistic compounds will map to the same CG representation, coarse-graining effectively reduces the size of CCS, allowing for faster construction of broadly encompassing structure–property relationships. Fig. 2.1 shows an implementation of the HTCG method in order to obtain a structure–property relationship for membrane permeability.

In this chapter, we outline our work in applying the HTCG method to obtain structure–property relationships for key features of the membrane PMF as well as membrane permeability using the CG Martini force field. First, in Section 2.2, we show how systematically running simulations of all 119 neutral Martini single-bead compounds (unimers) and two-bead compounds (dimers) led to the discovery of a linear relationship between the water-octanol partitioning free energy, $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, and the membrane PMF. In Section 2.3, we go on to use the partition free energy as well as the acidity of a compound to construct a structure–property relationship that related molecular structure to membrane permeability. Further extending this approach to Martini trimers and tetramers without any modification would prove computationally unfeasible, as the number of compounds to screen grows exponen-

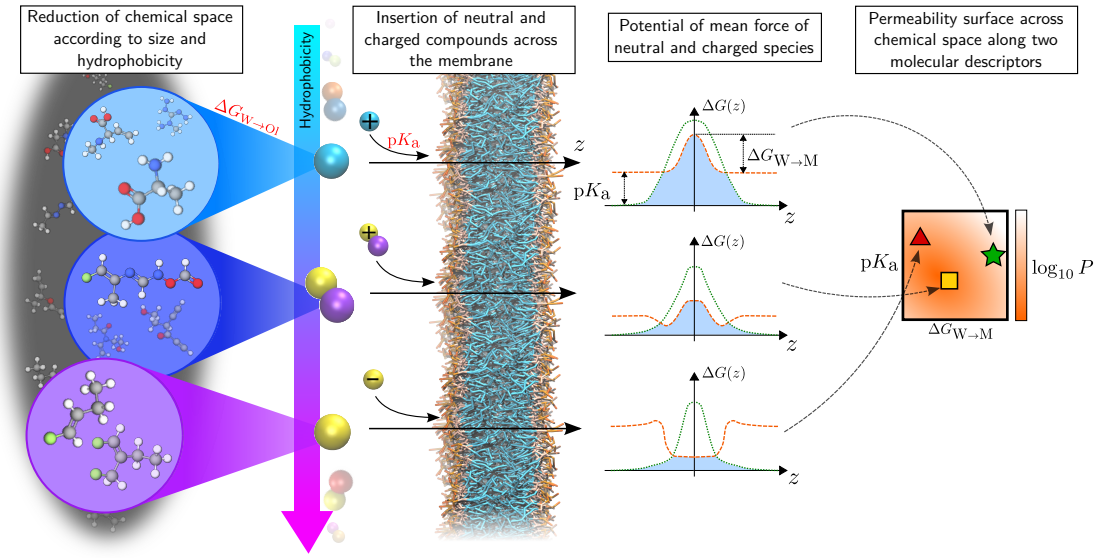

Figure 2.1: Schematic portraying the HTCG method used to calculate membrane permeabilities. The above figure and following caption are reproduced here with permission from Menichetti et al. [138]. From left to right: Coarse-graining reduces the size of chemical space, such that many small molecules of similar size and hydrophobicity get mapped to the same representation [60]. For each molecule, we model its passive translocation across a lipid bilayer (water not shown for clarity). The thermodynamics of the system is characterized by the potential of mean force (PMF), evaluating both the neutral and charged species, shifted according to the compound's pK_a. The major dependence of the PMF on the water/octanol partitioning and the pK_a motivates these as molecular descriptors to construct a permeability surface (Eq. 2.1). These two molecular descriptors, also highlighted in red, are experimental quantities directly fed into the physics-based simulations to yield a parameter-free estimation of the permeability coefficient.

tially with the number of particles per compound. Therefore, in Section 2.4, we discuss our implementation of a Monte-Carlo scheme, which, when coupled with a supervised machine learning technique, enables us to predict the membrane PMFs of Martini trimers and tetramers without fully sampling this CG compound space.

Note that, in all of the aforementioned sections, I was not the largest contributor to each project. Rather, my main role was to apply the algorithm developed by Bereau and Kremer, AUTO-MARTINI, on the computer-generated database of chemical compounds (GDB) developed by Reymond et al, which serves as a proxy for CCS [139, 140]. I could then apply the structure–property relationships obtained through the HTCG method to the set of GDB compounds that mapped to Martini representations, yielding predictions for over 500,000 compounds which mapped to Martini unimers and dimers, and over 1.3 million compounds when including trimers. This process as well as the resulting analyses are detailed in Sections 2.5 and 2.6.

Assume that a certain Martini representation outperforms all others when using the HTCG method to screen for a desired property. Because thousands of chemical compounds can be mapped to a single Martini representation, it remains unclear as to which chemical compounds should actually be chosen for further testing, and it would be costly to computationally screen all of these compounds at an atomistic resolution. In Section 2.7, we use a combination of clustering and dimensionality reduction techniques to suggest a hierarchical screening approach that enables continued sampling of CCS efficiently at higher resolutions. We also compare different molecular representations which depend only on the chemical structure of a molecule, and see if any of these representations enable us to directly link chemical structure with the property used to assign the molecule to a Martini bead type: $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$. Finally, we summarize our findings and discuss future avenues of study in Section 2.10

2.2 Linear relations between bulk partitioning and the potential of mean force

Disclaimer: These sections from the following work by Menichetti et al. are reproduced here with permission.

Roberto Menichetti, Kiran H. Kanekal, Kurt Kremer, and Tristan Bereau
In silico screening of drug-membrane thermodynamics reveals linear relations between bulk partitioning and the potential of mean force
The Journal of Chemical Physics 147(12):125101, 2017.
DOI: 10.1063/1.4987012

2.2.1 Methods **

Molecular Dynamics simulations in this work were performed in GROMACS 4.6.6,[141] using the Martini force field [142–145]. We relied on the standard force field parameters[146] with an integration time step of $\delta t = 0.02\tau$, where τ is the model's natural unit of time as dictated by the units of energy $\mathcal{E}$, mass $\mathcal{M}$ and length $\mathcal{L}$, $\tau = \mathcal{L}\sqrt{\mathcal{M}/\mathcal{E}}$.

A Parrinello-Rahman barostat[147] and a stochastic velocity-rescaling thermostat[67] provided control over the system pressure ($P = 1$ bar) and temperature ($T = 300$ K). The corresponding coupling constants were $\tau_{\mathrm{P}} = 12\,\tau$ and $\tau_{\mathrm{T}} = \tau$.

Bulk simulations consisted of $N_W = 450$ and $N_O = 336$ water and octane molecules, respectively. A DOPC membrane of 36 nm^2 was generated by means of the INSANE building tool[148] and subsequently minimized, heated up, and equilibrated. The total number of lipids in the membrane was $N_L = 128$ (64 per layer), immersed in $N'_W = 1590$ water molecules. As usual when using non-polarizable Martini water, we added an additional 10% of antifreeze particles in the simulations containing water molecules [143].

In the case of two-bead molecules, we first considered a representative subset of 40 coarse-grained compounds, roughly uniformly covering a range of transfer free-energies from water to bilayer midplane of $\Delta G_{\mathrm{W}\to\mathrm{M}} \simeq [-8.14]$ kcal/mol. We determined the corresponding potentials of mean force as a function of the distance z of the compound from the bilayer midplane, $G(z)$, by means of umbrella-sampling techniques [71]. We set biasing potentials with a harmonic constant of $k = 240$ kcal/mol/nm^2 every 0.1 nm along the normal to the bilayer midplane, for a total of 24 simulations. In each of them, two solute molecules were placed in the membrane in order to increase sampling and alleviate leaflet area asymmetry [59, 149, 150]. The total production time for each umbrella simulation was $1.2 \cdot 10^5\,\tau$. We estimated the free-energy profiles by means of the weighted histogram analysis method,[72, 151, 152] and the corresponding errors via bootstrapping [153]. The same calculations were performed in order to determine the potentials of mean force for all of the 14 single-bead compounds analyzed in this work. The computational cost for the reconstruction of each potential of mean force amounted roughly to 200 CPU hours.

We herein focus on calculating $\Delta G_{\mathrm{W}\to\mathrm{I}}$ and $\Delta G_{\mathrm{I}\to\mathrm{M}}$, the transfer free-energies between the three different environments—water (W), interface (I), and bilayer midplane (M)—along the potential of mean force. In terms of $G(z)$, these are defined as $\Delta G_{\mathrm{W}\to\mathrm{I}} = G(\bar{z}) - G(z \to \infty)$ and $\Delta G_{\mathrm{I}\to\mathrm{M}} = G(z = 0) - G(\bar{z})$, where $\bar{z} \approx 1.8$ nm is the position of the lipid-water interface with respect to the bilayer midplane ($z = 0$).

The transfer free energies for all 105 coarse-grained two-bead molecules were determined from alchemical transformations [154]. Given the excellent agreement

between the two end points of a potential of mean force and the water/octane partitioning (see Sec. 2.2.1 and Fig. 2.2), as already pointed out in Ref. [155], the latter was used as a proxy for the hydrophobic core of the membrane.

In the calculation of the $\Delta G_j^{A\to B}$, j = I,W,O, we employed the multistate Bennett acceptance ratio[156] (MBAR), a generalization of the BAR method [157]. MBAR determines the free energy difference $\Delta G_{\mathrm{j}}^{A\to B}$ by appropriately combining the results obtained from simulations that sample the statistical ensembles generated by a set of interpolating Hamiltonians $H(\lambda)$, $\lambda \in [0,1]$, with $H(\lambda=0)=H^A$ and $H(\lambda=1)=H^B$. Specifically, we made use of 21 evenly distributed λ-points between 0 and 1 for each alchemical transformation and in each environment (interface, water and octane). The production time for each λ point was $2\cdot 10^4\tau$ in bulk water and bulk octane and $4\cdot 10^4\tau$ at the interface. The cumulative computational cost of performing each alchemical transformation in water, interface and octane amounted roughly to 60 CPU hours.

2.2.2 Results and Discussion **

Fig. 2.2 shows the excellent agreement between the two end-points of a potential of mean force (i.e., $\Delta G_{\mathrm{W}\to\mathrm{M}} = G(z=0) - G(z\to\infty)$) and the water/octane partitioning, $\Delta G_{\mathrm{W}\to\mathrm{O}}$, which illustrates that bulk octane is a good proxy to represent the hydrophobic core of the bilayer, as already discussed in Ref [155]. A linear fit for the two quantities provided

$$\Delta G_{\mathrm{W}\to\mathrm{O}} = \Delta G_{\mathrm{W}\to\mathrm{M}} - \alpha, \tag{2.2}$$

$\alpha \approx 0.28$, 0.32 kcal/mol for one-bead and two-bead compounds respectively, with Pearson correlation coefficients $R^2 = 0.99$. As described in the Methods section, this allowed us to determine transfer free energies with respect to an octane environment, later converting them to the corresponding membrane values.

For every compound, the transfer free energies are subject to a thermodynamic cycle that links the three variables

$$\Delta G_{\mathrm{W}\to\mathrm{I}} + \Delta G_{\mathrm{I}\to\mathrm{M}} - \Delta G_{\mathrm{W}\to\mathrm{M}} = 0. \tag{2.3}$$

Fig. 2.3 illustrates the relationship between these three transfer free energies for all 119 coarse-grained molecules considered in this work inserted in a DOPC membrane. In both cases of one- and two-bead compounds, beyond the thermodynamic cycle linking $\Delta G_{\mathrm{W}\to\mathrm{I}}$, $\Delta G_{\mathrm{I}\to\mathrm{M}}$ and $\Delta G_{\mathrm{W}\to\mathrm{M}}$, we observe a collapse of the data onto two lines, indicative of a linear relationship between these transfer free energies. Moreover, the only difference between one- and two-bead results consists in the presence of a simple offset (i.e., same slope) between the profiles.

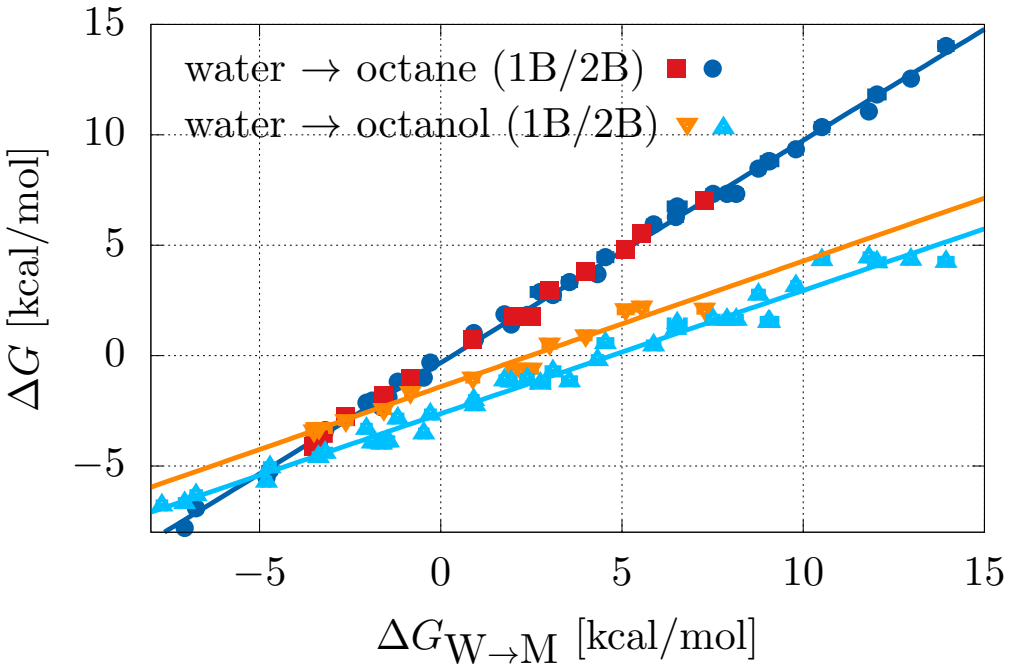

Figure 2.2: Relationship between the two endpoints of the potential of mean force (i.e., $\Delta G_{\mathrm{W}\to\mathrm{M}} = G(z = 0) - G(z \to \infty)$) with the water/octane and water/octanol partitioning free energies, in the case of one- (1B) and two-bead (2B) compounds. Figure and caption reproduced with permission from Menichetti et al. [60].

As a result, the thermodynamic cycle shown in Fig. 1.4 can be reconstructed from the knowledge of a single variable and the Martini bead representation of the compound. The error in doing so amounts to ≈ 0.4. kcal/mol. These relationships are validated from reference atomistic simulations of amino-acid side chains,[59], where we consider only atomistic compounds whose Martini representation consists of a single bead. While most points fall within the linear fit from the single-bead coarse-grained data, we observe three statistically significant outliers: asparagine (asn), isoleucine (ile), and glutamine (gln). These molecules lie on the data corresponding to two-bead compounds, although their Martini representation consists of a single bead. The origin of such discrepancies will be explained below. The comparison of atomistic and Martini potential of mean force for protein side-chains was already performed in Ref. [144].

Remarkably, the relationships between transfer free energies displayed in Fig. 2.3 can further be linked to a compound's water/octanol free-energy $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$, given its accurate linear relation with $\Delta G_{\mathrm{W}\to\mathrm{M}}$, see Fig. 2.2. A fit of the data provided

$$\Delta G_{\mathrm{W}\to\mathrm{M}} = \gamma \Delta G_{\mathrm{W}\to\mathrm{Ol}} + \delta, \tag{2.4}$$

with $\gamma = 1.70,\ 1.75$ and $\delta = 2.51,\ 4.69$ kcal/mol for one- and two-bead compounds, with $R^2 = 0.97$. Given a compound's experimentally determined bulk measurement and Martini representation,[139] we can thereby reconstruct the three main points of the potential of mean force, as shown in Fig. 2.4a. We rationalize these findings by noting the suitability of the octanol environment as a proxy for the membrane interface. Similarly, we showed the appropriateness of octane for the bilayer midplane (Fig. 2.2). Indeed, both water/alcohol and water/alkane coefficients correlate with blood-brain partitioning [158]. Therefore, the relationships in Fig. 2.3 stem directly from the linear correspondence between water/octane and water/octanol transfer free energies (which can be deduced from the linear relations shown in Fig. 2.2). From the model's perspective, the linear relations are not entirely unexpected, as Martini describes hydrophobicity by a set of equally-sized Lennard-Jones particles, with varying well-depths. Interestingly, these relationships also hold at the atomistic level. At infinite dilution, the difference in partitioning of a single small molecule between water and either octane or octanol is due to a single hydrogen bond. We suspect that, at the atomistic level, the impact of this hydrogen bond on the partitioning behavior strongly informs the linearity observed, although the exact mechanism remains unclear.

The statistical errors displayed by the coarse-grained simulations are marginal: less than 0.1 kcal/mol. However, a comparison of experimental measurements of the water/octanol partitioning free energies of several hundred small molecules against Martini predictions yielded a mean-absolute error of 0.79 kcal/mol [139]. Given the relation between the water/octanol and water/midplane curves of Fig. 2.2

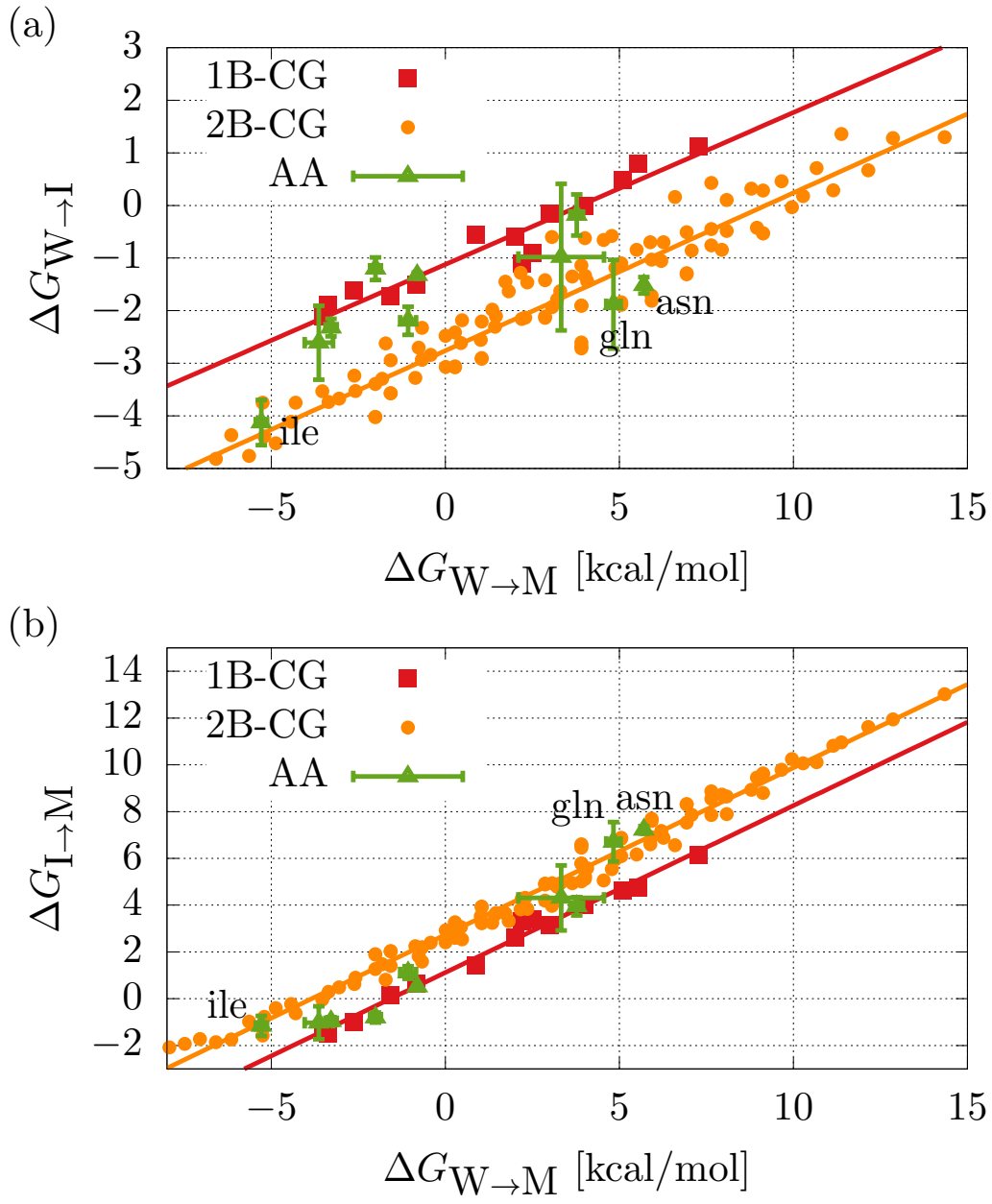

Figure 2.3: (a) Transfer free energies from water to interface $\Delta G_{\mathrm{W}\to\mathrm{I}}$ as a function of the compound's water/membrane partitioning free energy, $\Delta G_{\mathrm{W}\to\mathrm{M}}$. The red and orange curve correspond to coarse-grained estimates for one-bead (1B) and two-bead (2B) molecules, respectively. The green points (AA) depict corresponding atomistic references of amino-acid side chains [59]. (b) Transfer free energies from interface to the membrane $\Delta G_{\mathrm{I}\to\mathrm{M}}$ as a function of the compound's water/membrane partitioning free energy, $\Delta G_{\mathrm{W}\to\mathrm{M}}$. Color coding follows from (a). In both figures, statistically significant outliers (see text) are marked with a label (asn, ile, and gln). Figure and caption reproduced with permission from Menichetti et al. [60].

we deduce from it a mean absolute error on features of the potential of mean force of approximately 1.4 kcal/mol. Further, the error associated with the fitted lines on Fig. 2.3 amounts to an overall error of roughly 1.8 kcal/mol in reconstructing the main points of the potential of mean force—at the bilayer midplane and at the interface, see circles in Fig. 2.4a—by using as input only the experimental water/octanol partitioning free energy of a compound. At the atomistic level, too few potentials of mean force are available to provide errors across chemical compounds.

The linearity observed between the free-energy barrier $\Delta G_{\mathrm{W}\to\mathrm{I}}$ (equivalently $\Delta G_{\mathrm{I}\to\mathrm{M}}$) and the water/membrane partitioning free-energy $\Delta G_{\mathrm{W}\to\mathrm{M}}$ suggests the possibility of looking for an approximately smooth two dimensional free-energy surface $G(z, \Delta G_{\mathrm{W}\to\mathrm{M}})$ across chemical space, hence as a function of $\Delta G_{\mathrm{W}\to\mathrm{M}}$ as well as of the distance from the bilayer midplane z.

In the case of two-bead coarse-grained molecules, we then constructed a two dimensional map of the free-energy surface $G(z, \Delta G_{\mathrm{W}\to\mathrm{M}})$ starting from the set of 40 potentials of mean force that were determined by means of umbrella sampling simulations, covering a range $\Delta G_{\mathrm{W}\to\mathrm{M}} \simeq [-8, 14]$ kcal/mol. Results are shown in Fig. 2.4b.

The correlations shown in Fig. 2.3 between $\Delta G_{\mathrm{W}\to\mathrm{I}}$ and $\Delta G_{\mathrm{W}\to\mathrm{M}}$ for different compounds correspond, on this surface, to the set of points $G(\bar{z} = 1.8 \text{ nm}, \Delta G_{\mathrm{W}\to\mathrm{M}})$. Apart from minor fluctuations, it is evident how the overall smoothness of the surface on the lines with constant z allows us to identify the existence of an average functional relationship between $\Delta G_{\mathrm{W}\to\mathrm{M}}$ of a compound and its potential of mean force $G(z)$ for every value of z. As an example, a small free-energy barrier located at $z \approx 2.5$ nm is present for all the compounds with $\Delta G_{\mathrm{W}\to\mathrm{M}} \in [-8, 0]$ kcal/mol. Small shifts in z may result from bilayer-thickness discrepancies between atomistic and coarse-grained simulations [139].

In this work we focused on the reconstruction of key features of the potential of mean force (i.e., the water/interface and interface/membrane transfer free energies, $\Delta G_{\mathrm{W}\to\mathrm{I}}$ and $\Delta G_{\mathrm{I}\to\mathrm{M}}$). The results shown in Fig. 2.4b further suggests that a knowledge of the water/membrane partitioning free energy of a compound, which can be obtained from the corresponding water/octanol one via the linear relation reported in Eq. (2.4), allows for a semi-quantitative reconstruction of the whole potential of mean force $G(z)$.

2.3 High-throughput coarse-grained screening to obtain membrane permeabilities

Disclaimer: These sections from the following work by Menichetti et al. are reproduced here with permission.

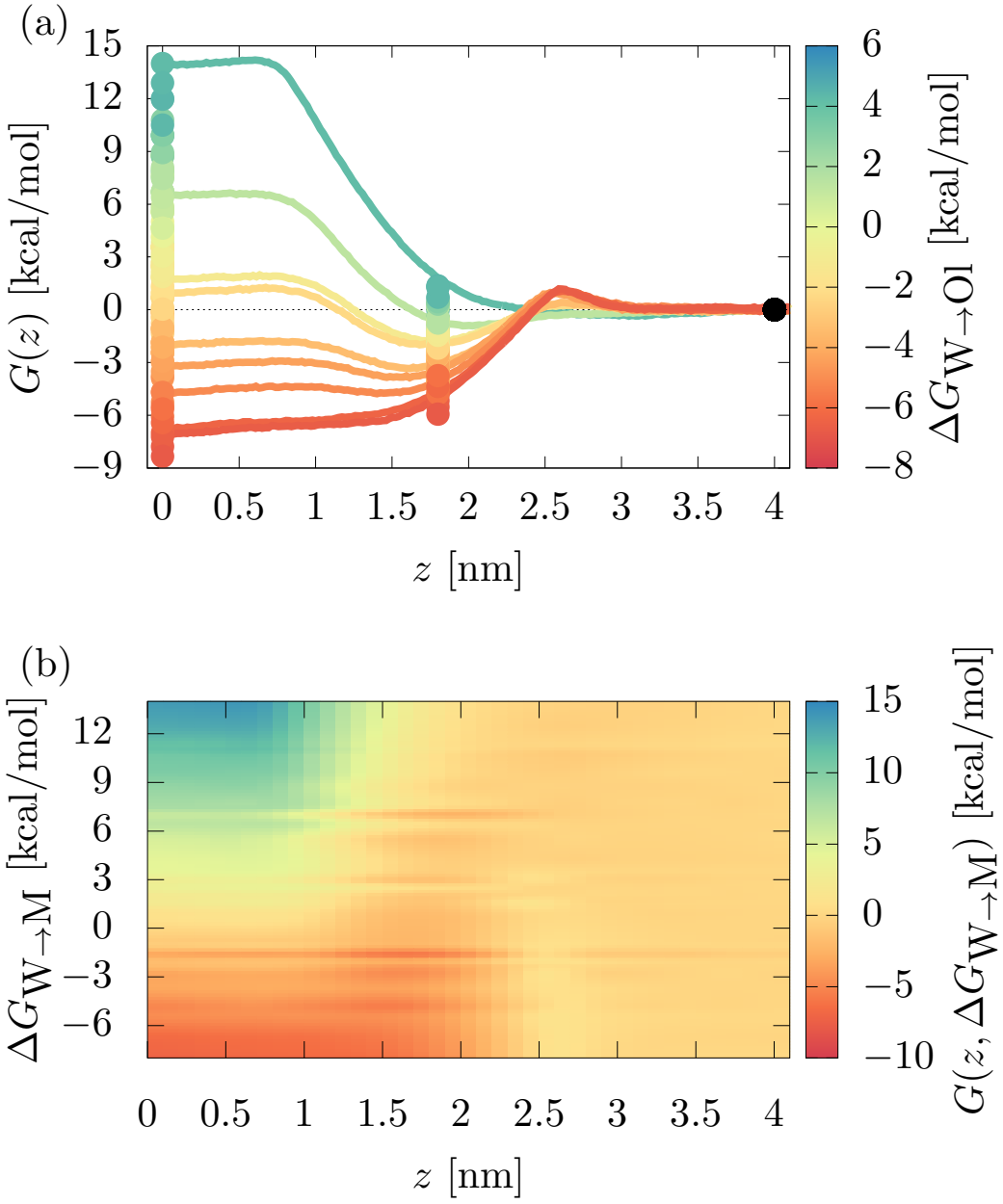

Figure 2.4: (a) Representative potentials of mean force of various Martini compounds as a function of the normal distance to the bilayer midplane. The color range denotes the water/octanol partitioning of the small molecule. Large circles correspond to estimates from the thermodynamic relation extracted in this work. (b) Two dimensional map of the free energy surface $G(z, \Delta G_{\mathrm{W}\to\mathrm{M}})$ for a small molecule, as a function of its distance from the DOPC bilayer midplane z and its membrane/water partitioning free energy $\Delta G_{\mathrm{W}\to\mathrm{M}}$.

Roberto Menichetti, Kiran H. Kanekal, and Tristan Bereau
Drug–membrane permeability across chemical space
ACS Central Science 5(2):290, 2019.
DOI: 10.1021/acscentsci.8b00718

2.3.1 Methods **

Molecular dynamics simulations **

Molecular dynamics simulations in this work were performed in GROMACS 4.6.6[141] and with the Martini force field, [86, 142, 143] relying on the standard simulation parameters [146]. The integration time step was $\delta t = 0.02\,\tau$, where τ is the model's natural unit of time dictated. Sampling from the NPT ensemble at $P = 1$ bar and $T = 300\ K$ was obtained by means of a Parrinello-Rahman barostat [147] and a stochastic velocity rescaling thermostat,[67] with coupling constants $\tau_P = 12\ \tau$ and $\tau_T = \tau$ respectively. We relied on the INSANE building tool[148] to generate a membrane of ≈ 36 nm^2 containing $N = 128$ DOPC lipids (64 per layer), $N' = 1890$ water molecules, $N'' = 190$ antifreeze particles,[143] and enough counterions to neutralize the box. The system was subsequently energy-minimized, heated, and equilibrated.

The potential of mean force $G(z)$ of each compound was determined by means of umbrella sampling [71]. We employed 24 simulation windows with harmonic biasing potentials ($k = 240$ kcal/mol/nm^2) centered every 0.1 nm along the normal to the bilayer midplane. In each of them, two solute molecules were placed in the membrane in order to increase sampling and alleviate leaflet-area asymmetry [59, 149]. The total production time for each umbrella simulation was $1.2 \cdot 10^5\ \tau$. We then estimated the free-energy profiles by means of the weighted histogram analysis method [72, 151, 152].

Permeability coefficients **

The permeability coefficient is obtained from the potential of mean force $G(z)$ and local diffusivity $D(z)$ in the resistivity $R(z) = \exp[\beta G(z)]/D(z)$, see Eq. 2.1. For compounds with multiple protonation states, both neutral and charged species contribute to the total flux, leading to the total resistivity R_T given by[26] $R_\mathrm{T}(z)^{-1} = R_\mathrm{N}(z)^{-1} + R_\mathrm{C}(z)^{-1}$, where R_N and R_C are the resistivities of the neutral and charged species, respectively. In calculating these quantities in the case of a single (de)protonation reaction, one has to offset the corresponding PMFs $G_\mathrm{N}(z)$ and $G_\mathrm{C}(z)$ by the free-energy difference for the acid/base reaction in bulk water[59]

$$G_\mathrm{base} = G_\mathrm{acid} + k_\mathrm{B}T(\mathrm{p}K_\mathrm{a} - \mathrm{pH}) \ln 10, \tag{2.5}$$

see Fig. 2.1, where we systematically consider neutral pH = 7.4. Beyond the distinction between acid and base, we consider both neutral and charged species (Fig. 2.1): (*i*) a neutral acid deprotonates into a charged conjugate base (acidic pK_a or apK_a) and (*ii*) a neutral base protonates into a charged conjugate acid (basic pK_a or bpK_a).

Estimation of the local diffusivity, $D(z)$, using the CG simulations is a priori problematic given the tendency of these models to inconsistently accelerate the dynamics [159].

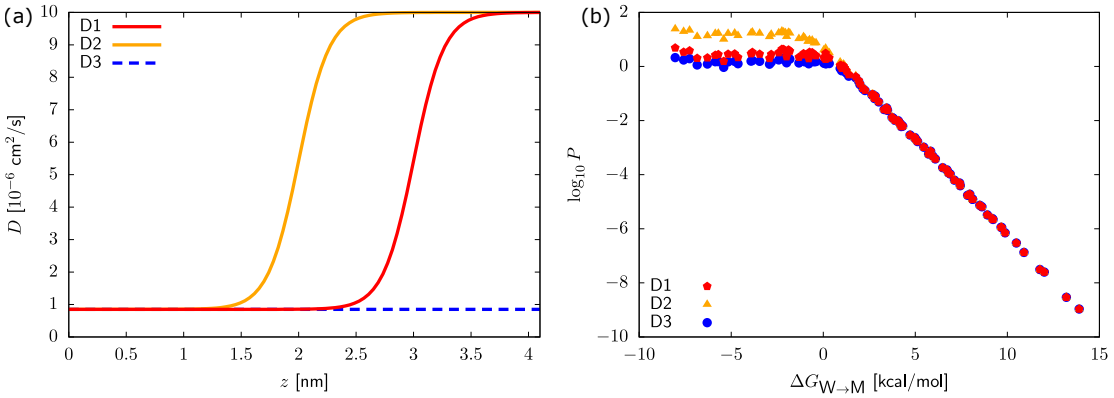

Figure 2.5: (a) Three diffusivity profiles used to determine the sensitivity of $\log P$ to the diffusivity $D(z)$. (b) Changes in $\log P$ as a function of $\Delta G_{\mathrm{W}\to\mathrm{M}}$ for each of the three diffusivity profiles shown in (a). It is clear that the diffusivity has a minor impact on the $\log P$ for apolar molecules only, even when a constant diffusivity value is used. Figure reproduced here with permission from Menichetti et al. [138].

One method for computing $D(z)$ for membrane permeability was implemented by Hummer, in which the diffusion coefficient is obtained via the velocity autocorrelation function for a harmonically restrained particle [160]. Carpenter et al. applied this method to several organic small molecules and noted that the diffusivity profile is highly uniform, even when the chemistry of the molecules varied significantly [26]. Because of this, we approximate the diffusivity profile for small organic molecules by fitting the diffusivity profiles calculated by Carpenter et al. to a sigmoidal function:

$$D(z) = \alpha + \frac{\beta}{e^{-\gamma(x-\delta)} + 1}. \tag{2.6}$$

where α, β, γ, and δ are all parameters obtained from the fit. We also performed a sensitivity analysis with respect to $D(z)$ by investigating how horizontal shifts

and vertical stretching of the function would influence $\log P$. We found that there is only a noticeable change in $\log P$ for hydrophobic molecules, but even this change was within a single log unit, which is an acceptable degree of error for permeability coefficient. This sensitivity analysis is shown in Fig. 2.5.

Permeability surfaces **

We obtained the permeability surfaces presented in Figs. 2.6 and S4 by first determining the PMF $G(z)$ for *all* possible neutral combinations of one and two CG beads, 119 in total. For each of them we then determined $G(z)$ for its charged counterparts, amounting to a total of 232 additional compounds. All PMF calculations required less than 10^5 CPU hours, on par with the typical computational time needed to run a *single* compound at an atomistic resolution [130]. At the CG level, protonating (deprotonating) a neutral chemical group amounts to replacing the bead type with a positive (negative) charge. We assume that the (de)protonation reaction always occurs in the chemical fragment represented by the more polar bead, and select the bead accordingly. By combining neutral and charged PMFs, we calculated the permeability coefficient of every compound as a function of the apK_a (or bpK_a) every 0.2 pK_a unit, and projected the results on the ($\Delta G_{\mathrm{W}\to\mathrm{M}}$, p$K_\mathrm{a}$) plane. The data consisted of a discrete set of permeabilities densely covering the partitioning free-energy axis located at the $\Delta G_{\mathrm{W}\to\mathrm{M}}$ of each CG compound, and were finally interpolated on a grid with gaussian weights resulting in the surfaces shown in Fig. 2.6.

Chemical space coverage *

Prediction of the water/octanol partitioning on both chemical databases considered in this work, GDB[140, 161] and ChEMBL,[162] was performed by means of the neural network ALOGPS [163]. apK_a and bpK_a predictions of neutral compounds were provided by the Calculator Plugin of CHEMAXON MARVIN [164]. The mean absolute error associated with the two prediction algorithms are 0.36 kcal/mol[163] and 0.86 units,[165] respectively. The aggregate predictions of water/octanol partitioning and pK_a on both databases required roughly 10^2 CPU hours. Functional groups were identifed using the CHECKMOL package [166]. 511,427 molecules were coarse-grained using the AUTO-MARTINI scheme [139]. AUTO-MARTINI automatically determines the coarse-grained force field in two steps: (*i*) the CG mapping is optimized according to Martini-based heuristic rules and (*ii*) interactions are set by determining a type for each bead, selected from chemical properties of the encapsulated atoms, especially water/octanol partitioning, net charge, and hydrogen-bonding.

2.3.2 Results and Discussion **

While drug permeation is known to depend on lipid composition,[135] in this work we only consider a single-component bilayer made of 1,2-dioleoyl-*sn*-glycero-3-phosphocholine (DOPC). The permeability coefficient, P, is readily estimated from the PMF and diffusivity profile (Eq. 2.1). The PMFs are extracted from HTCG simulations of *all* CG representations made of one and two beads, mapping to a representative subset of small organic molecules in the range $30-160$ Da [60]. For compounds capable of (de)protonating, we also model the corresponding charged species. For convenience, we distinguish the pK_a of a chemical group as being either acidic (apK_a) or basic (bpK_a), which quantifies the propensity of a *neutral* compound to deprotonate or protonate, respectively. The effective permeability coefficient is constructed by a combination of the two PMFs (Fig. 2.1), shifted according to the compound's pK_a in water, see Methods [59, 164]. The diffusivity profile is estimated from reference atomistic simulations [26].

Fig. 2.6 displays smooth permeability surfaces as a function of the drug's acidic and basic pK_a value in water. The $\log_{10}$ scale of the permeability surfaces indicates the wide timescale variations these molecular parameters exert on the thermodynamic process. For both panels, the horizontal behavior indicates that larger permeabilities are obtained toward the left—more hydrophobic compounds—while polar molecules experience more difficulties crossing the lipid bilayer, leading to a drastic reduction in P. The effect is compounded by (de)protonation: panel (a) across the vertical axis describes the effect of the compound's apK_a in water onto P. Extremely strongly acidic molecules (ap$K_\mathrm{a} \lesssim 2$) effectively remain charged across the membrane interface, leading to prohibitively large free energies along the PMF, such that their rate of permeation is strongly suppressed. Increasing apK_a shows a significant increase in P, up to ap$K_\mathrm{a} \approx 7$, beyond which P plateaus. This stabilization is due to the competition between neutral and charged PMFs, where the charged PMF is shifted to increasingly larger values, and therefore never contributes significantly compared to the more attractive neutral PMF. Of particular interest are the strong acids ($2 \lesssim \mathrm{ap}K_\mathrm{a} \lesssim 7$), which neutralize upon entering the membrane, effectively enhancing the permeability coefficient as compared to a compound that remains charged across the interface. An approximately symmetric behavior can be observed when switching from acidic to basic compounds (panel (b)). The impact of both apK_a and bpK_a on the permeability coefficient becomes even more pronounced in the case of zwitterions, where high permeation rates are only obtained for compounds containing both weak acidic *and* basic chemical groups.

The permeability surface also displays a comparison against atomistic simulations[26, 59, 133] for several compounds (symbols in Fig. 2.6). These points provide a validation of our methodology—we report a mean absolute error of 1.0 $\log_{10}$ unit across

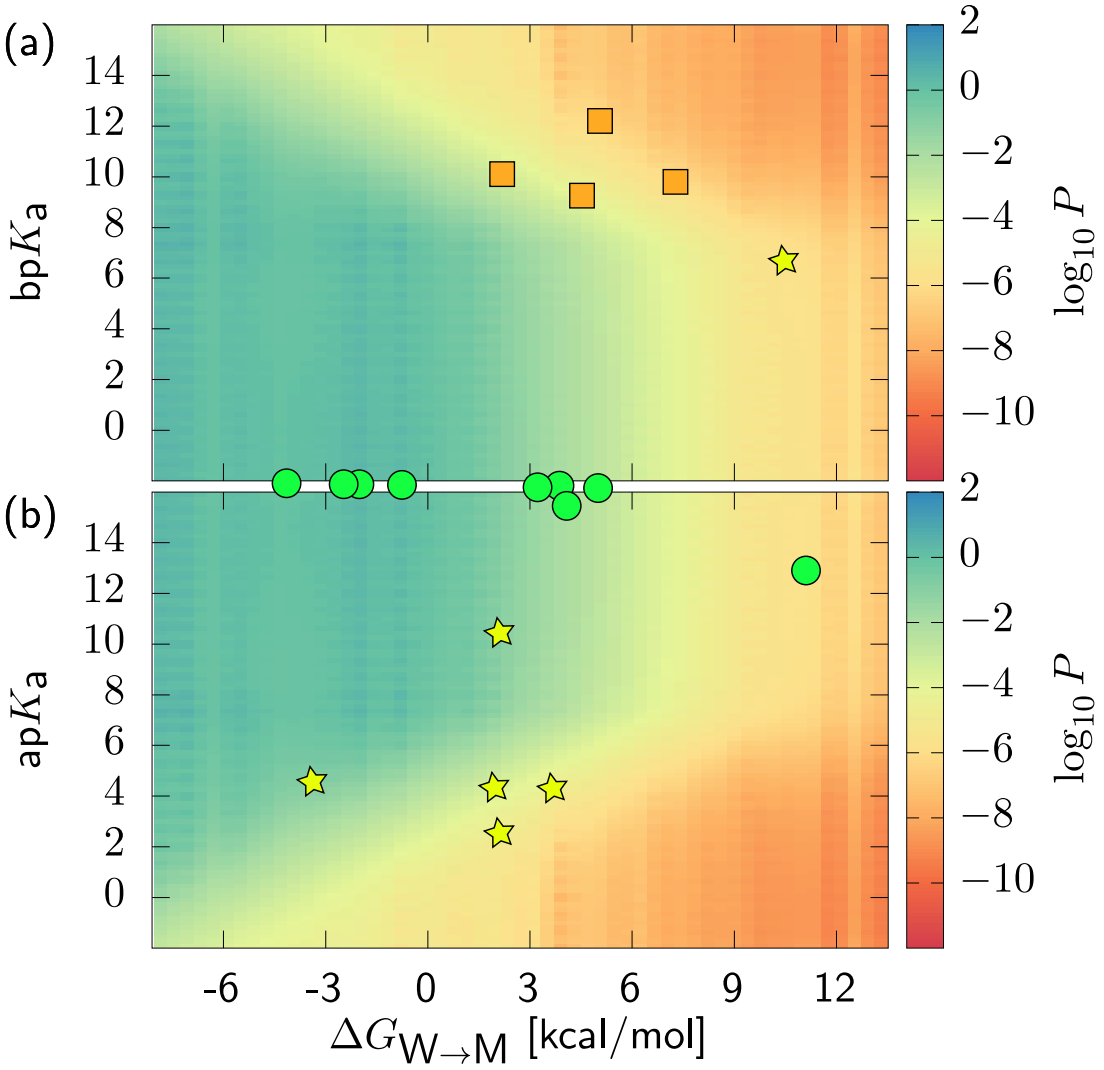

Figure 2.6: Permeability surfaces ($\log_{10}$ scale) calculated from HTCG simulations as a function of two small-molecule descriptors: the (a) basic or (b) acidic pK_a in water and the water/membrane partitioning free energy, $\Delta G_{W\to M}$. Cooler (warmer) colors correspond to faster (slower) permeating molecules. The intersection between the two surfaces corresponds to compounds that effectively always remain neutral. Green circles, yellow stars, and orange squares correspond to deviations from atomistic simulations within 0.5, 1.3, and 2.2 log units, respectively. Figure and caption reproduced from Menichetti et al. [138].

the two molecular descriptors. Most importantly, the few datapoints highlight the extremely limited exploration of chemical space using *in silico* simulations at an atomistic resolution.

2.4 Supervised machine learning applied to the coarse-grained exploration of chemical space

Disclaimer: These sections from the following work by Hoffmann et al. are reproduced here with permission.

Christian Hoffmann, Roberto Menichetti, Kiran H. Kanekal, and Tristan Bereau
Controlled exploration of chemical space by machine learning of coarse-grained representations
Physical Review E 100(3):033302, 2019.
DOI: 10.1103/PhysRevE.100.033302

2.4.1 Methods **

Coarse-grained simulations **

MD simulations of the Martini force field [137] were performed in GROMACS 5.1. The integration time-step was $\delta t = 0.02\ \tau$, where τ is the model's natural unit of time. Control over the system temperature and pressure ($T = 300\ K$ and $P = 1$ bar) was obtained by means of a velocity rescaling thermostat [67] and a Parrinello-Rahman barostat [147], with coupling constants $\tau_T = \tau$ and $\tau_P = 12\ \tau$. Bulk simulations consisted of $N_{\mathrm{W}} = 450$ and $N_{\mathrm{O}} = 336$ water and octane molecules, where the latter was employed as a proxy for the hydrophobic core of the bilayer [60]. As for interfacial simulations, a membrane of 36 nm^2 containing $N_{\mathrm{L}} = 128$ DOPC lipids (64 per layer) and $N'_{\mathrm{W}} = 1890$ water molecules was generated by means of the INSANE building tool [148], and subsequently minimized, heated up, and equilibrated. In all simulations containing water molecules we added an additional 10% of antifreeze particles.

Free-energy calculations **

Water/interface and interface/membrane transfer free energies $\Delta G_{\mathrm{W}\to\mathrm{I}}$ and $\Delta G_{\mathrm{I}\to\mathrm{M}}$ for all compounds investigated in this work were obtained from alchemical transformations, in analogy with Ref. [60]. This construction is based on the relation linking the transfer free energies of two compounds A and B ($\Delta G^A_{\mathrm{W}\to\mathrm{I}}$, $\Delta G^B_{\mathrm{W}\to\mathrm{I}}$

and $\Delta G^{A}_{\mathrm{I}\to\mathrm{M}}$, $\Delta G^{B}_{\mathrm{I}\to\mathrm{M}}$) to the free energies of alchemically transforming A into B in the three fixed environments, $\Delta G_{\mathrm{I}}^{A\to B}$, $\Delta G_{\mathrm{W}}^{A\to B}$ and $\Delta G_{\mathrm{M}}^{A\to B}$

$$\begin{aligned}\Delta G^{B}_{\mathrm{W}\to\mathrm{I}} &= \Delta G^{A}_{\mathrm{W}\to\mathrm{I}} + (\Delta G_{\mathrm{I}}^{A\to B} - \Delta G_{\mathrm{W}}^{A\to B}),\\ \Delta G^{B}_{\mathrm{I}\to\mathrm{M}} &= \Delta G^{A}_{\mathrm{I}\to\mathrm{M}} + (\Delta G_{\mathrm{M}}^{A\to B} - \Delta G_{\mathrm{I}}^{A\to B}).\end{aligned} \tag{2.7}$$

$\Delta G_{\mathrm{I}}^{A\to B}$, $\Delta G_{\mathrm{W}}^{A\to B}$, and $\Delta G_{\mathrm{M}}^{A\to B}$ were determined by means of separate MD simulations at the interface, in bulk water, and in bulk octane. For the calculation of each $\Delta G_{i}^{A\to B}$, $i = \mathrm{I}, \mathrm{W}, \mathrm{M}$ we again relied on the multistate Bennett acceptance ratio (MBAR) [156, 167]. We employed 24 evenly spaced λ-values for each alchemical transformation and in each environment (interface, water, octane). The production time for each λ point was $4 \cdot 10^4$ τ at the interface and $2 \cdot 10^4$ τ in bulk environments. To calculate $\Delta G_{\mathrm{I}}^{A\to B}$ we added a harmonic potential with $k = 1000\ \mathrm{kJ\,mol^{-1}\,nm^{-2}}$ between the compound and the bilayer midplane at a distance $z = 1.5$ nm to account for the spatial localization of the interface.

Monte Carlo sampling **

We perform a stochastic exploration of the chemical space of CG linear trimers and tetramers through the generation of Markovian sequences of compounds. Given the last compound A of a sequence, the new compound B is proposed by randomly selecting a bead of A and changing its type. The move from A to B is then accepted with probability

$$P_{A\to B} = \min\left\{1, \exp\left[-\beta(\Delta G^{B}_{\mathrm{W}\to\mathrm{I}} - \Delta G^{A}_{\mathrm{W}\to\mathrm{I}})\right]\right\}, \tag{2.8}$$

where $\Delta G^{A}_{\mathrm{W}\to\mathrm{I}}$ and $\Delta G^{B}_{\mathrm{W}\to\mathrm{I}}$ are the water/interface transfer free energies of A and B, respectively. This choice for $P_{A\to B}$ aims at driving the Monte Carlo (MC) sampling towards compounds that favor partitioning in the membrane. While in this work we set $\beta = 1/k_{\mathrm{B}}T$, we stress that β can in principle be chosen independently of the system temperature. The free-energy difference in Eq. 2.8 is derived from the alchemical free-energy differences of transforming A into B in the three fixed environments $\Delta G_{i}^{A\to B}$, $i = \mathrm{W}$, I, M (first relation in Eq. 2.7), which we compute from MD simulations.

We generated up to five independent Markovian sequences in parallel, each starting from a different initial compound. To avoid recalculating alchemical transformations already visited, we stored the history of calculations and looked up previously-calculated values when available.

Thermodynamic-cycle optimization **

We reconstructed the transfer free energies $\Delta G_{\mathrm{W}\to\mathrm{I}}$ and $\Delta G_{\mathrm{I}\to\mathrm{M}}$ for all compounds analyzed in this work as a summation over a sequence of alchemical trans-

formations by means of Eq. 2.7. The outcome of our Monte Carlo sampling consists of an "alchemical network" in which each node of the network represents a compound, and an edge connecting two nodes A and B corresponds to an alchemical transformation that was sampled via an MD simulation. Each edge is characterized by the free-energy differences $\Delta G_{\mathrm{i}}^{A\to B}$ in the three fixed environments, $i = \mathrm{W, I, M}$.

For each environment, the net free-energy difference along any closed cycle in the network must be zero, by virtue of a free energy being a state function. We thus enforced this thermodynamic condition to optimize the set of free-energy differences calculated from MD simulations. We employed the algorithm proposed by Paton [168] to identify the cycle basis that spans the alchemical network, i.e., each cycle in the network can be obtained as a sum of the N_C basis cycles. We denote the MD free-energy differences involved in at least one basis cycle by $\Delta G_i^j,\ j = 1,..,N_G,\ i = \mathrm{W, I, M}$, while nodes connected to only a single edge cannot be taken into account. For each environment, we optimized the set of free energies $\Delta\tilde{G}_i^j$ by minimizing the loss function

$$\mathcal{L}_i = \sum_{j=1}^{N_G}(\Delta G_i^j - \Delta\tilde{G}_i^j)^2 + \sum_{k=1}^{N_c}\omega\Big(\sum_{j\in k}(-1)^{s_{j,k}}\Delta\tilde{G}_i^j\Big)^2. \tag{2.9}$$

While the first term ensures that the optimized free-energy differences $\Delta\tilde{G}_{\mathrm{i}}^j$ remain close to the MD simulation results, the second term ($\omega = 10.0$) penalizes deviations from zero for each thermodynamic cycle within a basis cycle. The exponent $s_{j,\alpha}$ controls the sign of the free-energy difference in the cycle, taking values of 0 or 1. To minimize the cost functions, we employed the Broyden-Fletcher-Goldfarb-Shanno method (BFGS) [169].

Machine learning **

We use kernel ridge regression (KRR) [170], where the prediction of target property $p(\mathbf{x})$ for sample $\mathbf{x}$ is expressed as a linear combination of kernel evaluations across the training points $\mathbf{x}_i^*$:

$$p(\mathbf{x}) = \sum_i \alpha_i K(\mathbf{x}_i^*, \mathbf{x}). \tag{2.10}$$

The kernel consists of a similarity measure between two samples

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||_1}{\sigma}\right), \tag{2.11}$$

which corresponds to a Laplace kernel with a city-block metric (i.e., L_1-norm), and σ is a hyperparameter. The optimization of the weights α consists of solving for the

samples in the training with an additional regularization term λ: $\alpha = (\mathbf{K}+\lambda \mathrm{I})^{-1}\mathbf{p}$. Error bars are computed using the predictive variance

$$\epsilon = \mathbf{K}^{**} - (\mathbf{K}^{*})^{T}(\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{K}^{*}, \tag{2.12}$$

where $\mathbf{K}^{**}$ and $\mathbf{K}^{*}$ represent the kernel matrix of training with training and training with test datasets, respectively [170]. The two hyperparameters σ and λ were optimized by a grid search, yielding $\sigma = 100$ and $\lambda = 10^{-4}$.

The representation ought to include enough information to distinguish a compound's chemical composition and geometry, as well as encode the physics relevant to the target property [19]. Because the CG compounds all consist of beads arranged linearly and equidistant, we have found that encoding the geometry had no benefit to the learning. Instead we simply encode the water/octanol partitioning of each bead, yielding for linear trimers $\mathbf{x} = \left(\Delta G^{(1)}_{\mathrm{W}\to\mathrm{Ol}}, \Delta G^{(2)}_{\mathrm{W}\to\mathrm{Ol}}, \Delta G^{(3)}_{\mathrm{W}\to\mathrm{Ol}}\right)$. Note that while the problem we consider in this work contains reflection symmetry for the compounds (i.e., ABC is equivalent to CBA), we did not need to encode this in the representation. Instead we sorted the bead arrangement when generating compounds for the importance sampling and machine learning.

We consider the insertion of a small molecule across a single-component phospholipid membrane made of 1,2-dioleoyl-*sn*-glycero-3-phosphocholine (DOPC) solvated in water. The insertion of a drug is monitored along the collective variable, z, normal distance to the bilayer midplane (Fig. 1.4). We focus on three thermodynamic state points of the small molecule: the bilayer midplane ("M"), the membrane-water interface ("I"), and bulk water ("W"). We link these quantities in terms of transfer free energies, e.g., $\Delta G_{\mathrm{W}\to\mathrm{M}}$ denotes the transfer free energy of the small molecule from water to the bilayer midplane.

2.4.2 Results and Discussion **

Importance sampling **

We ran MC simulations across CG linear trimers, randomly changing a bead type, calculating the relative free energy difference between old and new compound in the three different environments, and accepting the trial compound using a Metropolis criterion on the water/interface transfer free energy $\Delta G_{\mathrm{W}\to\mathrm{I}}$. This criterion aimed at selecting compounds that favor partitioning in the membrane.

The MC algorithm yielded an acceptance ratio of 0.2. While initially most trial compounds contributed to expand the database, the sampling scheme quickly reached a stable regime where roughly half of the compounds had already been previously visited. Because each free-energy calculation is expensive, we avoid recalculating identical alchemical transformations to help efficiently converge the protocol.

Interestingly, we find a large number of closed paths within the network of compounds sampled via our MC algorithm. Since the free energy is a state function, the closed path represents a thermodynamic cycle—it must sum up to zero. We found negligible changes in the free energies regardless of whether or not this condition was enforced on the closed path, meaning that our chain of transformations does not compound significant statistical error.

Machine learning **

We trained an ML model using the vector of water/octanol partitioning of each Martini bead type—one of their salient properties [139]. When trained on most of the MC-sampled data, we obtained out-of-sample mean absolute errors (MAE) as low as 0.2 kcal/mol for $\Delta G_{\mathrm{W}\to\mathrm{I}}$ and $\Delta G_{\mathrm{I}\to\mathrm{M}}$, on par with the statistical error of the alchemical transformations. Remarkably, the prediction of $\Delta G_{\mathrm{W}\to\mathrm{M}}$ converges to an MAE lower than 0.05 kcal/mol, illustrative of the strong correlation between water/octanol and water/membrane free energies in Martini [60]. For all three quantities we monitor a correlation coefficient above 97%, indicating excellent performance.

Next, we train our ML model on the entire dataset of MC-sampled compounds. We use this model to predict all other CG linear trimers. Because of the importance-sampling scheme, the predicted compounds will typically feature different characteristics, e.g., more polar compounds that would preferably stay in the aqueous phase. As such the ML model is technically extrapolating outside of the training set. This can be seen in Fig. 2.7, where the projections on the top and the right highlight the distinct coverages of sampled and predicted compounds along each variable. Yet, the main panels (a) and (b) display strong linear relations between transfer free energies. These correlations are not built in the ML models, since we optimize independent weight coefficients for the different target properties. They also offer higher accuracy compared to simple linear fits: MAE of 0.3 and 0.5 kcal/mol for the ML and linear fit, respectively, across a small set of reference compounds in the sampled dataset. Importantly, these linear relations had already been highlighted in previous work for CG unimers and dimers (data reproduced on Fig. 2.7) [60]. The linear behavior displayed across both sampled and predicted compounds testifies to the robustness of the ML model, despite the extrapolation.

A systematic coarse-graining of compounds in the GDB using AUTO-MARTINI was performed to identify small organic molecules that map to CG linear trimers [139, 140]. We identified 1.36 million compounds, for which we can associate all three transfer free energies, $\Delta G_{\mathrm{W}\to\mathrm{M}}$, $\Delta G_{\mathrm{W}\to\mathrm{I}}$, and $\Delta G_{\mathrm{I}\to\mathrm{M}}$. We note that the sampled and predicted CG representations amount to similar numbers of compounds, such that the ML boosting introduced here offers an additional 0.8 million compounds to the database.

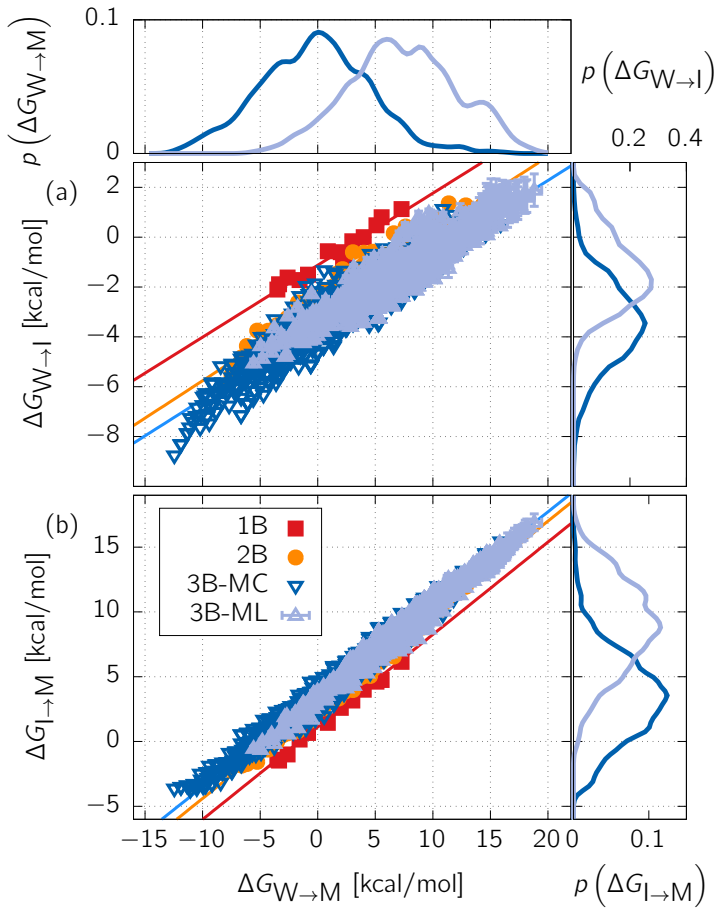

Figure 2.7: (a) Transfer free energies from water to interface $\Delta G_{\mathrm{W}\to\mathrm{I}}$ as a function of the compound's water/membrane partitioning free energy, $\Delta G_{\mathrm{W}\to\mathrm{M}}$. The red and orange curves correspond to coarse-grained estimates for unimers (1B) and dimers (2B), respectively, previously presented in Menichetti et al. [60]. The dark and light blue points depict corresponding quantities for trimers estimated from MC sampling (3B-MC) and the ML predictions (3B-ML), respectively. Linear fits highlight the molecular-weight dependence. (b) Transfer free energies from the interface to the membrane $\Delta G_{\mathrm{I}\to\mathrm{M}}$ as a function of the compound's water/membrane partitioning free energy, $\Delta G_{\mathrm{W}\to\mathrm{M}}$. The coverages are projected down along a single variable on the sides. Error bars for 3B-MC are on par with the datapoint sizes (not shown), while error bars for 3B-ML display the 95% confidence intervals from the predictive variance. Figure and caption reproduced with permission from Hoffmann et al. [110].

The overwhelming size of chemical space naturally calls for statistical techniques to analyze it. A variety of data-driven methods such as quantitative structure–property relationships (QSPR) and ML models at large have been applied to chemical space [17, 119, 171, 172]. While sparse databases easily lead to overfitting [130], a dense coverage can offer unprecedented insight [173]. Here we rely on tools from statistical physics to ease the exploration of chemical space: the application of importance sampling guides us toward the subset of molecules that enhance a desired thermodynamic property. The latter is akin to recent generative ML models [6], but without the a priori requirement for labeled training data.

A conceptually-appealing strategy to expand the MC-sampled distribution is through an ML model. Effectively we train an ML model on the MC samples and further boost the database with additional ML predictions. Unfortunately, the limited extrapolation behavior of kernel models means that accurate predictions can only be made for compounds *similar* to the training set. *How* similar is often difficult to estimate a priori. Similarity metrics are often based at the level of the ML's input space—here the molecular representation. For instance, the predictive variance estimates error bars based on the query sample's distance to the training set [170].

Instead of basing a similarity metric on the ML's input space, we focus on the target properties directly. Our physical understanding of the problem offers a clear requirement on the transfer free energies, through the linear relationships shown in Fig. 2.7 [60]. As such, the thermodynamics of the system impose a physically-motivated constraint on the predictions. Rather than specific to each prediction, this constraint is *global* to the ensemble of data points. Satisfying it grounds our predictions within the physics of the problem, ensuring that we accurately expand the database.

Remarkably, we find that we can significantly expand our database—doubling it for trimers and a factor of 10 for tetramers—while retaining accurate transfer free energies. Unlike conventional atomistic representations [174], our ML model is encoded using a CG representation, such that compounds need only be similar at the CG level. This CG similarity is strongly compressed because (*i*) of a more straightforward structure–property link [138] and (*ii*) coarse-graining reduces the size of chemical space [60]. All in all, backmapping significantly amplifies the additional region of chemical space reached by the ML model. Our work highlights appealing aspects of bridging physics-based methodologies and coarse-grained modeling together with machine learning, offering better robustness and transferability to explore significantly broader regions of chemical space.

2.5 Reduction of chemical space due to coarse-graining *

Disclaimer: This section from the following work by Menichetti et al. is reproduced here with permission.

Roberto Menichetti, Kiran H. Kanekal, Kurt Kremer, and Tristan Bereau
In silico screening of drug-membrane thermodynamics reveals linear relations between bulk partitioning and the potential of mean force
The Journal of Chemical Physics 147(12):125101, 2017.
DOI: 10.1063/1.4987012

Fig. 2.8 shows how the Martini model groups molecules into fewer coarse-grained representations, thereby effectively reducing the size of chemical compound space. This grouping stems from the discrete set of bead types of the Martini model, which assigns the same representation to groups that are chemically similar. To estimate this grouping, we have coarse-grained compounds from the Generated Database[140] of molecules up to ten heavy atoms. In Fig. 2.8, we show the distributions of compounds that map to any one of the one- and two-bead coarse-grained representations considered here, as a function of the water/octanol partitioning. The atomistic distributions of Fig. 2.8b,d were obtained using the ALOGPS neural network [175]. Despite the uneven spacing of the water/octanol partitioning free energies of the coarse-grained molecules, the atomistic distributions are roughly reproduced by the coarse-grained distributions in Fig. 2.8a,c, except for small artifacts in the strongly polar regime (i.e., $\Delta G_{\mathrm{W}\to\mathrm{Ol}} \gtrsim 2.0$ kcal/mol). In total, we identified 465,387 unique molecules, representing most synthetically-feasible small organic molecules between 30 and 160 Da. This many-to-few mapping arises solely from the limited representability of thermodynamic properties of chemical groups, rather than the coarser structural representation (i.e., atoms to beads).

The removal of chemically and structurally specific information present in atomistic simulations is traditionally viewed as a necessary drawback for access to otherwise computationally prohibitive simulations. However, it is precisely this drawback that enables a single coarse-grained simulation to be representative of a large number of small molecules, as degenerate chemical groups are mapped to the same bead type. As a counterexample, fixing a Martini-like mapping in combination with a non-transferable, chemically-specific parametrization (e.g., as in most bottom-up, structure-based models) would prevent any reduction in chemical space. This work thereby introduces the ability for transferable coarse-grained models to screen large numbers of small molecules.

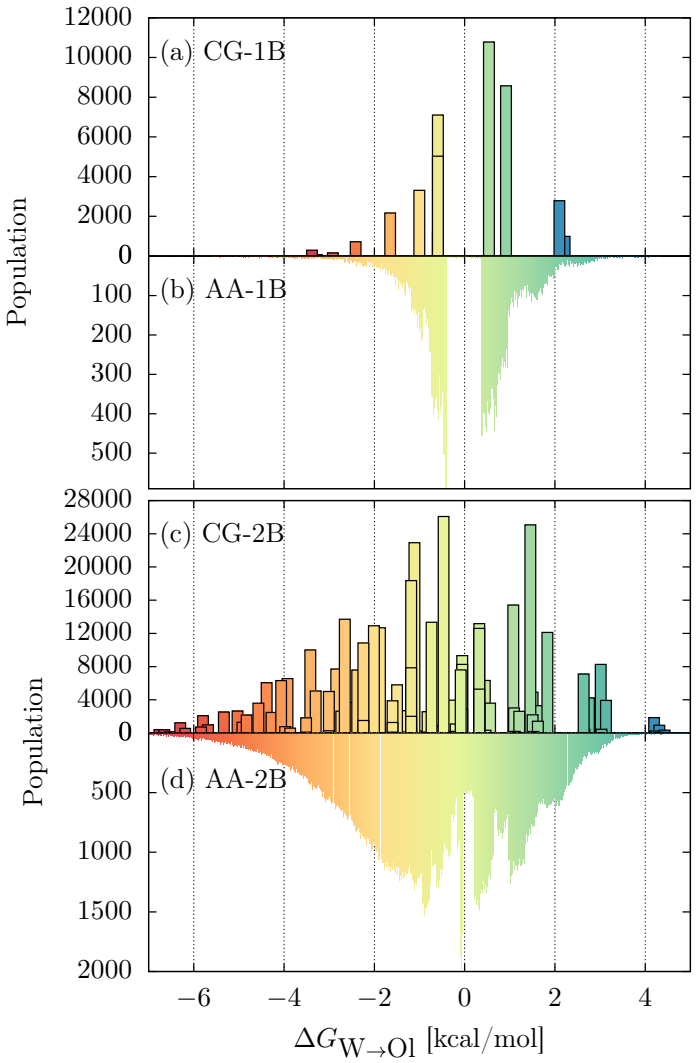

Figure 2.8: Histograms of 465,387 small molecules extracted from GDB that map onto one-bead or two-bead coarse-grained representations. (a),(c) Coarse-grained and (b),(d) atomistic populations as a function of water/octanol partitioning free energy. The width of the bars in (a),(c) have no physical significance and are simply for the reader's convenience.Figure and caption reproduced with permission from Menichetti et al. [60].

Additionally, we report predictions for the transfer free-energies from water to the bilayer midplane or interface estimated from both coarse-grained simulations and the thermodynamic relations displayed in Fig. 2.3. For the latter case, the water/octanol partitioning free energy predicted by ALOGPS is used as the input.

2.6 Relating property back to structure *

Disclaimer: These sections from the following work by Menichetti et al. are reproduced here with permission.

Roberto Menichetti, Kiran H. Kanekal, and Tristan Bereau
Drug–membrane permeability across chemical space
ACS Central Science 5(2):290, 2019.
DOI: 10.1021/acscentsci.8b00718

To better elucidate how the chemical structure impacts the permeability coefficient, we consider a large database of small organic molecules from combinatorial chemistry: the generated database (GDB) [140, 161]. It consists of a large set of stable molecules up to 10 heavy atoms made of the chemical elements C, O, N, and F, saturated with H. We pointed out how transferable coarse-grained models effectively reduce the size of chemical space by lumping many molecules into one coarse-grained representation [60]. This allows us to associate the above-mentioned one- and two-bead CG permeability results to 5×10^5 molecules. The distinction made between compounds that reduce to CG molecules made of a single bead ("unimers") from those made of two beads ("dimers") effectively amounts to a segregation between molecular weights [60]. We populate the permeability surfaces with these compounds—projecting them onto the two molecular descriptors: pK_a and water/octanol partitioning free energy $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$. By coarse-graining every single compound, we establish a map between chemical structure and its CG thermodynamic property.

Fig. 2.9 displays the chemical-space coverage of GDB compounds onto the molecular descriptors. For all panels, we have colored the points in terms of the permeability calculated using HTCG simulations. Top and bottom panels distinguish between bpK_a and apK_a, while left and right denote unimers and dimers, respectively. We first note that the cloud of points is not uniformly distributed, but is instead centered around zero in $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$. An increase in the molecular weight of the compound (left to right in Fig. 2.9) opens up new regions of chemical space, as we observe a significant broadening of the distribution along the water/octanol axis. This naturally arises due to the extensivity of the water/octanol partitioning,

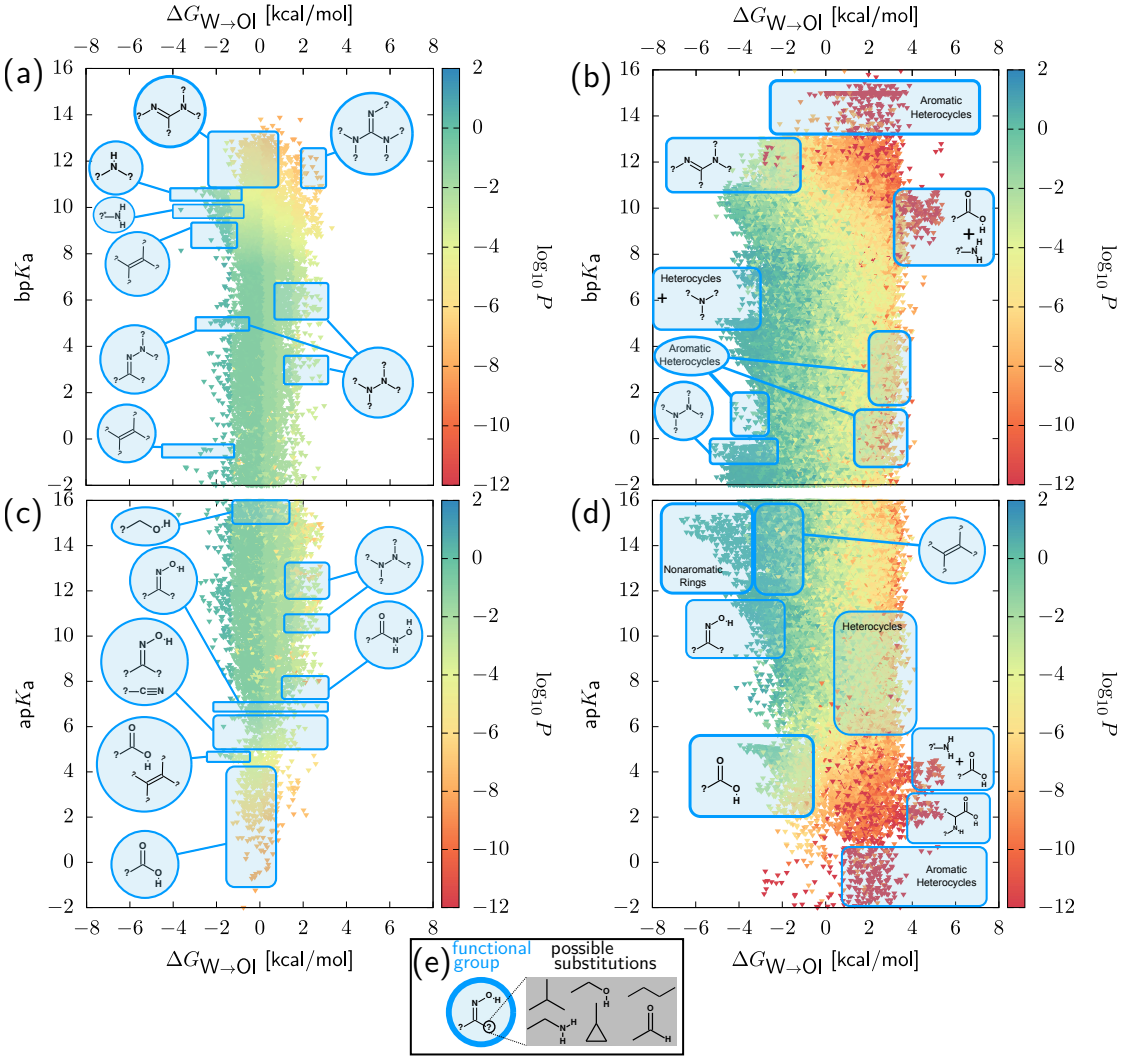

Figure 2.9: Chemical-space coverage of GDB projected onto pK_a and water/octanol partitioning free energies, $\Delta G_{W\to Ol}$. Basic and acidic pK_a are shown in panels (a,b) and (c,d), respectively. Panels (a,c) and (b,d) describe the coverage corresponding to coarse-grained unimers and dimers, respectively. Regions highlighted in light blue display several representative chemical groups. Substitutions denoted by "**?**" correspond to either H or a substitution starting with an alkyl or aryl carbon, while "**?***" only corresponds to substitutions that begin with an alkyl carbon. (e) Our analysis clusters molecules containing both a predominant functional group (blue), but also one or several substitutions (black), of which only a few possibilities are shown. Figure and caption reproduced with permission from Menichetti et al. [138].

the more complex combinatorics of atoms involved, and the additional presence of five-membered rings.

Unlike bulk partitioning, the pK_a of a compound is not significantly impacted by aggregate behavior, but is instead dominated by one or a few specific chemical groups capable of (de)protonating. As such, we investigated the presence of chemical groups representative of a subset of chemical space. The regions in blue highlight a chemical group that is predominant, appearing in at least 50% of the molecules in that subset. The localization of chemical groups remains largely similar from unimers to dimers (e.g., carboxylic group). Our high-throughput analysis offers an intuitive visualization of the link between chemistry and permeabilities via the pK_a. Fig. 2.9 reflects that oxygen-containing functional groups are generally more likely to be proton donors, whereas nitrogen-containing functional groups can serve as either proton donors or acceptors [176]. At low apK_a values, we mainly see carboxylic groups transitioning to nitrogen-containing functional groups (e.g., oxime derivatives) as we increase the apK_a. Contrastingly, the bpK_a chemical coverage displays no predominant oxygen-containing functional groups. Notable exceptions are the zwitterionic amino acid-like compounds and certain aromatic heterocyclic compounds shown in Fig. 2.9, which have both a low apK_a and a high bpK_a. These functional groups largely contribute to the chemical coverage of zwitterions.

2.6.1 Functional Group analysis and molecular design *

Fig. 2.9 enables a robust ad hoc method for both direct and inverse molecular design. The direct route amounts to estimating the permeability coefficient given a chemical structure. Fig. 2.9 simply requires an estimate for the two molecular descriptors, pK_a and $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$, either from experiments or prediction algorithms [163, 164]. More interestingly, our results allow us to focus on specific regions of chemical space compatible with a desired permeability coefficient. We effectively reduce the high dimensionality of chemical space by projecting down onto our molecular descriptors and identifying key scaffolds.

Fig. 2.9 offers a simple route at an inverse design procedure. For example, if designing a small molecule of 3 to 5 heavy atoms (i.e., mapping to a CG unimer) that requires a $\log_{10} P$ of -1.0, Fig. 2.9c suggests molecules containing either a terminal hydroxyl group or an oxime group. Indeed, small alcohols such as propanol and butanol match this target, although we are not aware of relevant experimental studies containing small oxime derivatives. Interestingly, we can also predict how small chemical changes will affect permeability: a change that impacts hydrophobicity (e.g., through hetereoatom substitions) will smoothly shift the compound horizontally on the surface. On the other hand, the introduction of new (de)protonatable groups might lead to large jumps on the surface, dictated by

the strongest acid or base present in the molecule. The different behavior across the horizontal and vertical axes is due to the extensive and intensive characters of the descriptors, respectively.

Critically, Fig. 2.9 shows remarkable transferability *outside* the range of compounds used in the screening. For example, while salicylate is made up of 10 heavy atoms, its aromatic ring leads to a four-bead representation. CG simulations using this parametrization result in $\log_{10} P = -4.21$, deviating only one $\log_{10}$ unit from the atomistic results (highlighted as one of the symbols in Fig. 2.6) [26]. Alternatively, we can easily read off the permeability from the surface: the carboxylic group is the main contributor for its descriptors $\mathrm{ap}K_\mathrm{a} = 2.8$ and $\Delta G_{\mathrm{W}\to\mathrm{Ol}} = -2.7$ kcal/mol (Fig. 2.9). This results in a *simulation-free* prediction for $\log_{10} P$ of -3.72, less than two log units away from the atomistic results. The discrepancy between the four-bead representation and the dimer surface we rely on is the main source of errors: we have observed a systematic shift between $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ and $\Delta G_{\mathrm{W}\to\mathrm{M}}$ as a function of the number of CG beads [60]. An even more challenging test case involved ibuprofen (206 Da, significantly outside our range of molecular weights), for which both CG simulations and the surface prediction yield an accuracy within $1\log_{10}$ unit within the atomistic results (symbol in Fig. 2.6). The transferability beyond the initial molecular weight considered speaks to the robustness of our physics-based approach. This feature contrasts radically with statistical methods that fit experimental data, such as QSPR: the transferability of a QSPR model hinges upon potential biases in the training dataset. Given the small dataset sizes available from experiments and the wider range of molecular weights, QSPR models tend to be limited to chemistries very close to those used in training [177, 178]. On the other hand, the HTCG method systematically spans a wide region of chemical compound space without resorting to parameter tuning, offering accurate predictions even beyond the range of molecular weight considered.

2.7 Unsupervised Machine Learning as a Route to Further Screening

Using the top-down CG Martini force field, we have now seen that coarse-graining effectively reduces the size of CCS, with thousands of molecules mapped to individual Martini representations. Furthermore, we have demonstrated that this HTCG method enables both direct and inverse molecular design. When applied in practice, however, the CG resolution of the structure–property relationship may not be sufficient when screening compounds. For example, if a suitable CG representation is found that satisfies the target design criteria, it remains unclear as

to which compounds would be the most suitable for synthesis and experimental testing. At this point, computational sampling at the atomistic resolution would be the next logical step, but with thousands of compounds mapping to a single CG molecule, problems of computational feasibility arise once again.

Recently, there has been a significant progress in applying supervised ML techniques to directly construct structure property relationships at the atomistic resolution, with several examples of highly accurate prediction of quantum-mechanical properties of small organic molecules [17, 119, 171, 172]. Many of these studies successfully used kernel ridge regression (KRR) to predict the heats of formation of these molecules with accuracy comparable to that of costly ab initio simulations which are normally used. Huang et al. has previously showed that, when using kernel ridge regression, the mathematical representation of the molecule plays a significant role in determining the accuracy of the model [111]. While the supervised learning approach is not easily implemented for soft-matter properties due to the difficulty in obtaining training data, it is possible that some of the molecular representations developed for the prediction of quantum-mechanical properties are also correlated with the thermodynamic properties of the same molecules.

In the following sections, we apply three molecular representations on a subset of CCS that maps to Martini molecules and use unsupervised ML techniques to determine whether these representations correlate with the water/octanol partition free energy ($\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$), a thermodynamic property. We choose $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ for two reasons: (*i*) the ALOGPS neural network provides an independent method to quickly predict the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ of the molecules used in the input data set, removing the need for additional experiments or simulation to obtain this property, (*ii*) $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ is the property used by the AUTO-MARTINI algorithm to assign chemical fragments to Martini bead types, and therefore, correlating a molecular representation with $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ is equivalent to predicting the corresponding Martini bead assignment. Successfully correlating these molecular representations with $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ would point towards a more efficient route to the prediction of thermodynamic properties using kernel ridge regression, as opposed to a more computationally costly neural network like ALOGPS. While some molecular representations will not strongly correlate with $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, they may correlate well with other properties relevant to the screening process. In the final section, we demonstrate that the combination of these molecular representations with unsupervised learning methods enables a hierarchical screening approach. Rather than exhaustively sampling the compounds that map to a single Martini representation, this would allow for the sampling of separate clusters as a computationally feasible alternative.

2.7.1 The set of unique fragments mapping to Martini beads

In our previous work, we were able to predict permeability coefficients for over 500,000 GDB compounds with 10-or-fewer heavy atoms made up of elements C,F,O, and N, mapping to Martini unimers or dimers. The set of compounds chosen for this work was obtained by first taking the set of over 400,000 compounds from this data set that mapped to dimers. Using the AUTO-MARTINI algorithm, each of these molecules was split into the two fragments, which were output as SMILES strings using the RDKIT package. All repeated fragments were removed, and the number of heavy atoms per fragment was restricted to 6 only. After applying all of these filtering steps, the final database consisted of 2035 unique fragments. Note that, despite referring to them as fragments, both during the coarse-graining process and for the subsequent analysis, hydrogen atoms are added by RDKIT to satisfy the valency criteria of each fragment, making them into whole molecules.

2.7.2 Molecular Representations

Each molecule in the data set is first converted from an input SMILE string to a 3-D structure that is then energy minimized using the UFF force field in a molecular mechanics based optimization scheme with the RDKIT package. The atomic numbers and internal geometries are then taken from this 3-D structure and used to create the three representations studied in this chapter: the Coulomb Matrix, the Spectrum of London Axilrod-Teller-Muto vector, and the alchemical smooth overlap of atomic positions (ASOAP) kernel [119, 120, 122]. For a full description of these representations, see Sec. 1.1.7. Since the number of heavy atoms is restricted to 6, the dimensionality of the Coulomb Matrices used in this work is 21. We used our own PYTHON code to convert the 3-D conformation generated by RDKIT into this representation. To convert our database of compounds into SLATM representations, we applied the QML package made for PYTHON 2.7 [179]. A cutoff value of 0.48 nm was used with a grid spacing of 0.003 nm and 0.03 radians used for the 2-body and 3-body spectra, respectively. As a result, the length of the SLATM vectors was 6694, making SLATM the representation with the highest dimensionality out of the three used in this work. Finally, the alchemical SOAP kernel was constructed using the 2035 molecules making up the data set, meaning that each molecule was represented as an array of 2035 distances to every other molecule. We used the GLOSIM code made available by the Laboratory of Computational Science and Modeling at the EPFL, Switzerland to compute the ASOAP distance matrix. While we have used the terms "similarity" matrix and "kernel" interchangeably in this section, we clarify that, in much of the KRR literature (for example as was shown in Eq. 2.11) usually the kernel matrix refers

to the kernelized similarity matrix, meaning a kernel function (e.g. Gaussian or Laplacian) has been applied to the similarity matrix. However, the input used for dimensionality reduction in this study was the un-kernelized similarity matrix, which is the matrix of pairwise dot-product distances between molecules.

2.7.3 Dimensionality Reduction Results

We initially converted the database of unique fragments mapping to Martini dimers with 6 heavy atoms into three corresponding high-dimensional databases for each representation studied in this work: the Coulomb Matrix, SLATM vector, and ASOAP kernel. In order to qualitatively determine how each of these representations transforms CCS, the next step was to apply dimensionality reduction methods in order to visualize this effect. However, it is unclear as to which dimensionality reduction technique would be best suited to accurately understand how the data is structured in the high-dimensional space. Just as the choice of representation could highlight different features of the input molecules, the choice of dimensionality reduction techniques will also bias the resulting 2-D visualization, possibly obscuring or exaggerating key structural features of the high-dimensional space. To assess the extent to which the choice of dimensionality reduction technique plays a role in projecting the high-dimensional data set into 2-D, we applied three different methods for each of the three high-dimensional databases: Principal Component Analysis (PCA), SKETCH-MAP, and Uniform Manifold Approximation and Projection (UMAP) [99, 104, 106]. The results are shown in Fig. 2.10. What follows is a discussion of each of these dimensionality reduction techniques and how well they can be used to visualize our input data sets. For an in-depth explanation of each of these techniques, we refer you to Chapter 1.

The first dimensionality reduction technique we applied was PCA, a highly popular linear method [99]. Given a high-dimensional input data set, PCA fits the data to the hyperplane that minimizes the least squares error over the whole data set. The vectors tangent to each hyperplane are known as the principal components. An equivalent framing of the method is that the principal components are a set of orthogonal vectors that maximize the variance in the high dimensional data set. In practice, this means that the principal components are found by calculating a covariance matrix over the entire data set and then performing an eigenvalue decomposition on this matrix. The resulting eigenvectors are the principal components and their corresponding eigenvalues denote the fraction of the total variance in the data set that is projected onto that eigenvector. The eigenvalues and eigenvectors are decreasingly ordered by the magnitude of the eigenvalues. Therefore, projecting the data onto the first principal component will give a 1-D plot maximizing the variance of the data as much as possible, projecting the data onto the first two components will do the same but projected onto a plane, and so

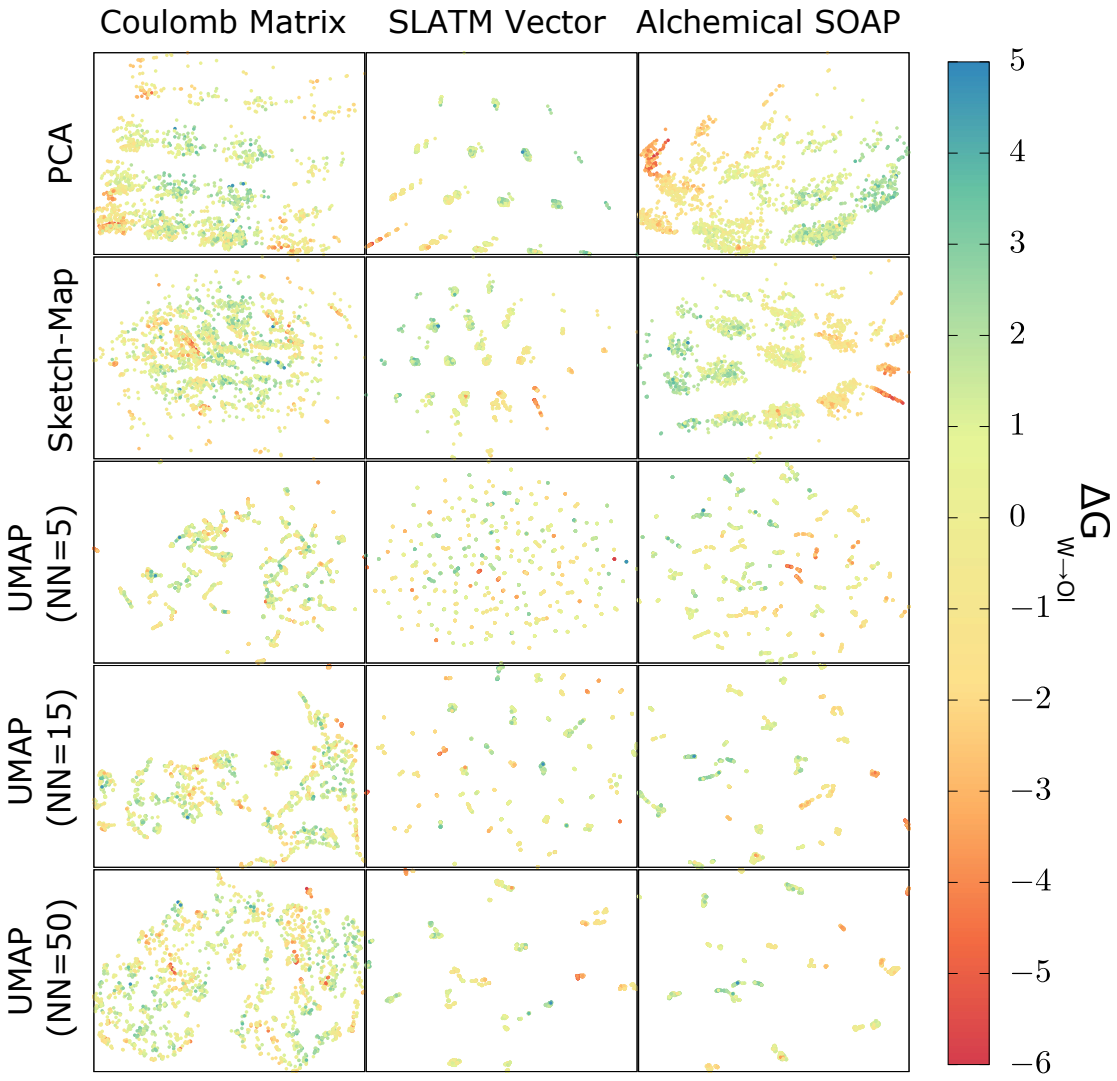

Figure 2.10: Plots showing the molecular fragment data set used in this work projected into 2-D using a variety of different dimensionality reduction techniques (corresponding to rows in the grid) after encoding the fragments using three different molecular representations (corresponding to columns in the grid). The molecules are colored based on their predicted $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ value (in kcal/mol) from ALOGPS.

on. Fig. 2.10 shows PCA plots for each of the three representations, with points colored by their predicted $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ value using ALOGPS. The amount of variance captured by the first two principal components was approximately 40% for the Coulomb Matrix and approximately 70% for both the SLATM and ASOAP data. While the Coulomb Matrix data set seems to show rough clustering behavior, none of these clusters strongly correlate with $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$. On the other hand, both SLATM and ASOAP appear to correlate nicely with $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, with well-defined clusters. This is somewhat expected, as both of these representations encode local environments which are defined from the perspective of individual heavy atoms. However, the SLATM data exhibits much cleaner separation between clusters compared to ASOAP. This is likely due to the composition of the SLATM vector, which is structured into separate spectra for each type of 2-body and 3-body interaction (for example all many-body interactions that contain an Oxygen will be set to zero if a molecule doesn't have O in it). Because we use an alchemical kernel in addition to the regular SOAP kernel, this separation due to the discretization of different interaction types seen in SLATM becomes blurred, which is why the ASOAP clusters are not as clearly defined. Because both ASOAP and SLATM correlate strongly with $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, one might naively assume that the clustering seen in the PCA plots might correspond to the Martini bead types. However, Fig. 2.10 clearly shows that this is not the case. All of the apolar fragments, which would map to five different apolar Martini bead types, are found within one or two clusters only. Similarly, the nonpolar and polar molecules, which would map to the remaining 9 bead types, dominate the populations of the remaining clusters, with some of these clusters containing roughly equal numbers of non-polar and polar molecules. This motivates a data-driven approach for optimizing top-down coarse-grained models for chemical transferability, which we discuss further in Chapter 3.

The inherent assumption when using PCA to visualize in 2-D is that the underlying manifold from which the data is sampled in the high-dimensional space is a plane, which may not be the case. Many projections of CCS exhibit highly nonlinear behavior, as a small perturbation in chemical composition (for example replacing a C with an N) can result in disproportionately large changes in the resulting properties. The molecular representations studied in this work may also contain some of this nonlinear character, meaning that PCA would not accurately represent the data in 2-D. Therefore, we decided to apply SKETCH-MAP, a nonlinear dimensionality reduction technique, to our data sets [104]. Specifically, SKETCH-MAP is essentially a nonlinear version of metric multidimensional scaling, which aims to preserve the high-dimensional distances in the data set when projecting to lower dimensions. Note that, if the euclidean distance is used, this is equivalent to PCA. Instead, SKETCH-MAP nonlinearly transforms the high-dimensional

distances using the following sigmoidal function:

$$F(R) = 1 - (1 + (2^{A/B} - 1)(R/\sigma)^{A})^{-B/A} \tag{2.13}$$

where R is the high-dimensional distance, and σ, A, and B are fitting parameters of the sigmoid that are specified by the user [104]. The choice of σ is especially important, as distances much smaller than σ are set to 0 while distances much larger than σ are set to 1. The design philosophy behind implementing this sigmoidal function is to prioritize the preservation of the distances between clusters of data, which should correspond to the chosen σ value, rather than preserving intra-cluster distances, which would tend to 0 after applying the sigmoid function. In this way, SKETCH-MAP aims to better represent the global structure of the data.

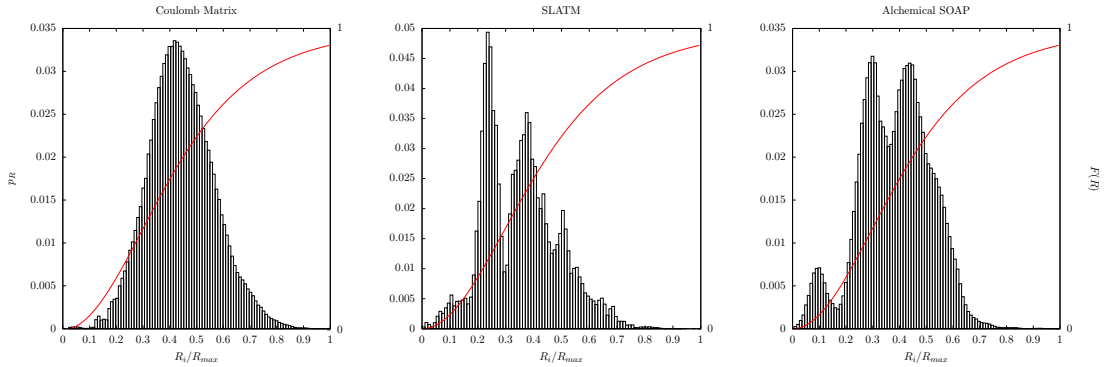

Figure 2.11: Histogram of Euclidean distances (or dot-product distances for ASOAP) in the high-dimensional space created when expressing the data set used in this work in terms of their Coulomb Matrix (left), SLATM (middle), and ASOAP (right) representations. Additionally, the red line in each case is the sigmoidal function used to transform these distances using SKETCH-MAP.

Fig. 2.11 shows the distribution of Euclidean distances between all data points for each of the three molecular representations. Also shown are the sigmoidal functions that transform these high dimensional distances for each of the corresponding SKETCH-MAP plots shown in Fig. 2.10. The SKETCH-MAP plots show similar qualities to those observed from the PCA plots. The weakly defined clusters seen in the Coulomb Matrix data set are further blurred by applying the sigmoidal function, and the lack of global correlation with $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ is similarly exacerbated. On the other hand, the clustering seen in the PCA plots is mostly reproduced for the SLATM and ASOAP cases, with the clusters appearing slightly more compressed here. This compression effect is likely due to the fact that the intra-cluster distance is much smaller than the σ values of the sigmoidal function, making these

distances tend to zero in the low dimensional output, whereas these distances are linearly scaled when applying PCA.

As an aside, we now discuss the applicability of the "curse of dimensionality", a term coined by Bellman referring to the exponential increase in volume that occurs when increasing the dimensionality of data points and using a Euclidean distance metric [180]. Aggarwal et al. confirmed this effect, demonstrating that the ratio between a data points nearest neighbor and farthest neighbor approaches 1 as dimension increases, meaning that all data points are roughly equidistant when using a Euclidean distance metric [181]. The curse of dimensionality can therefore be a major barrier to identification of clusters in a high-dimensional space. However, the results of these works assumed that the input data set had a uniform distribution, corresponding to a unimodal distribution of distances, becoming narrower as dimension increased. Remarkably, Fig. 2.11 shows that only the Coulomb Matrix falls into this trap, with a single unimodal distribution of distances despite having the lowest dimensionality of the three representations used by two orders of magnitude. Using both SLATM and ASOAP, however, there appear to be multiple peaks present, suggesting that, even with dimensionalities of 6694 and 2035, respectively, there is clustering occurring. The plots of the data projected into 2-D seen in Fig. 2.10 also show a significantly higher degree of clustering when using these two representations as compared to the Coulomb Matrix. Bennett et al. showed that, even for high-dimensional data, multimodal distributions in the Euclidean distances between points indicate that the curse of dimensionality does not apply [182]. Essentially, if there is a high pairwise cluster stability, meaning that the distance between points within a cluster is drastically reduced compared to the inter-cluster distance, the euclidean distance remains useful. Houle et al. further noted that this effect will occur if there is a high degree of contrast between the "important" and "unimportant" dimensions in the data set [183]. This points to the hierarchical nature that is implicitly built into both the SLATM and ASOAP representations. The stochiometric contribution to the representation shows the most variance, causing molecules with similar chemical composition to first be clustered together, with the geometrical differences appearing within individual clusters. For example, the 1-body terms in the SLATM vector show the greatest variance due to the fact that they are simply a sum of the total number of valence electrons for each atom type. This means that they will either have values of 0 if the specific atom type is not present in the molecule, or they will be equal to the number of valence electrons times the number of the specific atom type present in the molecule. On the other hand, the $1/r^6$ and $1/r^9$ scaling applied to the 2-body and 3-body terms dictates that the variance of these spectra is significantly reduced by comparison, ensuring that major clusters will be defined based on chemical composition and intra-cluster distributions will depend on the

internal geometry of the molecule.

One of the major limitations of the SKETCH-MAP approach is that it assumes a single length-scale, and thus a single σ value, is necessary to separate clusters in the high-dimensional data set. This means that SKETCH-MAP is not useful for visualizing data with multiple inter-cluster length-scales. For example, the compression effect seen in the SKETCH-MAP plots in Fig. 2.10 makes it difficult to deduce how many sub-clusters there may be in each cluster, and several SKETCH-MAP plots may be necessary, each with different σ values, in order to ascertain the localized structure of the data within a cluster. As an alternative to this approach, we applied UMAP as our third and final dimensionality reduction method in this work [106]. UMAP constructs a fuzzy topological manifold using the high-dimensional data, assuming that the data can itself be reduced to a set of locally connected topological sets, each with its own distance metric. This distance metric is determined by fitting a Gaussian function to each point as well as its k-nearest neighbors, where k is a user-specified input parameter. After constructing this high-dimensional topological manifold, the algorithm aims to construct a corresponding low-dimensional manifold and minimizes the cross-entropy between the two. The data is then embedded into 2-D using the optimized low-dimensional manifold. Because the definition of distance varies across the high-dimensional manifold, UMAP is highly successful for visualizing local clusters of similar data points. However, for the same reason, the degree to which UMAP can approximate the global structure of the high-dimensional data cannot be easily assessed using a global metric. Two factors affect this: (*i*) the number of nearest neighbors considered when defining local distance, with more neighbors resulting in the preservation of more global structure, and (*ii*) the optimization of the cross-entropy is not a convex problem, but is solved using stochastic optimization methods. Therefore, even when using a large k value, it is still possible to get stuck in a local minimum. These two effects result in well defined clusters that can seem randomly placed in comparison to each other when using UMAP.

To demonstrate this, we first vary the number of nearest neighbors considered by UMAP when constructing the locally connected sets used to define the high-dimensional manifold. The results are shown in the bottom three rows of Fig. 2.10. First, for $k = 5$, we see many small disconnected clusters. This is especially true for SLATM, where many single points or pairs are scattered over the plot. For both SLATM and ASOAP, as k increases from 5 to 50, many of these points are coalesced into fewer clusters overall. On the other hand, even with $k = 5$, the CM plot suggests large amounts of clustering. However, this indicates that all of the points are somewhat uniformly distributed in the high-dimensional space, and each individual point has much more than 5 points that are all equidistant to it. This also agrees with the distance distribution seen in Fig. 2.11.

We further investigate the extent to which each of these dimensionality-reduction methods preserve the global structure of the data by calculating the joint probability density functions for the distances in high-dimensional spaces versus the distances in 2-D, shown in Fig. 2.12 [105]. Unsurprisingly, PCA does the best job at preserving the distances, despite the low percentage of the variance captured for each representation. SKETCH-MAP systematically underestimates the 2-D distance, but maintains the trend of large distances remaining large even after reducing the dimension. As expected, there is very little correlation between the high-dimensional and 2-D distances when using UMAP, as, by construction, UMAP defines a different distance metric for each point in the high-dimensional space. However, the fact that the improvement is only slight despite increasing the number of nearest neighbors by a factor of 10 may suggest that the algorithm was caught in a local minimum when optimizing the lower-dimensional manifold.

This poses a problem when using UMAP. The probability density functions for PCA and SKETCH-MAP shown in Fig. 2.12 allow us to trust that the overall global structure of the data is being captured when using these methods. Since UMAP emphasizes local clustering rather than global structure, a different approach is needed for validation of the UMAP. One possible means to accomplish this could be to find clusters of points in the high-D space and see how well UMAP reproduces these clusters in the low-D space.

In order to test this hypothesis, we performed a clustering analysis on the SLATM data set. Both Figs. 2.10 and 2.11, indicate that this data set consists of well-separated clusters in the high-dimensional space. We use a clustering algorithm that relies on a hierarchical density-based approach, called HDBSCAN originally developed by Campello et al. [97, 98]. For a full description of how the HDBSCAN algorithm works, please refer to Chapter 1, as we provide only a cursory description here. The data is treated as a connected graph with data points representing nodes. The edges connecting these nodes are weighted according to a localized distance metric that depends on the nearest neighbor distances for each point in the data set. Rather than take a single cut-off length-scale or cut-off density as input, HDBSCAN requires the size of the smallest possible cluster to be defined in addition to the number of nearest neighbors accounted for when reweighting the graph edges. A dendrogram is then calculated that spans the entire data set, and the stability of clusters is determined by how "long-lived" they are as the furthest points from the cluster center are systematically removed until the minimum cluster size is reached. This "lifetime" metric essentially answers the following question: if the furthest assigned data point were removed, would the remaining data set still be considered a single cluster, or would it have to be split into separate clusters? The final clusters that are chosen are those that are the most stable under this criterion. We applied HDBSCAN as implemented

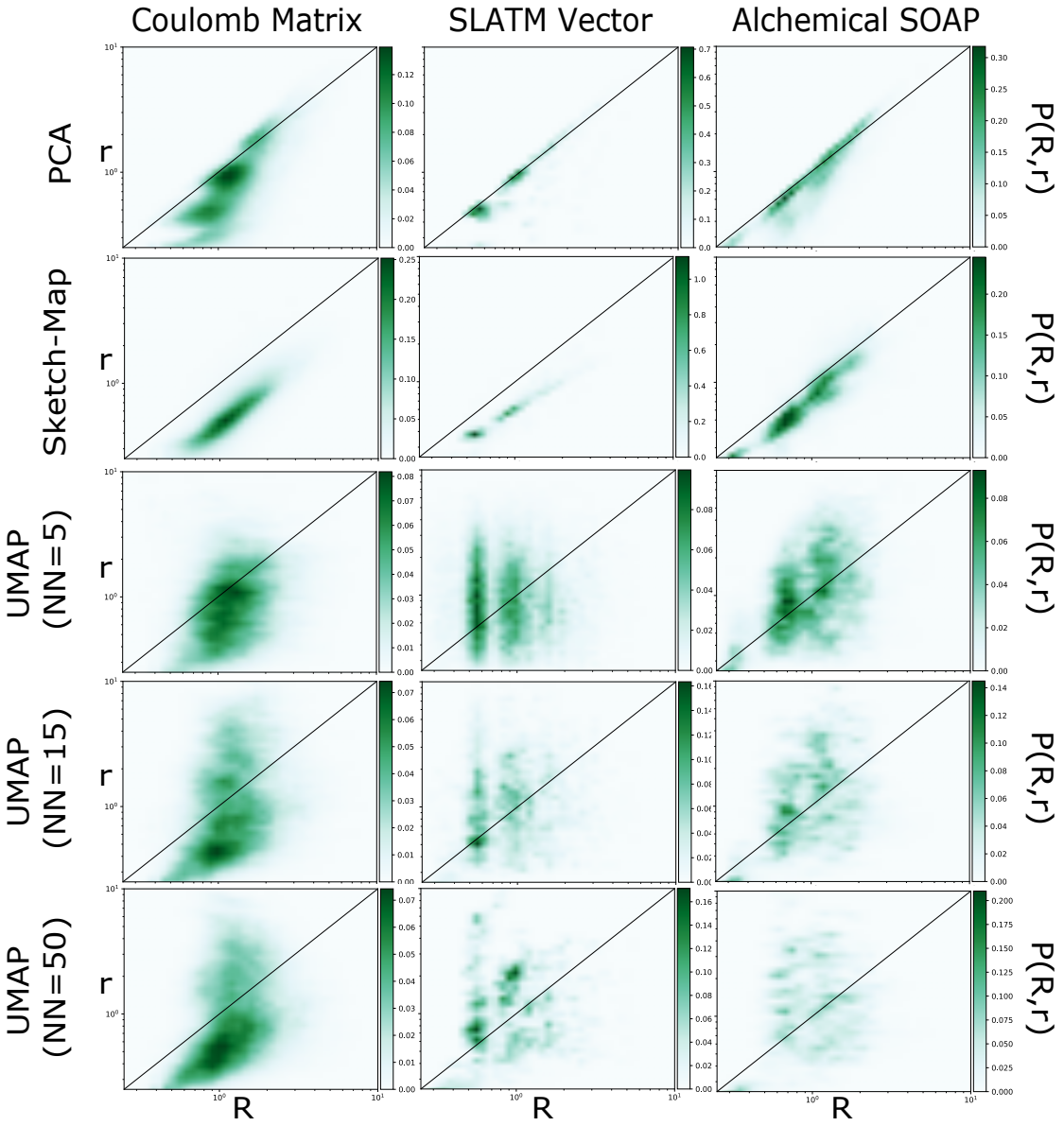

Figure 2.12: Smoothed 2-D histograms portraying the joint probability density functions between the high-dimensional (R) and low-dimensional (r) Euclidean distances for each of the dimensionality reduction techniques and molecular representations used in this work. The distances are scaled for each method such that the maximum distance is 10. The black line denotes $R = r$.

in PYTHON 2.7 to the high-dimensional SLATM data set, as well as all of the 2-D plots of the SLATM data set, setting the minimum cluster size to 5 points for all cases. As previously discussed, the curse of dimensionality does not seem to play a significant role when using the SLATM representation. In addition, the Mutual Reachability Distance metric used by HDBSCAN to initally transform the data implicitly takes into account the shared nearest neighbor distances between points, which has been shown to be a well-performing metric in high-dimensional spaces [183]. The number of clusters found in the high-D data set was 176. Unsurprisingly, both PCA and SKETCH-MAP showed far fewer clusters with 138 and 132 clusters found, respectively. Interestingly, the number of clusters found when using UMAP varied significantly as a function of the number of nearest-neighbors considered, from 195 when $k = 5$ to 140 when $k = 15$ and 103 when $k = 50$. This suggested a means of choosing the k parameter such that the number of clusters would match what was given from the high-dimensional data. Indeed, we found that choosing $k = 9$ resulted in a UMAP with the number of clusters being equal to 175, a promising match to the high-dimensional result of 176. We then compared the cluster assignments between the high-dimensional data set and this new UMAP. Out of the 176 clusters found in the high-dimensional space, 160 of these were perfectly reproduced in the 2-D space. For 7 of these remaining 16 clusters, over 70 % of the cluster was correctly assigned. This means that approximately 95 % of the clustering was preserved when using UMAP to construct the 2-D space, with only 155 out of 2035 points incorrectly labeled as noise or as belonging to a different cluster. This shows that using UMAP in combination with a clustering scheme like HDBSCAN provides a means to correctly tune the parameters used for the visualization and assess the validity of the result by using the number of clusters, rather than distance, as a global metric. However, our success in applying this clustering approach was also due to the use of the SLATM vector, which demonstrated strong clustering behavior even in the high-dimensional space. We recommend a similar level of investigation (i.e., high-dimensional distance distributions, preliminary linear dimensionality reduction plots) to assess the quality of other high-dimensional data sets when using this approach. We further note that, by increasing the minimum cluster size input parameter used by HDBSCAN, it is possible to identify the larger clusters seen in the PCA and SKETCH-MAP plots, though the information regarding the sub-clusters is consequently lost. Overall, these unsupervised machine learning methods are highly useful tools for determining what underlying structure (if any) exists in a given data set and obtaining a more intuitive sense of the data.

The contrast between UMAP and the other two methods of dimensionality reduction, PCA and SKETCH-MAP, is similar to the contrast between the bottom-up and top-down approaches used in another dimensionality-reduction technique: coarse-

graining. The top-down approach seeks to develop a CG model that reproduces certain macroscopic properties. For example, the Martini model aims to correctly model the hydrophobicity of molecules over a wide range of CCS. Similarly, both PCA and SKETCH-MAP emphasize the global structure of the high-dimensional data, and show a global trend with respect to a macroscopic property, $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, for both the SLATM and ASOAP representations. Therefore, dimensionality reduction techniques that emphasize global structure are a good starting point for coarse-graining CCS in a top-down manner. Fig. 2.10 shows that the Martini bead types do not seem to match the clusters shown when using PCA or SKETCH-MAP. This suggests that a data-driven top-down coarse-graining approach using $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ may result in a partitioning of the CCS that is more efficient than Martini for exploring CCS. We apply this data-driven optimization and determine its effect on coarse-grained screening efficiency in Chapter 3. Bottom-up coarse-graining, however, takes a higher-resolution model and coarsens it by removing extraneous degrees of freedom. In this analogy, the higher-resolution model is the high-dimensional data set taken as input. UMAP is then able to visualize the localized clusters that are identified in this space, which is validated using HDBSCAN. In this work, we were able to reduce the data from 2035 input fragments to 176 representative clusters, potentially paving the way for a bottom-up approach to coarse-graining CCS. We validate this approach in Chapter 4, in which we construct a chemically-transferable bottom-up coarse-grained model using the unsupervised machine learning methods described here.

2.8 Predicting partition free energies using SLATM

Both PCA and SKETCH-MAP plots seen in Fig. 2.10 demonstrate a clear correlation with $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ when using the SLATM vector. In order to further investigate this correlation, we parameterize a KRR model trained on experimental data and compare it's accuracy to that of the ALOGPS neural network. The experimental data was obtained from the National Cancer Institute (NCI) database and was restricted to molecules with twenty or fewer heavy atoms, resulting in 2324 molecules total [184].The data itself consisted of the SMILES strings of each molecule as well as the log of the water/octanol partition coefficient, $log_{10}P$. Note that this is separate from the log of the membrane permeability coefficient discussed previously, and it is related to $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ by the following Arrhenius-type relation:

$$\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}} = \frac{1}{\beta}\left(\frac{\log_{10} P}{\log_{10} e}\right) \tag{2.14}$$

where k_B is the Boltzmann constant and the temperature T is set to 300 Kelvin. After generating 3-D structures from an input SMILES string, the molecules were

then converted into molecular SLATM representations using the QML package. Just as was done for the prediction of partition free energies, as seen in Eqn. 2.11, we used a Laplacian kernel with the same σ and λ values (although these were also optimized in a separate grid search) but with a Euclidean L_2 norm rather than the city-block metric used in the earlier work. We then constructed the learning curve shown in Fig. 2.13, with the error bars denoting the standard deviation corresponding to the distribution of predicted mean absolute errors when applying 10-fold randomized cross validation. The resulting model showed an MAE of 0.32 log units with a standard deviation of 0.03 log units, corresponding to an error in $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ of 0.44 ± 0.05 kcal/mol. The ALOGPS neural-network-based program was also used to predict the error of the same experimental data set, and an MAE of 0.24 log units was obtained, corresponding to an error in $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ of 0.33 kcal/mol, only slightly outperforming our model.

We now provide a description of the ALOGPS program so as to better compare the two models tested. ALOGPS uses an associative neural network approach to predict $\log_{10} P$ [163, 175]. Given an input SMILES string, a 3-D conformation of the molecule is constructed, and the input vector consists of the number of non-Hydrogen atoms, the number of Hydrogen atoms, and then a series of 73 electrotopological state (e-state) indices, which were initially developed by Hall and Kier [185]. These e-state indices can be subdivided into atom-type and bond-type indices, which take into account the electronic properties, topology, and geometry of the molecule as it pertains to a specific atom or bond. For the calculation of atom-type indices, each atom is given an intrinsic value, which is the ratio of valence to sigma orbital electrons while covalently bonded in the input molecule. The e-state index for the atom is then given by summing over all pairwise differences in intrinsic value between that atom and every other atom, divided by the squared distance between the two atoms. An intrinsic bond value for a bond between two atoms is calculated by averaging the intrinsic values of the atoms that make up the bond. The bond-type e-state index is then obtained by doing a similar pairwise-sum over bonds as was done for the atomic indices. This input feature vector is then fed into an ensemble of 64 dense, feed-forward neural networks, with each network being trained on a different portion of the total data set. For a new input molecule, the output property is then given by performing a weighted average over the output of each neural network such that the variance of the individual networks is accounted for. Each neural network consisted of a single hidden layer of only 5 neurons. Therefore, the total parameter space for the entire neural network ensemble was 24,384 weights and biases in total. The network was trained on half of the 12,908 molecules randomly selected from the PHYSPROP database, with the remaining half used as a test set [186]. The resulting MAE was found to be 0.26 log units. Unfortunately, we were unable to find this specific

version of the PHYSPROP database, and so were unable to test the accuracy of our KRR model on this data set.

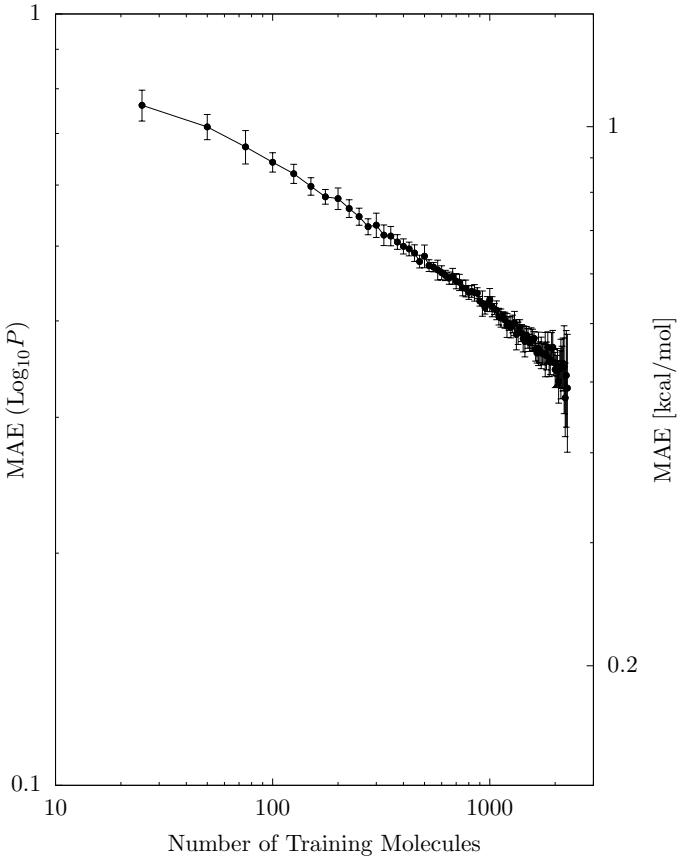

Figure 2.13: Learning curve corresponding to the KRR model that used the SLATM representation to predict the log of the water/octanol partition coefficient. The error bars correspond to the standard deviation when running 10-fold randomized cross validation.

Remarkably, our model achieves an accuracy within 0.11 kcal/mol of ALOGPS despite being trained using a data set roughly 1/3rd the size of that used to train ALOGPS. Furthermore, the number of parameters optimized when training the KRR model is equivalent to the number of molecules used in training, 2324, which is approximately a full order of magnitude less than the number of parameters optimized in the neural network approach. On the other hand, the dimensionality of the input feature used in the KRR model is approximately two orders of magnitude greater than the 75 input features fed into the neural network. This speaks to the fundamental differences between the KRR approach versus the neural network

approach. The KRR model is successful due to the fact that the SLATM vector already encodes much of the relevant information, minimizing the amount of data needed for the model to attain high predictive power. On the other hand, the high parameter space of the associative neural network used by ALOGPS, as well as the increased training set size, allows for the network to "learn" an optimal representation before predicting the target property. Notably, both our KRR model and ALOGPS only require a single conformation of the input molecule to predict a thermodynamic property that is normally obtained by averaging over many conformations. This may indicate that the conformations chosen were already very close to the equilibrium conformations for the given molecules. Further testing to see how including a conformational average rather than a single conformation, as was recently done by Rauer et al., is necessary [187].

2.9 Hierarchical Screening

Fig. 2.10 shows that clustering is a viable means to hierarchically screen compounds that may all map to a single Martini representation. Using the SLATM vector or the ASOAP kernel, enables a finer exploration of CCS with respect to $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ when compared to Martini. However, since Martini already accounts for $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, a second level of screening using the same property may be redundant, and other properties, such as the molecule's size and shape may be of greater interest at this point. While the Coulomb Matrix didn't correlate well with $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, it was the cheapest out of all representations to generate, with the lowest dimensionality of the representations investigated. In order to determine whether or not this type of representation could be applicable for screening based on molecular size and shape, we applied a modified version of the Coulomb Matrix that did not account for atom type (i.e., setting all atom types to C) to only encode the internal geometry of the molecules in terms of their pairwise distances. Note that this is essentially equivalent to the Weyl Matrix of pairwise distances [118]. Fig. 2.14 shows the application of this representation to the set of GDB molecules with 7 heavy atoms that all mapped to a single Martini bead, numbering 2177 molecules in total. The 28 dimensional space was projected into 2-D using SKETCH-MAP and HDBSCAN was used on the high-dimensional data to identify the largest clusters, with the minimum cluster size now set to 15. The clustering analysis revealed that each cluster corresponded to a particular molecular scaffold, as shown in Fig. 2.14. Since molecules that mapped to every Martini bead type are found in every cluster, filtering first by bead type and then by molecular scaffold could be an effective means to screen compounds. Rather than exhaustively sample all of the compounds mapping to a single Martini representation, representative molecules from each cluster can be sampled. The clusters to

which the best performing molecules are assigned could then be further clustered using a different representation that highlights other relevant properties, and so on, until a reasonable number of test compounds is achieved. If coupled with a back-mapping scheme similar to those recently proposed that rely on generative adversarial neural networks, the sampling of these clusters could be integrated into an online high-throughput workflow [188]. This hierarchical screening approach is currently being tested in our group and will be the subject of a future work.

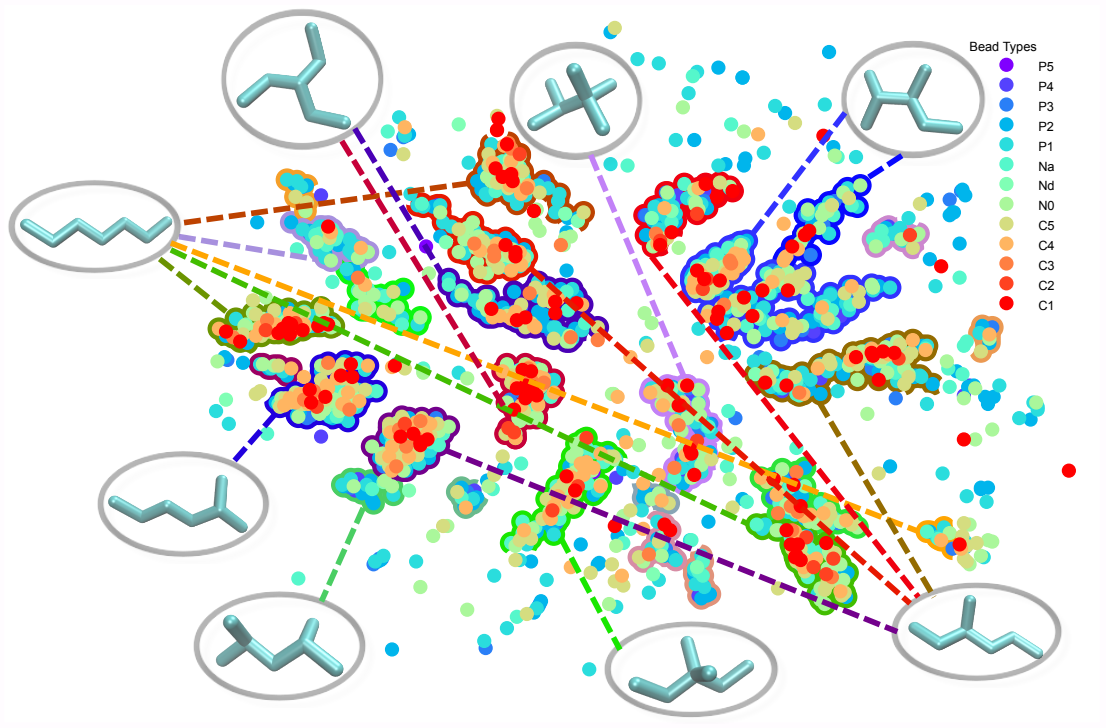

Figure 2.14: Sketch-map of molecules corresponding to Martini unimers with 7 heavy atoms. The molecules have been expressed as Coulomb Matrices, but without accounting for atom type (essentially the same as a Weyl matrix). Each point is colored by its Martini bead type. The boundary colors correspond to clusters identified using HDBSCAN. These clusters correspond to different molecular scaffolds, which allow for different scaffolds to be sampled when performing hierarchical screening.

2.10 Conclusions

In this chapter, we first introduced the HTCG approach as a means to quickly construct structure–property relationships that span CCS. By applying this method using the top-down Martini force field, we were able to identify linear relationships between key thermodynamic state points when modeling the behavior of small molecules in a lipid bilayer membrane environment. A single, easily-accessible parameter, $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ is the only required input in order to predict the transfer free energies between these state points. We extended this structure–property relationship by introducing a second descriptor, the acidity, onto which we could then project the coarse-grained permeabilities for all Martini unimers and dimers. We then demonstrated that further exploration of coarse-grained CCS, corresponding to Martini trimers and tetramers, was possible by implementing a Monte-Carlo scheme that used alchemical transformations to construct and optimize thermodynamic cycles that efficiently sampled the CG compound space. A KRR model was then trained on these results to further expand the transferability of these structure property relationships. In implementing the HTCG approach, we also demonstrated a drastic reduction of CCS when coarse-graining using Martini due to the degeneracy of molecules that were mapped to the same Martini representation. Approximately 1.8 million molecules in total were mapped to Martini unimers, dimers, and trimers, using the AUTO-MARTINI algorithm. By performing a functional group analysis on these compounds, we were able to provide a means to implement inverse molecular design when targeting a specific membrane permeability. Next, we assessed three different molecular representations as well as three different dimensionality reduction techniques in order to determine whether unsupervised ML could provide a means for further screening in a hierarchical manner. We found that PCA and SKETCH-MAP preserve the global structure of the high-dimensional data while the UMAP visualizations consisted of well-separated clusters which were randomly placed in relation to each other. Additionally, the SLATM and ASOAP representations were able to relate chemical structure to $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, whereas the Coulomb Matrix did not show a strong correlation to this property. This insight led to the parameterization of a KRR model using the SLATM vector that could compete with the ALOGPS program, although it remains unclear as to why a single input configuration was sufficient to achieve such high accuracy when predicting a thermodynamic property. We showed that even relatively low-dimensional representations, like the modified Coulomb Matrix, could identify molecular scaffolds which could be used in a hierarchical screening approach. Further work is currently underway to apply the HTCG approach for other target properties. Finally, it is clear that the clusters obtained from this unsupervised ML approach do not always correspond to specific Martini bead types. This suggests a more efficient top-down assignment of hydrophobicity values for each bead

type that could be useful for better covering CCS. We detail our approach towards deriving a data-driven version of Martini that more efficiently covers CCS, as well as the inherent limits that arise when relating Martini-like top-down models to specific chemical structures, in the next chapter.

3 Resolution limit of data-driven top-down coarse-grained models spanning chemical space

In the previous chapter, we put forward two methods for coarse-graining chemical compound space (CCS), analogous to two different dimensionality reduction techniques. We saw that, with the correct choice of representation and dimensionality reduction technique, a clear trend was seen with respect to a macroscopic thermodynamic property, $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$. The chemical structure of the molecules, and specifically the presence of certain hetero-atom substitutions were primarily responsible for the overall clustering, in addition to dictating the partitioning behavior of these compounds. This overall correlation between the encoded chemistry and the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ suggested that even projecting the CCS onto the 1-D $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ axis would still preserve sufficient chemical information, such that dominant chemical motifs could be easily inferred from a given $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ value. In this chapter, we expand on the idea that these global dimensionality reduction techniques can be likened to a top-down coarse-graining of CCS because both aim to capture the global structure of an input data set of compounds.

Although the data set of compounds in the previous analysis corresponded to fragments which mapped to Martini dimers, the clustering observed did not neatly correspond to the Martini bead types, with several apolar bead-types found predominantly in two clusters. This was somewhat expected, as the Martini model was not designed to optimally represent CCS projected onto a descriptor quantifying hydrophobicity, like $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$. This observation, along with the insights highlighted in the previous chapter, motivate us to develop multiple top-down coarse-grained models that more effectively represent CCS when projected onto $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$. We use a data-driven approach that allows us to vary the number of bead types in each model, corresponding to varying resolutions in $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ space. This enables a hierarchical-screening approach from a top-down perspective, with models of higher resolution used to screen increasingly narrower regions of CCS with similarly hydrophobic compounds. While increasing the resolution of these models can provide a stronger indication as to the chemical structures that give rise to the desired properties, we demonstrate that this top-down approach

is limited by the use of a single descriptor, with negligible increases in screening efficiency as the number of bead types increases.

This chapter has been previously published as the following research article listed below. The article is reproduced here with kind permission from the other authors and the Journal of Chemical Physics which published this work.

Kiran H. Kanekal and Tristan Bereau
Resolution limit of data-driven coarse-grained models spanning chemical space
The Journal of Chemical Physics 151:164106, 2019.
DOI: 10.1063/1.5119101

We again follow the convention of the previous chapter of using a * symbol to specify which sections are taken from the above publication whereas sections that are unique to this chapter have no such symbol.

3.1 Introduction *

Molecular design is a cornerstone of materials science, requiring a fundamental understanding of the relationships between molecular structure and the resulting properties. Traditionally, these structure–property relationships[189] only arise after multiple rounds of screening and discovery of new materials [190–194]. These screening approaches constitute examples of direct molecular design, in which the space of all chemical compounds, known as the chemical compound space (CCS), is explored to determine the most suitable chemistry for the target application. Direct molecular design can be interpreted as projecting a hypersurface in the high-dimensional CCS onto a lower dimensional space defined by certain key molecular descriptors that strongly correlate with the desired property. In contrast, inverse molecular design, in which a structure–property relationship is used to infer a suitable chemical structure from a desired property, remains a "holy grail" of materials science. The main obstacle to achieving this goal is the inability to quickly establish structure–property relationships that can span broad regions of CCS. This is an exceedingly difficult task, given that the size of CCS was estimated to be 10^{60} for drug-like molecules less than 500 Da [4]. Experimentally, this process is inhibited due to both the material and time cost associated with synthesizing and testing a large variety of chemistries that are necessary to infer a relation that is both robust and accurate enough to enable inverse molecular design.

Computationally, recent advancements in processing power and in machine learning have enabled several efficient methods for estimating the electronic properties of a large variety of materials [195–200]. These methods have the added benefit of

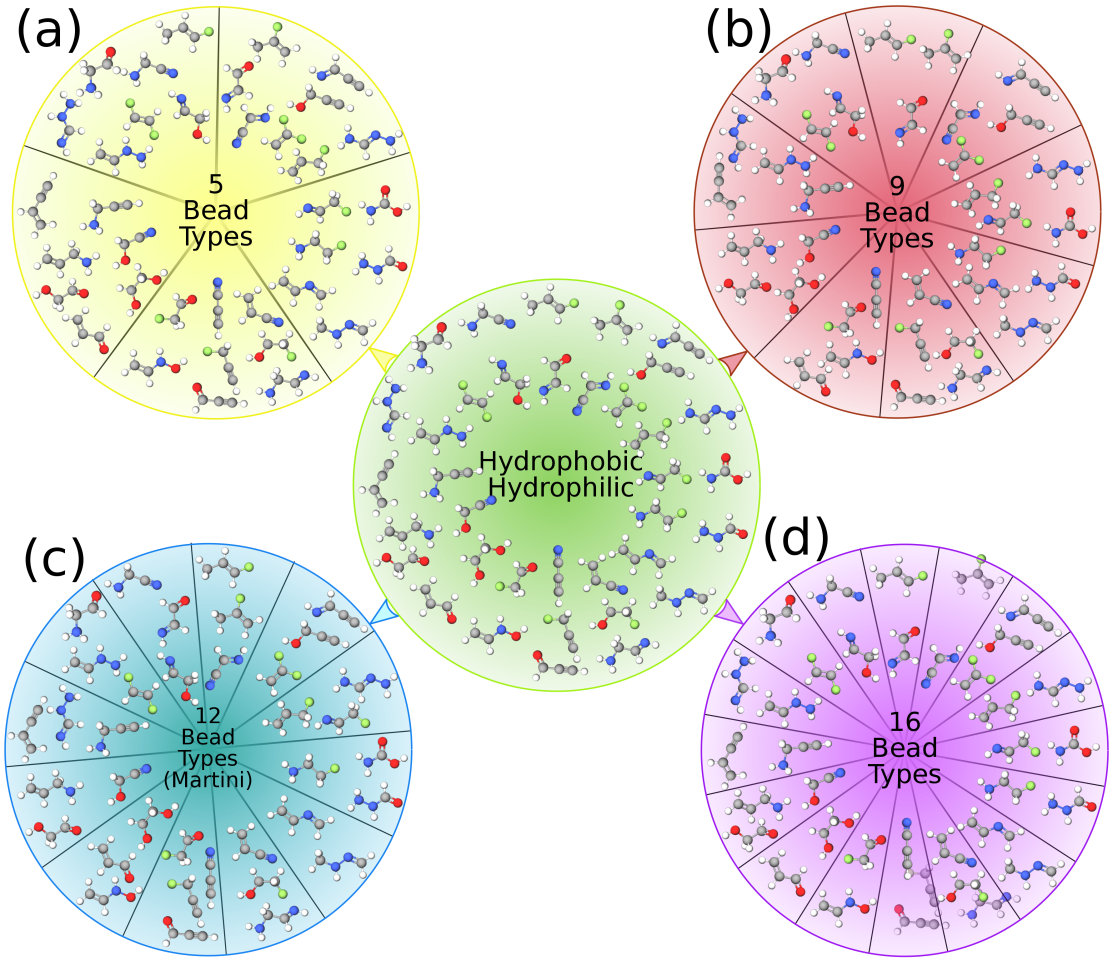

Figure 3.1: A cartoon schematic showing the projection of CCS onto the hydrophobicity descriptor $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, allowing for the creation of top-down chemically-transferable coarse-grained models with a) five, b) nine, c) twelve, and d) sixteen bead types. The number of bead types included in these models defines the degree to which CCS is partitioned on the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ axis. By varying the number of bead types in each model, we obtain greater insight as to the range of chemistries spanned by a single bead type.

screening molecules that cannot be easily synthesized, and can thus motivate (or demotivate) the experimental exploration of these chemistries. However, there has been relatively little success in applying computational high-throughput screening methods to determine the stability of chemical compounds in soft-matter systems for which thermal fluctuations play a critical role [201, 202]. Force-field based methods, such as molecular dynamics simulations, are typically used to account for the immense number of configurations that result from thermal fluctuations in these systems. Unfortunately, due to the extensive computational resources required, a high-throughput scheme based on atomistic molecular dynamics simulations is currently unfeasible for spanning the large regions of CCS needed to obtain broadly applicable structure–property relationships.

Coarse-grained molecular dynamics simulations provide a means to significantly reduce the computational expense relative to fully atomistic simulations while still capturing the relevant physical properties [32, 34, 136, 203]. Coarse-grained representations of molecules result from mapping groups of atoms to coarse-grained "pseudo-atoms" or beads. The governing interactions between beads are determined such that the desired properties of the atomistic system are retained. This usually corresponds to a smoothing of the underlying free-energy landscape, allowing for more efficient sampling. Conventionally, coarse-graining is applied to a single molecule with the goal of efficiently sampling a specific system of interest. The coarse-grained potentials are obtained via one of several possible methods (e.g., iterative Boltzmann inversion[76, 204], force-matching[77, 205]). However these methods are computationally expensive, requiring an initial atomistic simulation that sufficiently explores the underlying free energy landscape of the system of interest [206]. Therefore, adapting coarse-grained molecular dynamics simulations to high-throughput screening of chemical compounds requires flexible yet reliable mapping and force field parameterization methods that do not rely on results from higher-resolution simulations for each compound screened.

The coarse-grained Martini force field has become widely used to simulate biological systems as it provides a robust set of transferable force field parameters by constructing biomolecules from a small set of bead types [142, 144, 207]. The Martini model is a top-down model, which maps an atomistic compound or molecular fragment to a coarse-grained site based on its partitioning between aqueous and hydrophobic environments. In the context of molecular design, the main advantage that Martini provides is its chemical transferability. While the force field was explicitly parameterized for a set of specific molecules, a single Martini bead can represent several different chemistries that share similar oil/water partitioning characteristics. Thus, the main feature captured by the Martini model is hydrophobicity, which can act as a key driving force in the physics of soft-matter systems. Rather than running a single atomistic simulation that yields a

single data point in CCS, a Martini coarse-grained molecular dynamics simulation provides a representative point in CCS, corresponding to the average behavior of all the chemistries that lay in the region surrounding that point. Thus, high-throughput coarse-grained (HTCG) simulations that use chemically-transferable force fields, such as Martini, are advantageous because they span vast regions of CCS to quickly infer the structure–property relationships and chemical descriptors that can be used to enable inverse molecular design at any resolution. Menichetti et al. recently demonstrated this by running Martini HTCG simulations to construct a structure–property relationship describing the thermodynamics of the insertion of a small organic molecule into a biological membrane across CCS [208, 209]. In doing so, they discovered a linear relationship between the bulk partitioning behavior of the solute and its potential of mean force. They were then able to identify a structure–property hypersurface to obtain membrane permeabilities for these solute molecules. Using the Generated DataBase[73, 161] (GDB), a systematically computer-generated set of organic drug-like compounds, as a proxy for CCS, we then related the regions of this surface to regions of CCS that were dominated by specific chemical moieties, enabling inverse molecular design of small molecules given a desired permeability. The question remains: how representative of CCS is the Martini force field? Given that Martini was designed to reproduce the partitioning behavior of certain solvents as well as the properties of lipid-bilayer membranes, is there a way to accurately parameterize a transferable coarse-grained force field with the goal of optimizing its coverage of CCS? In the context of high-throughput coarse-grained simulations that use Martini, creating a structure–property relationship that enables inverse design requires an understanding of the chemistry that is representative of a specific bead type. The metrics used in assigning specific chemical fragments to Martini bead types mainly consist of several water/oil partitioning free energies, although bulk liquid densities and membrane-specific properties have also been used [89, 210]. Here, we focus specifically on the water/octanol partitioning free energy (although other water/oil partitioning free energies could also be used as they also effectively encode hydrophobicity). Therefore, an intuition for which chemistry maps to a given bead type can only be obtained by understanding how $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ varies as a function of chemistry. Given that the number of heavy (non-hydrogen) atoms that usually map to a Martini bead is around four, we can think of each bead as representing a small carbon scaffold perturbed to some degree by either replacing carbons with other heavy atom types (e.g., oxygen, nitrogen, or fluorine) or by replacing single bonds with double or triple bonds. We define a functional group as being one or a localized combination of these types of perturbations.

In this work, we quantify the information loss that occurs when a top-down coarse-grained model, like Martini, is used to reduce the resolution of CCS. Ad-

ditionally, we parameterize three sets of coarse-grained force fields in the Martini framework. In this context, we use the terms "force field" and "model" interchangeably, defined as a set of parameters which describe the interactions between a fixed number of coarse-grained representations called bead types. Each force field developed in this work consists of five, nine, and sixteen neutral bead types, as well as two extra types to account for hydrogen-bond donors and acceptors. We observe that Martini does not provide the most efficient reduction of CCS. We show that the nine-bead force field reduces CCS to the same degree as Martini despite having three fewer bead types, and that further increasing the number of bead types yields negligible improvements in the performance of the model. The models are validated by performing coarse-grained simulations to calculate the water/octanol partitioning free energies of approximately 500 compounds for which experimental data is available. Finally, we demonstrate that the main advantage of a force field with a large number of bead types is the reduction of uncertainty when back-mapping these coarse-grained representations to real chemical functional groups. Just as decreasing the resolution of the coarse-grained mapping reduces the resolution of the potential energy landscape, a reduction in the number of bead types of a chemically-transferable coarse-grained force field allows for an increased degeneracy of chemical fragments that map to a single bead type, illustrated in Fig. 3.1. Ideally, a well-designed chemically-transferable coarse-grained force field would contain some number of bead types that can be intuitively back-mapped to single chemical functional groups. However, the size of a single functional group is small relative to the size of a Martini bead, such that many functional groups could be identified within a fragment mapping to a single Martini bead. Here, we demonstrate that this mismatch between the size of a Martini bead and a single functional group requires additional constraints in order to identify the unique chemistry that maps to each bead type. Incorporating these constraints into a Bayesian formalism yields probabilities of specific chemistries mapping to a given bead type, further promoting inverse molecular design. However, even these additional constraints allow for the same functional groups to be present in multiple bead types, indicating a natural resolution limit when using $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ as the sole basis for a chemically-transferable, top-down coarse-grained model.

3.2 Methods *

Note that a large amount of data referenced in this chapter is available in a ZENODO repository for download. [211]

3.2.1 The Auto-Martini Algorithm *

This work relies on the AUTO-MARTINI algorithm initially developed by Bereau and Kremer [139]. The algorithm first determines an optimal mapping for an organic small molecule. The mapping provides the number of coarse-grained beads used to represent the molecule as well as their placement. A mapping cost function is minimized for each molecule so as to optimize both the number and placement of beads used in its coarse-grained representation. The assignment of coarse-grained potentials to each bead (bead-typing) occurs by assigning an existing Martini bead type that has the closest matched water/octanol partitioning free energy ($\Delta G_{\mathrm{W}\to\mathrm{Ol}}$) with that of the molecular fragment encapsulated by the bead. The partition coefficients of these fragments are obtained by using ALOGPS,[163, 175] a neural network algorithm that predicts these values given the chemical structure of the fragment.

Several changes were made to the AUTO-MARTINI code in order to increase its accuracy when applied to a large and varied database such as the GDB. The "lonely atom penalty"[139], which weights the effect of leaving single heavy atoms outside the van der Waals radii of the Martini beads, was increased slightly from 0.20 to 0.28. Additionally, the "additivity check" was removed for molecules that map to single beads. This additivity check was designed to ensure that the voronoi decomposition of molecules into fragments and the subsequent selection of bead types for each fragment was sensible (the sum of the $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ values for each bead should be within a cutoff value when compared to the $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ of the entire molecule). This was enacted in order to resolve an issue in which molecules that were meant to be mapped to a single bead (e.g. propanol) were unable to be successfully mapped using the code. The effect of these two changes on the distributions of $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ is shown in Fig. 3.2a and b.

The removal of the additivity check for molecules mapping to single beads caused the gap in the distribution in Fig. 3.2a to no longer appear, meaning that several molecules that would normally map to a single bead were excluded because they failed the additivity check, which should not be applied for single beads. Note that there is a noticeable dip in the coarse-grained distribution of Fig. 3.2. This corresponds to the N0 bead type, which is underpopulated when compared to the corresponding region in the atomistic distribution. We found that this was an artifact due to a cut-off value in the code that caused molecules to be mapped to a donor-acceptor type of bead even if their $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ was closer to the N0 value. By reducing this cut-off value, we were able to obtain the distribution shown in Fig. 3.7, and is also shown in Fig. 3.8c. The final change has to do with the assignment of ring molecules. The standard approach for ring molecules was to use the entire set of atoms in the ring for each fragment and weight each bead's contribution by a scaling factor. For all ring molecules, this was previously set to 2/3 so as

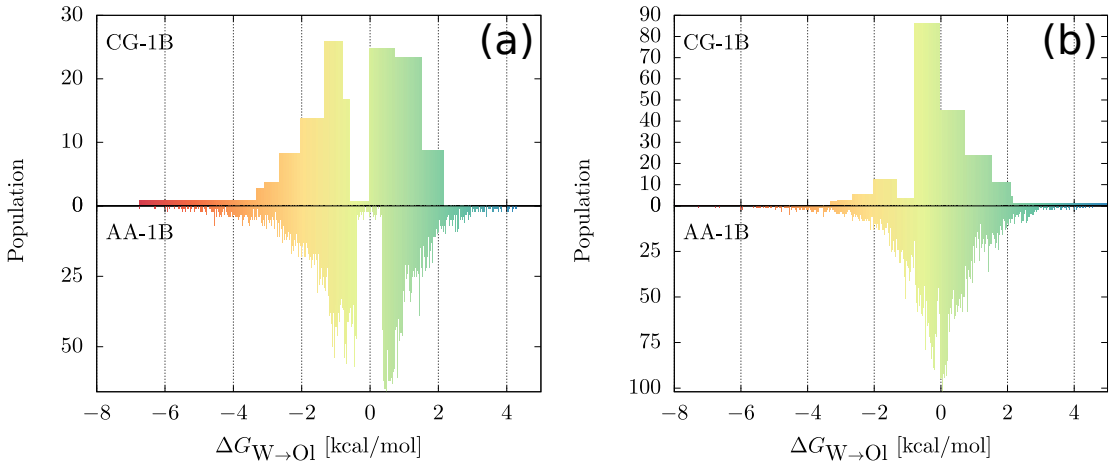

Figure 3.2: Comparison of the $\Delta G_{W\to Ol}$ distributions for molecules mapping to a single Martini bead using the (a) originally published AUTO-MARTINI code and (b) after increasing the lonely atom penalty and removing the additivity requirement for single beads.

to reproduce the Martini parameterizations for benzene and cyclohexane [139, 210]. However, in order to optimize the mappings for the multitudes of ring-containing molecules in the GDB, we found that a factor of 1/2 for 5-membered rings and 1/3 for six-membered rings yielded much better agreement with respect to the ALOGPS predictions for the ring molecules. The results are shown in Fig. 3.3 for both 5-membered and 6-membered rings. Changing these scaling factors resulted in an decrease in the MAE from 1.64 kcal/mol to 0.946 kcal/mol for 6-membered rings and a decrease from 0.893 kcal/mol to 0.807 kcal/mol for the 5-membered rings. All of these updates are included in the latest version of the code which is freely available via a GITHUB repository [212].

Using the refined AUTO-MARTINI algorithm, approximately 3.5 million molecules with ten heavy atoms or less that make up the GDB were mapped to coarse-grained representations for four different force fields. The molecules contain carbon, nitrogen, oxygen, fluorine, and hydrogen atoms only. Of these 3.5 million compounds, approximately 340,000 were successfully mapped to both coarse-grained unimers (1 bead representations) and dimers (2 bead representations) for all of the force fields described in this work. The majority of the remaining compounds were mapped to coarse-grained representations with a higher number of beads, and a small fraction of compounds were unable to be successfully mapped by the algorithm. Histograms comparing the distributions of $\Delta G_{W\to Ol}$ for each set of atomistic compounds mapping to coarse-grained unimers and dimers and their coarse-grained

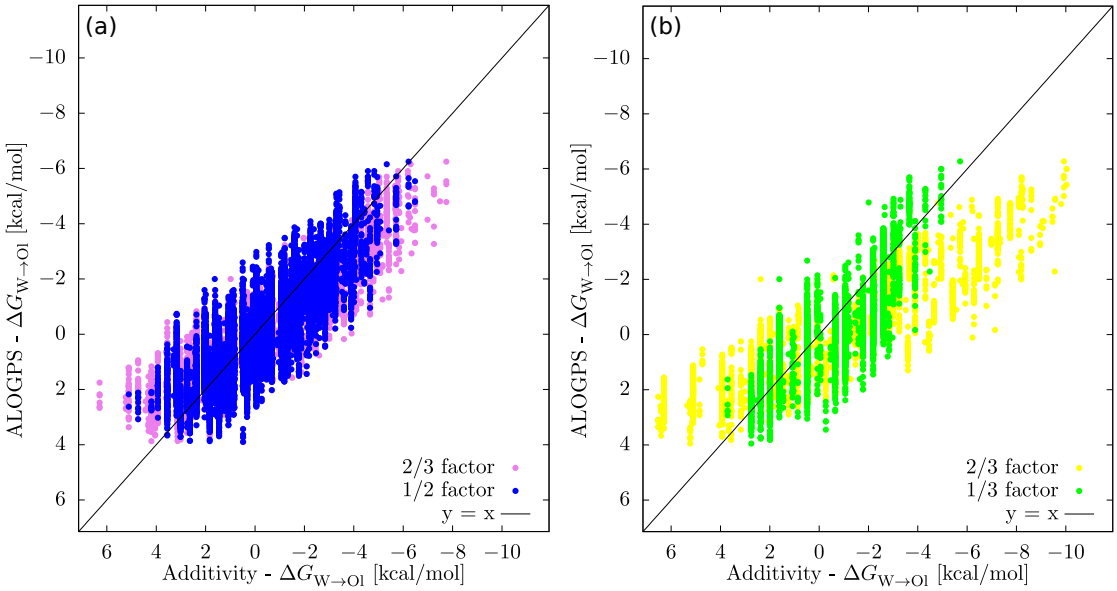

Figure 3.3: Correlation curves showing the agreement between the predicted partitioning free energy values from ALOGPS for ring molecules and the partitioning free energy of the coarse-grained Martini representation assigned by AUTO-MARTINI for (a) five-membered and (b) six-membered ring-containing molecules.

counterparts were constructed using the NUMPY histogram function[213], with the number of bins equal to 1000 and 1050 for unimers and dimers, respectively. These histograms are shown in Fig. 3.7a-d for Martini, and are repeated in Fig. 3.8 along with all of the other histograms computed using the different force fields for easy comparison.

3.2.2 Under-mapping of molecules *

In Fig. 3.4, we show the absolute populations of molecules/fragments that map to unimers/dimers as a function of the number of heavy atoms per fragment. Note that the fragment size distribution is roughly centered between four and five heavy atoms. While a four to one mapping scheme is prescribed by the Martini model for normal sized beads, coarser mappings are also possible. For example, Butanol maps to a single Nda bead in the Martini model, and Octanol maps to only a C1 and a P1 bead. Furthermore, we note that 98% of the molecules/fragments that have six heavy atoms have either a double bond, triple bond, or a branching structure, with most of these fragments having some combination of these molecular features. The presence of any of these features causes a reduction in the radius of gyration and a reduction in the internal degrees of freedom of each fragment, justifying the mapping to a single bead. The remaining molecules/fragments which have more than six heavy atoms make up 14% of the total set of molecules.

Auto-Martini starts by finding a set of optimal mappings for an input molecule ranked by how well they minimize the cost function defined in the original auto-martini paper [139]. If the molecule or fragment includes some of the molecular perturbations mentioned above, it is highly likely that one of the mappings (though not the best ranked one) will correspond to a single bead. The algorithm then tries to find combination of bead types such that the sums of the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ values of each molecule add up to the overall $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ of the molecule as predicted by ALOGPS, within a threshold defined in the aforementioned additivity check. If the algorithm cannot find any combination of bead types that satisfies this check, it repeats this process with the next optimal mapping until it can find a mapping that allows a bead type combination that satisfies the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ criteria. This results in a small fraction of the chemical compound space being under-mapped, as seen in Fig. 3.4. Coarse-graining involves balancing the entropy reduction that comes from reducing the atomistic degrees of freedom by modifying the enthalpic terms (e.g., the potential energy functions) such that the overall thermodynamic properties of interest are preserved. In this case, the thermodynamic property of interest is the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, the AUTO-MARTINI algorithm only maps molecules to coarse-grained representations that best match the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ as determined by ALOGPS, which accounts for the effect of the molecule size when making its prediction. Thus, even if the molecule is under-mapped, the partitioning behavior

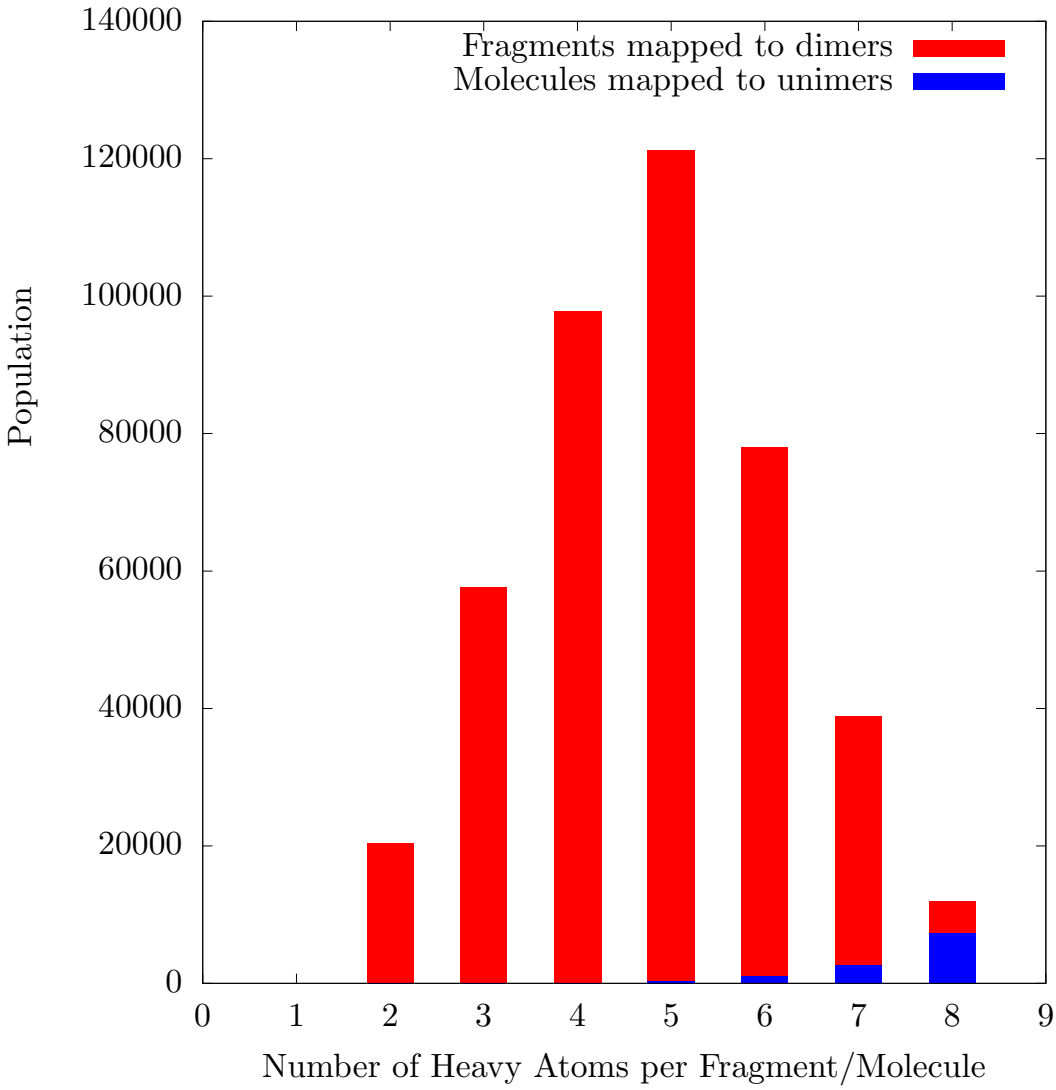

Figure 3.4: Histograms showing the population distribution of molecules (for Unimers) or fragments (for Dimers) mapping to single Martini beads based on the number of heavy atoms in each molecule/fragment.

of the resulting coarse-grained representation is still matched.

The AUTO-MARTINI algorithm also assigns a two-bead mapping for five-membered rings and a 3-bead mapping for six-membered rings. Because we are interested in the region of the CCS that maps to unimers and dimers in this work, we only include database entries for molecules containing up to 5-membered rings. This is consistent with the recommendation of the Martini model, which requires a 2-3 heavy atom mapping for the S bead types. We did not take steps to enforce ring planarity for these 5-membered rings by adding a third bead, as we found that doing so caused large errors in matching $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ because of the inability of ALOGPS to accurately predict the contribution of single-heavy-atom fragments that resulted when trying to map a 5-membered ring to 3 beads. Furthermore, there are issues with the Martini model itself when modeling 5-membered rings using 3 S-type beads. The thickness of the coarse-grained ring structure is significantly larger than that of the atomistic ring. Additionally, it was recently found that the lack of cross-parameterization for interactions between Martini beads of different sizes can lead to errors when calculating partitioning free energies. Due to these concerns, we have again placed more importance on achieving the overall thermodynamic accuracy with respect to the ALOGPS prediction rather than the specific mappings.

3.2.3 The Jensen-Shannon Divergence *

In this work, the main tool used to quantify information loss when going from atomistic to coarse-grained resolution is the relative entropy in the form of a Jensen-Shannon divergence (JSD or D_{JS}) [214]. The relative entropy framework has been previously established as a useful tool for evaluating the quality of coarse-grained models [78, 215]. The JSD is a variation of the well-known Kullback-Leibler divergence[216] (D_{KL}) used to calculate the relative entropy between two distributions. It offers two advantages over the Kullback-Leibler divergence in that it is symmetric and always has a finite value. Rather than directly relating two distributions, as is the case for the Kullback-Leibler divergence, the JSD computes the relative entropy by comparing each of these distributions to a third distribution which is the average of the other two distributions, as shown in the following equations

$$D_{\mathrm{JS}} = \frac{1}{2} D_{\mathrm{KL}}\left(P_{\mathrm{CG}} || P_{\mathrm{avg}}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(P_{\mathrm{AA}} || P_{\mathrm{avg}}\right), \tag{3.1}$$

$$\text{where } D_{\mathrm{KL}}(A||B) = \sum_{i=1}^{N} a_i \ln\left(\frac{a_i}{b_i}\right),$$

$$\text{and } P_{\text{avg}} = \frac{1}{2}(P_{\text{CG}} + P_{\text{AA}}).$$

In the above equations, we define D_{KL} in terms of two arbitrary distributions, A and B with N elements a_i and b_i. Here, we use the JSD to evaluate how well the distribution of the water/octanol partitioning free energies for the coarse-grained molecules (P_{CG}) match the corresponding distribution at the atomistic resolution (P_{AA}). A value of 0 indicates that the two distributions are the same. The use of the average distribution (P_{avg}) conveniently prevents divisions by zero when comparing histograms like those shown in Fig. 3.7a-d.

3.2.4 Basin Hopping and Minimization Schemes *

In this work, we use multiple methods to optimize the coarse-grained partitioning free energies to best match the atomistic distribution of free energies. The first such method is the basin-hopping method,[113] which is a variation of Metropolis-Hastings Monte Carlo. The algorithm proceeds in the following steps. Given a set of initial coordinates and objective function, the initial coordinates are first randomly perturbed and subsequently minimized. The results of the minimization are either accepted or rejected based on a predefined Metropolis criterion. These two steps form a single iteration of the algorithm, and a large number of iterations may be required to find the desired minima. Here, we use the JSD as our objective function and a set of possible water/octanol partitioning free energies for each coarse-grained bead type as our initial coordinates. Each move then corresponds to shifting the values of $\Delta G_{\text{W}\to\text{Ol}}$ for each coarse-grained bead type in a given force field. The optimizations were performed in order to define the desired $\Delta G_{\text{W}\to\text{Ol}}$ values for the five-bead-type force field, using the BASINHOPPING function provided by SCIPY[217] with a Broyden-Fletcher-Goldfarb-Shanno local minimizer,[218] a Metropolis temperature parameter of 0.008, and a step size of 0.024 kcal/mol. For the reference atomistic distribution, we applied the ALOGPS neural network to predict $\Delta G_{\text{W}\to\text{Ol}}$ for all molecules in the GDB with eight heavy atoms or less that were known to map to single bead Martini representations using the AUTOMARTINI algorithm. However, finding the optimal set of $\Delta G_{\text{W}\to\text{Ol}}$ values for the sixteen-bead-type force field using this approach proved to be computationally unfeasible, as the dimensionality of the problem scales with M^N, where N is the total number of bead types in the force field and M is the range of $\Delta G_{\text{W}\to\text{Ol}}$ values spanned by the Martini bead types divided by the step size. To parameterize the sixteen-bead-type force field, we used the SCIPY minimize function[217] with the modified Powell method,[218] starting with an initial set of eighteen bead types that were evenly distributed along the $\Delta G_{\text{W}\to\text{Ol}}$ axis. The results of the minimization indicated two sets of two bead types that were within 0.1 kcal/mol

of each other, and so each pair was combined into a single bead type, resulting in sixteen bead types total.

3.2.5 Clustering the GDB *

In addition to optimization of the JSD, a new set of coarse-grained water/octanol partitioning free energies was also proposed by clustering the GDB, leading to the 9-bead-type force field. Specifically, all GDB molecules with eight heavy atoms or less were grouped based on the number and type of hetero-atom substitutions present in the molecule (i.e., the number of times that a C was replaced with N, O, or F). The resulting atomistic molecular populations as well as the mean and standard deviation of their water/octanol partitioning free energies are shown in Fig. 3.9. Detailed information on each of the distributions (beyond what is provided in Fig. 3.9) is available in the ZENODO repository. The distributions are constructed based on the number and type of heavy atom substitutions that exist in the molecules. For example, the file named "GDB02to08_HAstats_fooo_subs.pdf" shows the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ distribution for all molecules containing one fluorine and three oxygen substitutions. Also included in the repository is a single file called "GDB02to08_HAstats.dat" which contains the mean and standard deviation for each of the distributions provided, which were used to make Fig. 3.9. The desired water/octanol partitioning free energies are determined by clustering the points on this graph, starting from the highest populated points and accepting anything that was within plus or minus 0.5 kcal/mol of these points. For example, the first point with the highest population in Fig. 3.9a is chosen as a starting point for the first bead type. All points that fall within 0.5 kcal/mol are assigned to this bead type and the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ is determined by taking a population-weighted average of all of these points. The next bead type is determined by selecting the highest point on Fig. 3.9a that is not already assigned to a bead type and repeating the process. For both this clustering and for the numerical optimization methods discussed in the previous section, the maximum number of heavy atoms per molecule was limited to eight. For all other data-driven calculations, all GDB molecules with up to ten heavy atoms were included.

3.2.6 Functional Group Analysis *

A statistical analysis of the functional groups found in the molecular fragments mapping to single beads is necessary in order to obtain a more detailed picture as to which chemistries are representative of specific bead types. The enumeration of functional groups was achieved through the use of the CHECKMOL software developed by Haider [166]. This software uses the 3D coordinates of each atom and the corresponding atom labels in a given molecule to identify common chemical

functional groups. A full list of the functional groups identified can be found in the ZENODO repository. Using CHECKMOL, we determine the degeneracy of specific functional-group pairs with respect to single bead types for the set of molecular fragments that mapped to a single bead. This amounts to counting the number of fragments containing a specific functional group pair and mapping to a single bead type. This population is then normalized with respect to the total number of fragments containing that same functional group pair across all bead types. It is useful to frame this statistical analysis in terms of conditional probabilities, as this yields specific information relevant for molecular-design applications. For example, the aforementioned counting and normalization is equivalent to calculating the likelihood of assigning a bead type (T) given a specific functional group pair (F), defined as $P(T|F)$. We use the fragment population distributions for each bead type and each functional group pair to obtain probabilities $P(T)$ of a bead type and $P(F)$ of a functional group pair. We then calculate the posterior probabilities $P(F|T)$ of a given bead type back-mapping to a specific functional group pair using Bayes' theorem

$$P(F|T) = \frac{P(T|F)P(F)}{P(T)}. \tag{3.2}$$

The results are shown as a series of heat maps for each force field in Fig. 3.10 and the corresponding heat maps for four-heavy-atom fragments is shown in Fig. 3.11.

3.2.7 Parameterization of New Bead Types *

The new force fields share most of the parameters defined by the Martini force field [210]. For the intra-molecular interactions, bonded, angle, and dihedral force constants remain the same as those prescribed by Martini. The non-bonded interactions only deviate from Martini through the strength of the potential used. We linearly interpolate across the interaction matrix defined in Martini,[210] utilizing the distance between the established Martini $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$[139] and the desired $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ for the interpolation. Fig. 3.5 shows the relationship between the Lennard-Jones ϵ parameter (related to the depth of the attractive well) for three given Martini bead types and the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ for all Martini bead types [210]. It is evident that there is no clear underlying functional form that can be applied to all Martini bead types. While there are localized regions that can be easily fit to lines, there are sharp discontinuities for each of the bead types at the boundaries of these localized regions. Therefore, linear interpolation is used to preserve these discontinuities in the new models, using the desired $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ as the target. The results of this interpolation are shown in Fig. 3.6. To construct this plot, we parameterized a new bead type for a series of $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ values evenly spaced along the range of $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ covered by Martini and ran simulations to

calculate the $\Delta G_{W \to Ol}$ for each using the methods described in Section 3.2.9. The results clearly show that the interpolation was successful for recovering the desired $\Delta G_{W \to Ol}$. There are some slightly larger deviations close for $\Delta G_{W \to Ol}$ closer to the P4 and P5 beads. This is probably due to the fact that, despite having a more attractive interaction with Martini water, the P5 bead has a slightly more positive $\Delta G_{W \to Ol}$ than the P4 bead type [139]. Since this interpolation was validated by calculating partitioning free energies between Martini water and pure Martini octanol, we also tested the effect that using a water-saturated octanol phase would have on the results [219]. This new phase consisted of 256 total solvent molecules, with 64 water molecules and 192 octanol molecules. The results are shown as the green points in Fig. 3.6. While the apolar and nonpolar bead types seem completely unaffected by this change in the octanol phase, the polar bead types show a $\Delta G_{W \to Ol}$ offset of approximately -0.5 kcal/mol compared to the interpolation target. We have included the epsilon values that make up the interaction matrices for all the new force fields as text files in the ZENODO repository.

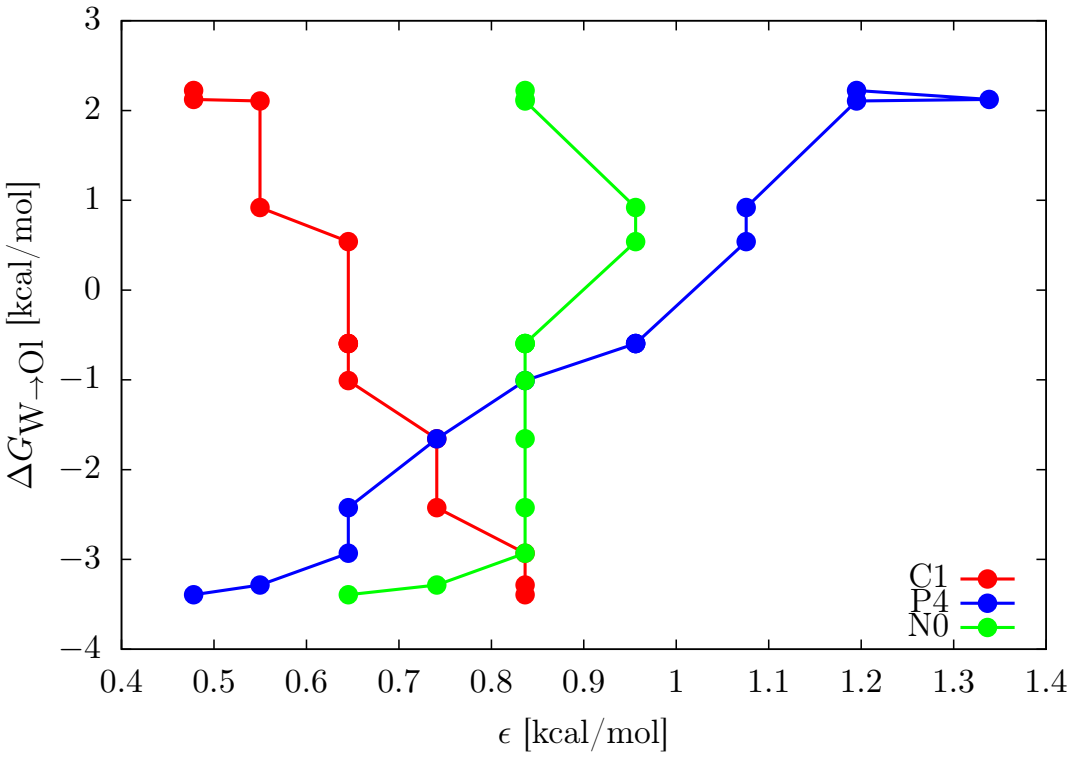

Figure 3.5: Relationship between the Lennard-Jones ϵ parameters for the Martini C1, P4, and N0 bead types and the $\Delta G_{W \to Ol}$ values for every Martini bead type.

Using this interpolation method, we parameterized three coarse-grained force fields. For the donor and acceptor types, we assigned the bead type which had the $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ closest to 0.0 as the bead type corresponding to molecules containing both donor and acceptor groups. Note that the donor/acceptor bead types are not labeled with "da" lettering, as is done for the Nda Martini bead type. We then followed the example set in the Martini interaction matrix [210]. The donor-only and acceptor-only bead type were assigned the same parameters as the da bead type but with a decrease in the ϵ value of 0.5 kJ/mol (making the interaction slightly more repulsive) when interacting with like bead types. For each force field, the bead types and corresponding $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ and ϵ values are given below. For all new bead types in this work, the Lennard-Jones σ matched that of a normal-sized Martini bead, $\sigma = 0.47$ nm.

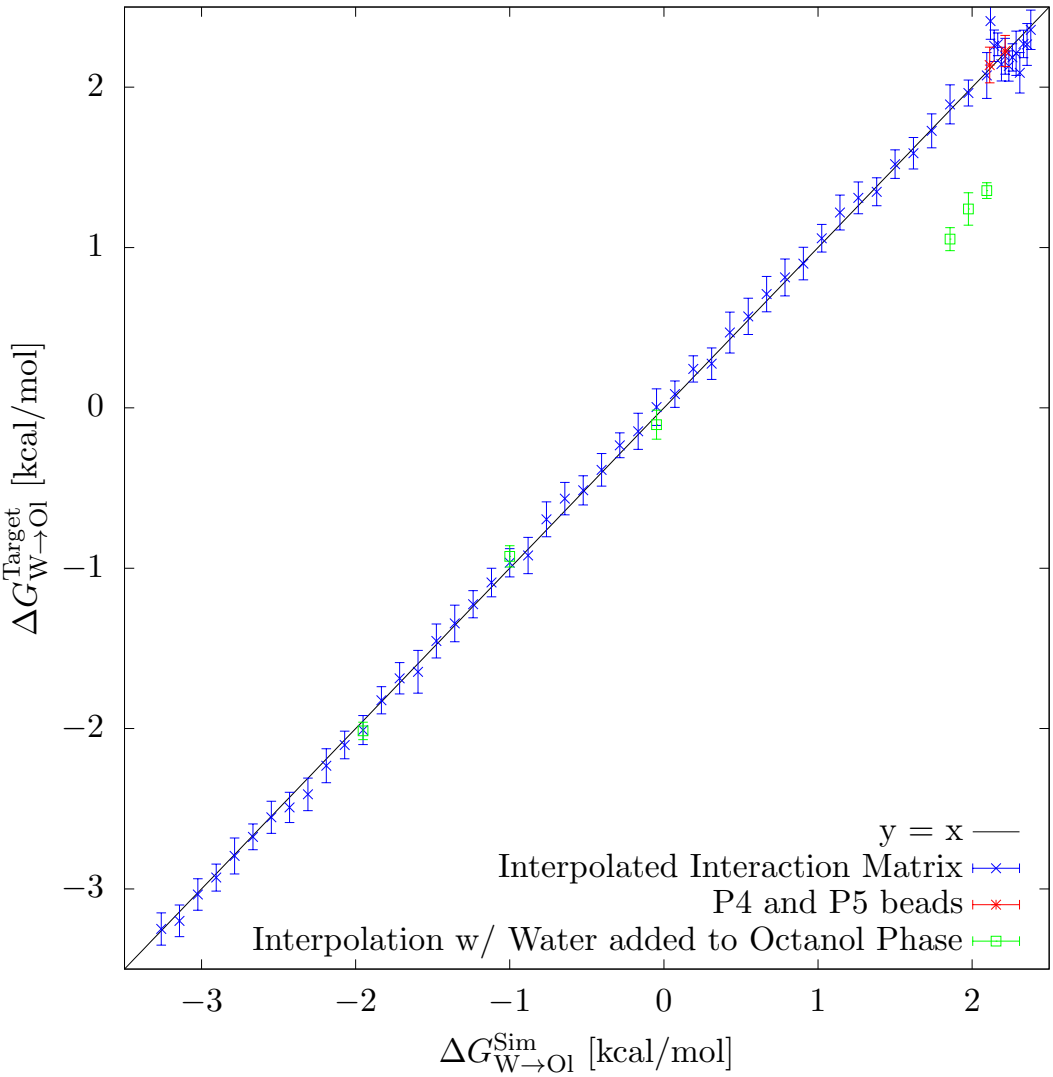

Figure 3.6: Calibration curve showing desired (target) $\Delta G_{\text{W}\to\text{Ol}}$ values on the vertical axis and $\Delta G_{\text{W}\to\text{Ol}}$ values obtained from simulations using beads parameterized by interpolating across the Martini interaction matrix on the horizontal axis.

Bead Type Name	Polar/Nonpolar/Apolar, Donor/Acceptor	$\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ [kcal/mol]
T1	Polar	2.05
T2	Polar	1.91
T3	Nonpolar Donor+Acceptor	0.098
T3d	Nonpolar Donor	0.098
T3a	Nonpolar Acceptor	0.098
T4	Apolar	-2.46
T5	Apolar	-3.13

Table 3.1: Names, characteristics, and $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ values for each bead type in the five-bead-type force field. For all beads, $\sigma = 0.47$ nm.

Bead Type Name	Polar/Nonpolar/Apolar, Donor/Acceptor	$\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ [kcal/mol]
T1	Polar	2.14
T2	Polar	1.39
T3	Polar	0.672
T4	Nonpolar Donor+Acceptor	-0.074
T4d	Nonpolar Donor	-0.074
T4a	Nonpolar Acceptor	-0.074
T5	Nonpolar	-0.899
T6	Apolar	-1.36
T7	Apolar	-2.17
T8	Apolar	-2.76
T9	Apolar	-3.51

Table 3.2: Names, characteristics, and $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ values for each bead type in the nine-bead-type force field. For all beads, $\sigma = 0.47$ nm.

The partitioning free energies of each bead type was then confirmed by running coarse-grained molecular dynamics simulations of single beads of each new bead type. These results show that this method yields an accurate force field without relying on an iterative scheme. The new bead types are named as Ti types, with i ranging from 1 to N where N is the total number of bead types in the force field. The numbering is also ordered by polarity. For example, the T1 bead type for all new force fields is the most polar type. Conversely, the T5, T9, and T16 bead types are the most apolar bead types in the five, nine, and sixteen-bead-type force fields, respectively.

Bead Type Name	Polar/Nonpolar/Apolar, Donor/Acceptor	$\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ [kcal/mol]
P4	Polar	2.22
P5	Polar	2.12
P3	Polar	2.11
P2	Polar	0.92
P1	Polar	0.54
Nda	Nonpolar Donor+Acceptor	-0.595
Nd	Nonpolar Donor	-0.595
Na	Nonpolar Acceptor	-0.595
N0	Nonpolar	-1.00
C5	Apolar	-1.66
C4	Apolar	-2.42
C3	Apolar	-2.93
C2	Apolar	-3.28
C1	Apolar	-3.39

Table 3.3: Names, characteristics, and $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ values for each neutral bead type in the Martini force field. For all beads, $\sigma = 0.47$ nm.

3.2.8 Extension to the Polarizable Martini Model

Due to the high popularity of the Martini model, many different modifications have been implemented, which allow for the modelling of other important physical interactions at this resolution. One such "flavor" of Martini was the introduction of a polarizable Martini water model by Yesylevskyy et al. [220]. This new Martini water consisted of three particles, which consisted of a standard neutral Martini bead type which interacted with other beads via a Lennard-Jones interaction only, as well as two charged beads (one positively and one negatively charged) that were bound to this neutral bead and only interacted with other charged particles via a Coulombic interaction only, with no other intermolecular interactions enabled. By replacing the previous Martini water (the P4 bead type) with this new water model, they were able to replicate the dielectric screening effect of bulk water in a Martini environment [220]. This model was improved by Michalowsky et al, with further modifications including new bead types that accurately represented monovalent ions and their electrostatic screening behavior in water [221, 222]. In order to determine whether an efficient HTCG approach could be implemented for more complex systems where charge screening plays a significant role, we implemented our interpolation scheme with the five-bead model using this improved polarizable Martini model, known as the "RefIon" model.

It was first necessary to calculate the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ values of each bead type using

Bead Type Name	Polar/Nonpolar/Apolar, Donor/Acceptor	$\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ [kcal/mol]
T1	Polar	2.18
T2	Polar	1.85
T3	Polar	1.03
T4	Polar	0.507
T5	Polar	0.335
T6	Nonpolar	0.126
T7	Nonpolar Donor+Acceptor	-0.061
T7d	Nonpolar Donor	-0.061
T7a	Nonpolar Acceptor	-0.061
T8	Nonpolar	-0.627
T9	Apolar	-0.838
T10	Apolar	-1.33
T11	Apolar	-1.62
T12	Apolar	-1.82
T13	Apolar	-2.20
T14	Apolar	-2.62
T15	Apolar	-2.81
T16	Apolar	-3.60

Table 3.4: Names, characteristics, and $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ values for each bead type in the sixteen-bead-type force field. For all beads, $\sigma = 0.47$ nm.

the RefIon model in order to perform the interpolation as was previously done for the standard Martini model. These values were obtained using the methods described in Section 3.2.9, but with the RefIon model used instead of standard Martini, and without the addition of antifreeze particles, as these were no longer necessary when using the RefIon model. Because the calculation of the hydration free energy is performed as a prerequisite for obtaining $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, we validated our simulations by comparing our obtained hydration free energies with those previously published by Michalowsky et al. for the neutral bead types, and find excellent agreement with their values. These results, as well as the corresponding $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ values for each bead type in the RefIon model are shown in Table 3.5. Unfortunately, we were unable to find previously reported hydration free energies for the charged bead types using the RefIon model.

After obtaining the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ values for the individual bead types of the RefIon model, we applied the interpolation scheme detailed in Section 3.2.7 to obtain force field parameters for the five-bead-type model, using their $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ calculated from the standard Martini model as a target. We then calculated the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$

Bead Type	Charged/(A/Non)Polar	ΔG_{Hydr} [kcal/mol]	$\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ [kcal/mol]
PQa	Charged	-27.51	25.45
Qda	Charged	-25.48	21.52
PQd	Charged	-22.52	20.47
Q0	Charged	-22.8	20.41
Qd	Charged	-23.96	19.99
Qa	Charged	-23.96	19.99
POL	Polar	-3.47	3.30
P4	Polar	-4.30	2.23
P3	Polar	-4.27	1.89
P5	Polar	-5.42	1.67
P2	Polar	-3.13	0.78
P1	Polar	-3.07	0.33
Nda	Nonpolar Donor+Acceptor	-1.97	-0.829
Nd	Nonpolar Donor	-1.97	-0.829
Na	Nonpolar Acceptor	-1.97	-0.829
N0	Nonpolar	-0.66	-1.44
C5	Apolar	0.41	-2.34
C4	Apolar	1.00	-2.89
C3	Apolar	1.03	-3.47
C2	Apolar	1.83	-3.82
C1	Apolar	2.59	-4.12

Table 3.5: Names, characteristics, and $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ values for each neutral and charged bead type in the RefIon force field. For all beads, $\sigma = 0.47$ nm.

of the newly parameterized model and compared these values to the target values used in the interpolation. The new $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ values are shown in Table 3.6 and closely match the targeted values shown in Table 3.1. This model is currently being used for HTCG screening schemes that will be detailed in a future work.

3.2.9 Coarse-Grained Simulations *

Coarse-grained molecular dynamics simulations were performed in GROMACS[141] version 4.6.6 using the standard Martini force-field parameters as well as the new force-field parameters derived in this work. A time step of $\delta t = 0.03\ \tau$ was used for all simulations, where τ is the natural time unit for the propagation of the model defined in terms of the units of energy $\mathcal{E}$, mass $\mathcal{M}$ and length $\mathcal{L}$ as $\tau = \mathcal{L}\sqrt{\mathcal{M}/\mathcal{E}}$.

Bead Type Name	Polar/Nonpolar/Apolar, Donor/Acceptor	$\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ [kcal/mol]
T1	Polar	2.03
T2	Polar	1.94
T3	Nonpolar Donor+Acceptor	-0.05
T3d	Nonpolar Donor	-0.05
T3a	Nonpolar Acceptor	-0.05
T4	Apolar	-2.21
T5	Apolar	-3.12

Table 3.6: Names, characteristics, and $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ values for each bead type in the five-bead-type force field for use with the RefIon model. For all beads, $\sigma = 0.47$ nm.

The simulations were run in an NPT ensemble with a Langevin thermostat and Andersen barostat[223] to keep the temperature and pressure at 300 K and 1 bar, respectively. The corresponding coupling constants were $\tau_T = \tau$ and $\tau_P = 12\tau$.

Water/octanol partitioning free energies were obtained by simulating approximately 500 coarse-grained molecules in octanol and water. Approximately 250 octanol molecules and 350 Martini water molecules were simulated for their respective systems, with the appropriate number of antifreeze particles [210]. The free energies were computed using the Bennett acceptance ratio method[157] in which the coarse-grained solute was incrementally decoupled from the solvent via the coupling parameter, λ. Twenty-one simulations were run for each molecule at evenly spaced λ values ranging from 0 to 1, with each simulation run for 200,000 time steps. Finally, the partitioning free energies were calculated using the relation $\Delta G_{\mathrm{W}\to\mathrm{Ol}} = \Delta G_{\mathrm{W}} - \Delta G_{\mathrm{Ol}}$. As this method does not take into account the saturation of the octanol phase with water that occurs in experimental systems,[219] we also ran some test simulations using an octanol phase which contained 25% molar water molecules. We found that only coarse-grained molecules containing a majority of highly polar beads would show a reduction in their $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ values due to increased contact with water in the water-saturated octanol phase, as shown in Fig. 3.6.

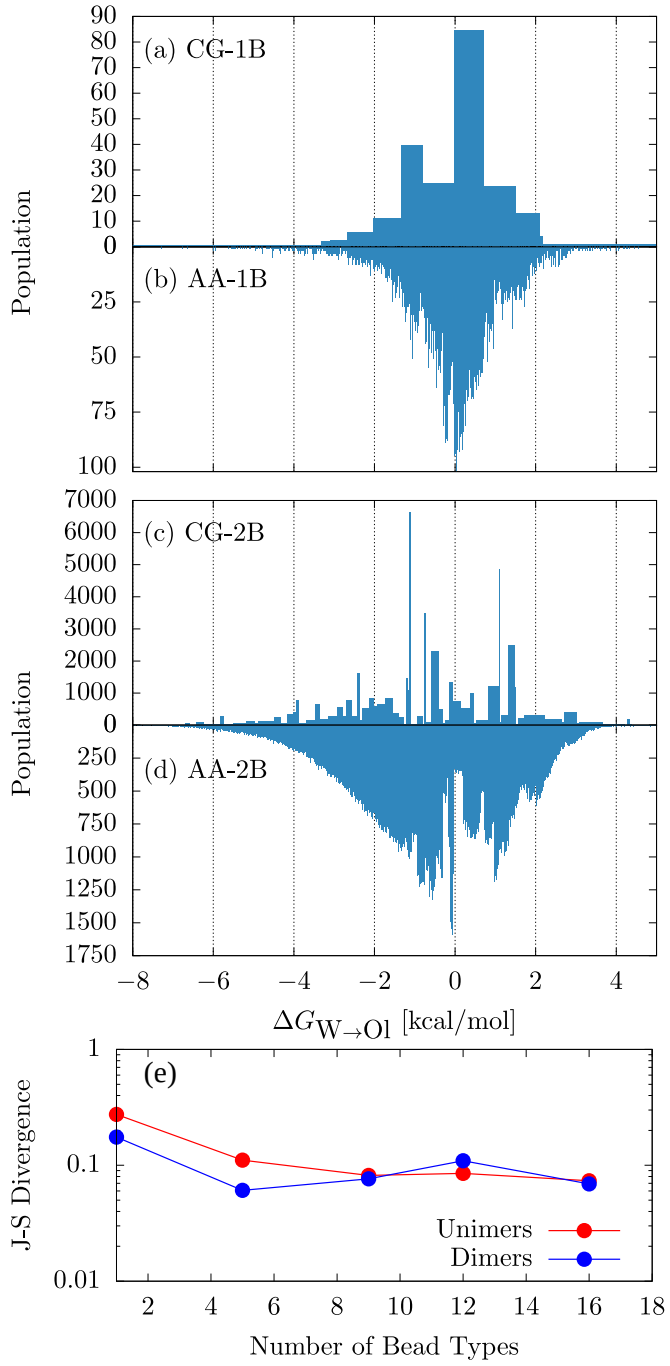

Figure 3.7: Histograms of 343,700 small molecules extracted from GDB that map onto one-bead or two-bead coarse-grained Martini representations. (a),(c) Coarse-grained and (b),(d) atomistic populations as a function of water/octanol partitioning free energy. The width of the bars in (a),(c) corresponds to the range of atomistic water/octanol partitioning free energies that can map to that coarse-grained representation. (e) Jensen-Shannon divergence calculated for the histograms corresponding to those in (a)-(d) for all force fields described in this work.

3.3 Results *

3.3.1 Quantifying information loss of coarse-grained models with varying number of bead types *

The updated AUTO-MARTINI algorithm was used to first map and subsequently assign bead types to 3.5 million molecules of the GDB containing ten or fewer heavy atoms using the Martini force field as well as the other three force fields parameterized by interpolating the Martini interaction matrix. Fig. 3.7 shows a comparison of the atomistic and coarse-grained $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ distributions for molecules mapping to Martini unimers (Fig. 3.7a,b) and dimers (Fig. 3.7c,d). In Fig. 3.8, we show all of the histograms used to compute the JSDs shown in Fig. 3.7e. The width of the coarse-grained bars reflects the range of $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ values within which a molecule must fall in order to be assigned that bead type, or, in the dimer case, a combination of bead types. The height of the bars is set such that the area covered by each bar is equal to the total number of molecules that were assigned that coarse-grained representation. We then calculate the JSD between the coarse-grained and atomistic histograms for each force field to quantify the information loss as a function of the number of bead types present in each force field (Fig. 3.7e). Increasing the number of bead types reduces the information loss when going from atomistic to coarse-grained resolution, though this reduction becomes insignificant after reaching nine bead types. The JSD comparing the unimer histograms (red curve in Fig. 3.7e) changes negligibly when increasing the number of bead types from nine to sixteen, with only a small increase for the Martini case (12 bead types). This is expected due to the fact that the atomistic histogram of GDB molecules mapping to a single bead is a simple, unimodal distribution with a peak at $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}} = 0$. Since all of the force fields have at least one amphiphilic bead type with a $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ close to 0, they all capture this defining feature of the histogram, and, comparatively, further information gains are negligible. Remarkably, we find the JSDs were largely insensitive to the choice of numerical optimization technique used for the derivation of each force field, as they all capture this prominent feature.

The JSDs calculated from the dimer histograms (blue curve in Fig. 3.7e) show a variety of interesting features. Both the nine and sixteen-bead-type force fields maintain roughly the same JSD, suggesting that the combinatorial explosion that results from doubling the molecular weight is captured by these force fields. The slight increase seen in the unimer JSD for Martini is noticeably amplified for the dimer case, indicating that careful placement of bead types on the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ axis is necessary to maximize chemical transferability.

Surprisingly, the greatest deviation in the JSD going from the unimer to dimer histogram comes from the five-bead-type force field, dropping well below the val-

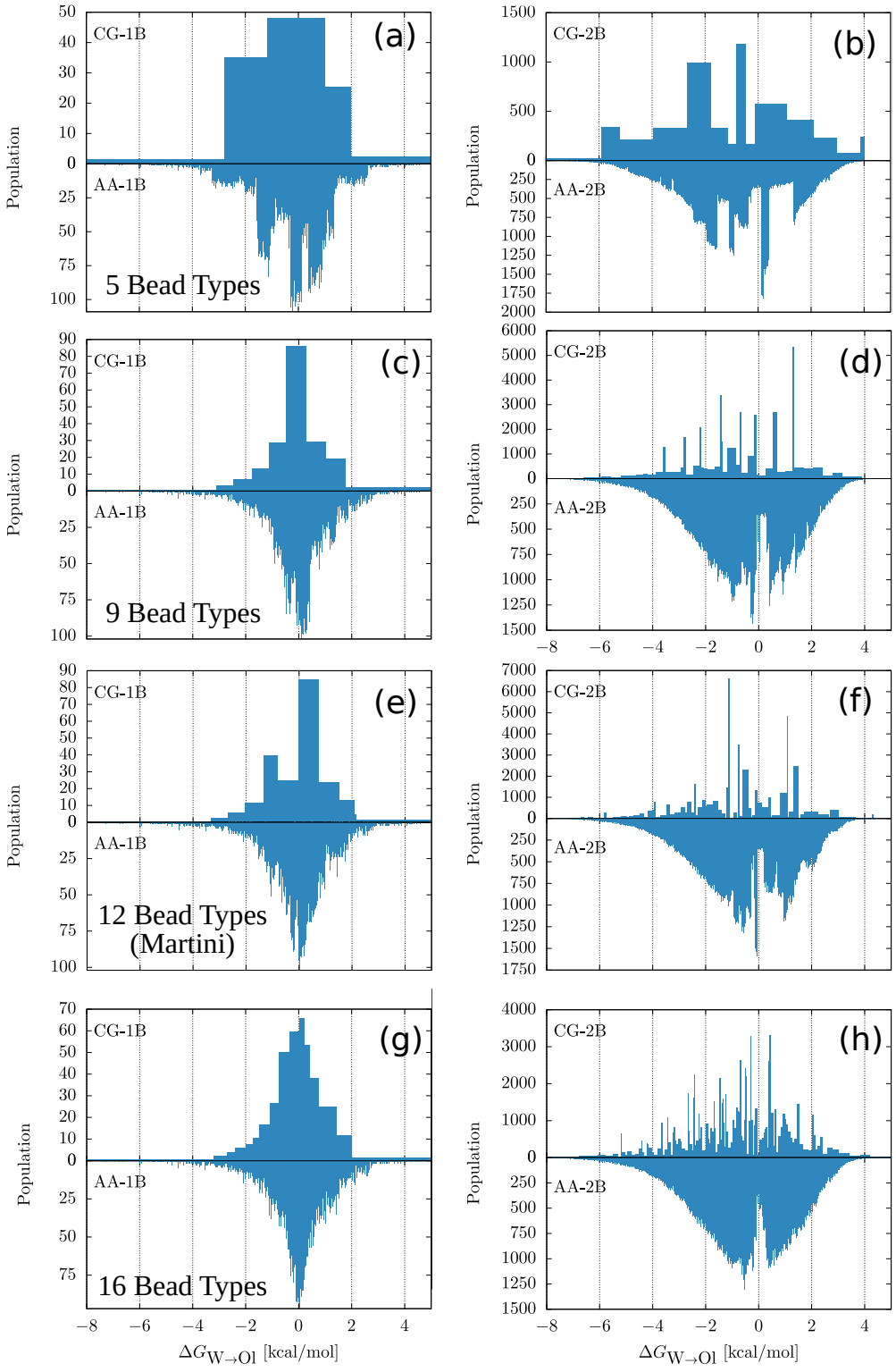

Figure 3.8: Histograms used to calculate JSD values shown in Fig. 3.7. The unimer and dimer distributions are shown for the five-bead-type (a,b), nine-bead-type (c,d), Martini (e,f), and sixteen-bead-type (g,h) force fields.

ues for the higher bead type force fields. The reason for this can be seen in Fig. S3b, which shows that the distribution of atomistic compounds mapping to dimers in the five-bead-type force field is significantly different from its analogs for the other force fields. While the other distributions contain populations ranging from $3.3 \cdot 10^5$ to $3.4 \cdot 10^5$, the 5-bead-type force field has only $3.0 \cdot 10^5$ molecules. Furthermore, even though the shapes of the distributions for the other three force fields are far more similar to each other than to the five-bead-type force field, the intersection of the sets of atomistic compounds mapping to each force field consists of $2.3 \cdot 10^5$ molecules. Including the set of molecules mapping to the 5-bead-type force field reduces this intersection to $1.8 \cdot 10^5$ molecules. This explains why the JSD value for the five-bead-type model is significantly lower than all of the others. This indicates that a significant number of molecules that would map to dimers when using one of the other force fields are mapped to trimers or tetramers using the 5-bead-type force field. Unfortunately, we were unable to compute histograms of molecules corresponding to coarse-grained trimers or tetramers due to computational constraints: in order to get a converged distribution that could represent the chemical space corresponding to molecules mapping to trimers, we would need to run the AUTO-MARTINI algorithm on the GDB molecules containing up to at least 15 heavy atoms (assuming a 5 heavy atom to 1 bead mapping), which is computationally unfeasible due to the exponential growth of CCS as a function of molecule size.

3.3.2 Relating chemistry to bead types *

As an alternative to purely numerical methods for determining the optimal $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ values for the bead types of a coarse-grained force field that best partitions CCS, we cluster the GDB itself and use the weighted average of $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ for each cluster. Fig. 3.9a shows the two descriptors upon which we project and subsequently cluster the GDB. Each point in Fig. 3.9a represents the set of molecules in the GDB that have a specific number and type of heavy atom substitutions (i.e., N, O, or F). The points are placed on the $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ axis according to the average of their $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ distribution. The error bars represent the standard deviation of the $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ in each distribution. One of the corresponding distributions is shown in Fig. 3.9b, with the rest available in the ZENODO repository. The points are clustered hierarchically with respect to population and average as shown in Fig. 3.9a. The highest-populated points are all chosen as cluster centers as long as they are separated by at least 0.5 kcal/mol, which is an arbitrarily chosen length-scale for the clustering to ensure a reasonable number of bead types in the final force field. After the points are clustered, the desired $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ of each bead type is determined by taking the population-weighted average of all the points in a cluster. This intuitively provides a basic understanding of the chemistry that

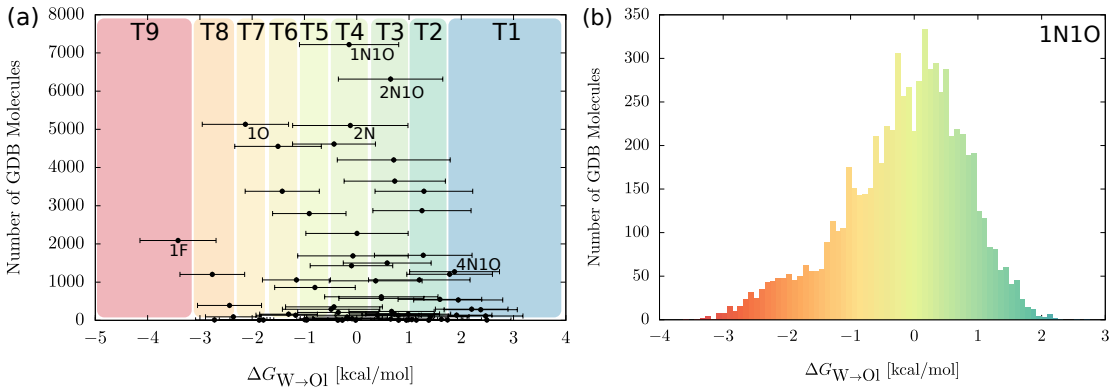

Figure 3.9: (a): Population versus average values of the distributions of water/octanol partitioning free energies corresponding to GDB molecules with up to 8 heavy atoms and a specific number and type of hetero-atom substitutions. The error bars refer to the standard deviations of each distribution. The colored backgrounds denote how these average values are clustered to obtain new bead types that more efficiently divide CCS. (b): Example distribution corresponding to top-most point labeled in the graph on the left. The label refers to the number and type of hetero-atom substitution, in this case 1 nitrogen and 1 oxygen substitution. The color applied to the histogram corresponds to the colors shown in (a), indicating the bead types to which these molecules would be assigned.

maps to a specific bead type. For example, a T4 bead is more likely to back-map to a molecule with one N and one O substitution compared to two N substitutions because of the difference in the GDB populations of each molecule type.

It is important to characterize the degree to which unique chemistries are captured by the bead types of each force field. Using the GDB as a proxy for CCS enables a quantitative understanding of the chemical transferability of each bead type through the calculation of conditional probabilities. Fig. 3.10 shows a series of heat maps corresponding to each of the four force fields investigated in this work. These heat maps are constructed by counting all fragments containing only five heavy atoms and assigned to a specific bead type, such that two functional groups are detected by the CHECKMOL software package. The fragment population distributions are then used to calculate the Bayesian likelihood $P(T|F)$ and posterior $P(F|T)$ for each bead type/functional pair combination in every force field. The numbers on the horizontal axis for each heat map denote specific pairs of functional groups found in the chemical fragments that are assigned to a bead type, while the color corresponds to either the likelihood or posterior probabilities. We see the localization of functional-group pairs to specific bead types mainly because of the constraint of only including fragments with five heavy atoms. This constraint limits the combinatorics of hetero-atom and bond substitutions that result in functional-group pairs. Despite the addition of these constraints, a large number of functional-group pairs are still split across multiple bead types.

Fig. 3.11 shows the likelihood and posterior values calculated for fragments containing only four heavy atoms and two functional groups as specified by CHECKMOL. The total number of bead types of each force field is not reflected in these heat maps, with the most apolar bead types missing. This is because all of the fragments that map to these bead types consist of saturated hydrocarbons or single alkene/alkyne substitutions only, and thus are not detected as having a functional group pair by CHECKMOL. Furthermore, there are no values calculated for the T7 beads in the sixteen-bead-type force field because there were no donor/acceptor/donor+acceptor fragments that also had two functional groups within the narrow range of $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ covered by the T7 bead types. Similar reasoning can also be applied to explain the lack of values for the T11 bead type in the same force field.

Table 3.7 provides additional quantification of the trends seen in Fig. 3.10, displaying the average number of functional-group pairs per bead type, as well as the number of likelihood and posterior values above cutoff values of 0.99 and 0.2, respectively. As the number of bead types increases, both the average number of functional-group pairs per bead type and the number of likelihood values greater than 0.99 decrease, indicating that fewer bead types in a force field increases the coverage of CCS for each bead type. The opposite trend is observed for the num-

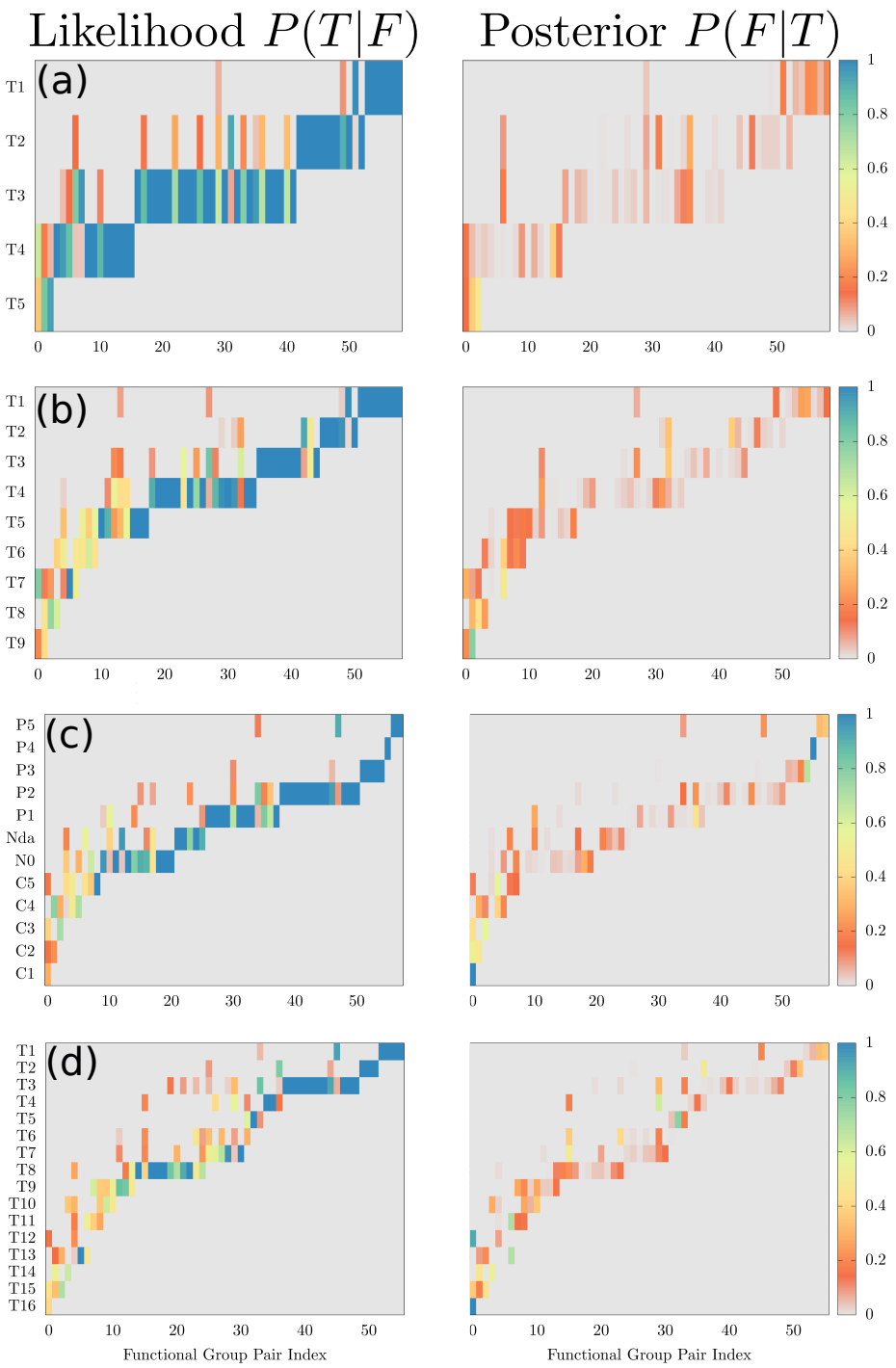

Figure 3.10: Heat maps portraying the degeneracy of specific pairs of functional groups for a given bead type for force fields containing five (a), nine (b), twelve (c), or sixteen (d) bead types. The horizontal axes denote specific functional-group pairs that exist in a chemical fragment with five heavy atoms only. The color corresponds to either the column-normalized or row-normalized probabilities. The column-normalized probabilities (left side) are equivalent to the Bayesian likelihood of a given functional group mapping to a specific bead type. The row-normalized heat maps (right side) show the Bayesian posterior probabilities of obtaining a specific functional group given a bead type.

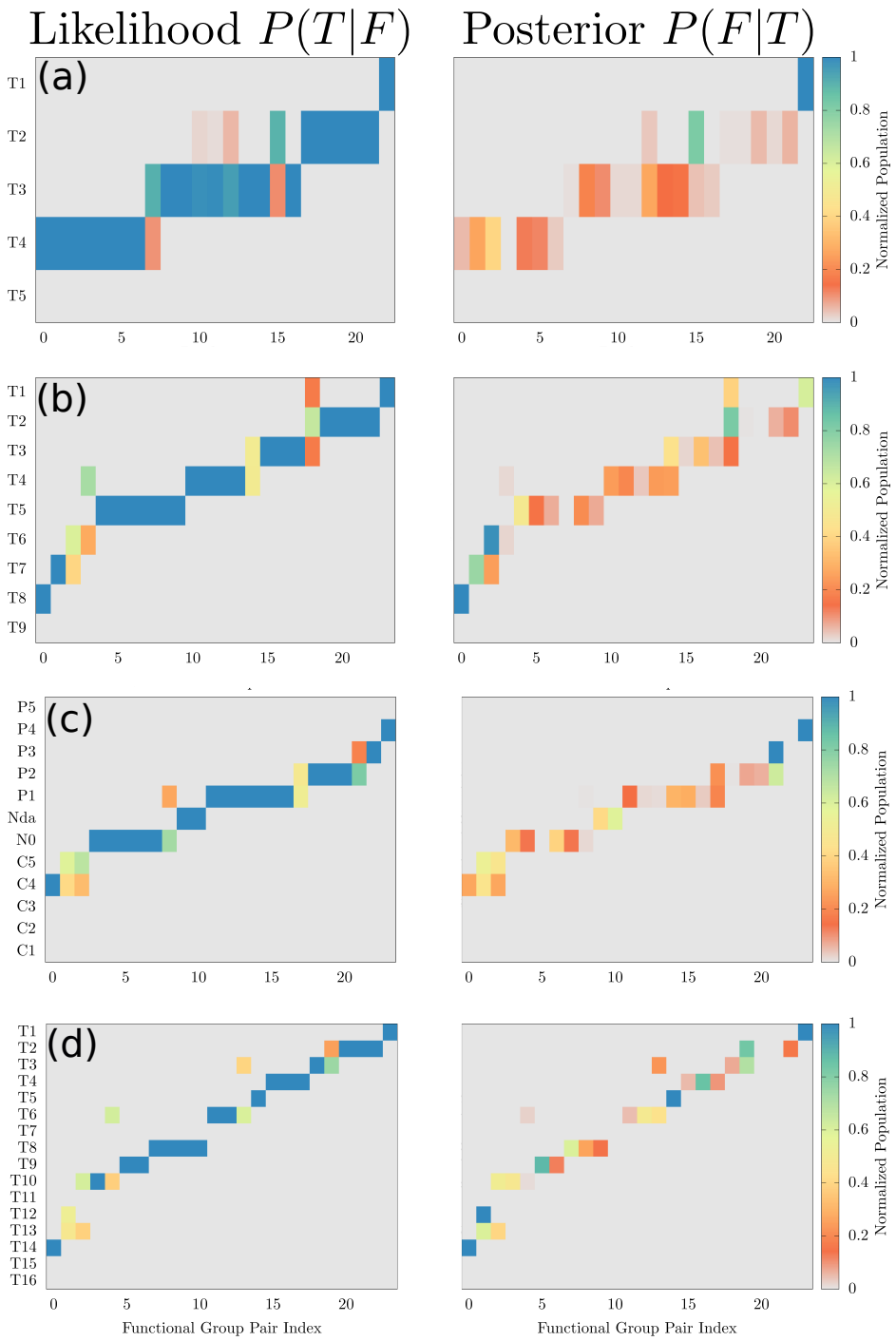

Figure 3.11: Heat maps portraying the degeneracy of specific pairs of functional groups for a given bead type for force fields containing five (a), nine (b), twelve (c), or sixteen (d) bead types. The horizontal axes denote specific functional-group pairs that exist in a chemical fragment with four heavy atoms only. The color corresponds to either the Bayesian likelihood (left side) or posterior (right side) probabilities.

ber of posteriors greater than 0.2, indicating that more bead types result in higher chemical specificity for each bead type.

# of Bead Types	Avg. # of Func. Group Pairs	# of Likelihoods > 0.99	# of Posteriors > 0.20
5	16.4	40	8
9	10.1	33	17
12 (Martini)	7.4	35	20
16	6.4	27	26

Table 3.7: For each force field, the number of bead types, the average number of functional-group pairs per bead type, the total number of likelihood values over 0.99, and the total number of posterior values over 0.2.

Over the course of this work, certain idiosyncrasies were discovered when using CHECKMOL. One such issue was the fact that the code tended to double-count some functional groups. For example, fragments with only a single fluorine substitution were counted as both a "halogen derivative" and as a "alkyl fluoride". This was only observed for the aforementioned fluorine substitutions as well as for dialkyl ethers. Other examples were also found for which the software could not correctly identify the functional groups contained in the fragment. This is probably due to the fact that CHECKMOL was not tested on some of the less common chemistries encountered in the GDB. The most egregious example of this was found for fragments containing the smiles string "NC=N" which were incorrectly labeled as a carboxylic acid derivatives by CHECKMOL. For this reason, we did not explicitly label the horizontal axes with their corresponding chemistries in Fig. 3.11 and Fig 3.10. For full transparency, we have included the smiles string for each unique fragment used in the Bayesian analysis for both four-heavy-atom and five-heavy-atom fragments as well as the corresponding values for $P(F)$, $P(T)$, $P(T|F)$, and $P(F|T)$ in the ZENODO repository. While the functional group labels given by CHECKMOL are incorrect in a few cases, the overall trends reported in this work are unaffected: namely, that increasing the number of bead types also provides increased values of the Bayesian posterior probabilities for back-mapping specific chemistries.

3.3.3 Coarse-grained force field validation *

While we have demonstrated that the careful placement of bead types on the $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ axis leads to more chemical transferability, the force fields themselves must be validated. Because $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ was used as the target property for the

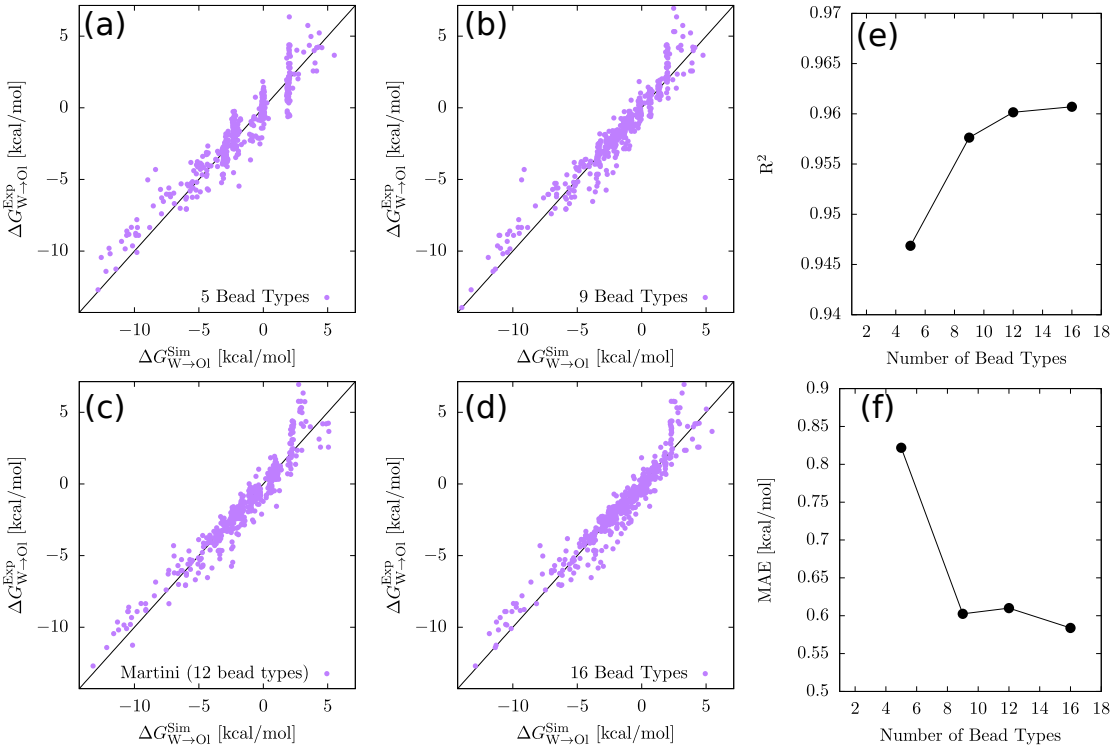

Figure 3.12: Correlation curves comparing the $\Delta G_{W\to Ol}$ calculated from coarse-grained MD simulations of approximately 500 molecules to their measured values from experiment. The results are shown for each of the force fields described in this work (a)-(d) as well as their respective Pearson correlation coefficients and mean absolute error (MAE) values (e)-(f).

interpolation of the Martini interaction matrix, we must ensure that this property is indeed captured by the resulting models and determine to what extent the accuracy of these models changes as the number of bead types increases. Because we used the Martini interaction matrix in our parameterization, our force fields are, by construction, fully compatible with the existing Martini model. This allows us to use the Martini solvent models for both water and octanol in our validation, without having to derive new solvent models for each new force field. Essentially, this means that, using the methods outlined in this work, it is possible to parameterize any number of new bead types with desired $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ values within the $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ range covered by Martini that will also be compatible with the Martini model. Fig. 3.12 shows correlation plots comparing $\Delta G_{\mathrm{W}\to\mathrm{Ol}}$ values computed from coarse-grained MD simulations with experimental values for approximately 500 ring-less molecules obtained from the National Cancer Institute database [184]. The comparison is made for all four of the models examined in this work. The number of compounds varies for each model, as the AUTO-MARTINI algorithm was able to successfully find mappings for more molecules in the database when using a model with a higher number of bead types, ranging from 479 compounds mapped when using the five-bead-type model to 505 when using the sixteen-bead-type model. The full set of compounds as well as their corresponding coarse-grained representations is provided in the ZENODO repository. The vertical series of points prominently seen in Fig. 3.12a are a consequence of the increased degeneracy of CCS for the 5-bead-type model: they represent many compounds mapping to the same coarse-grained representation. As expected, the correlation becomes less discretized as the number of bead types increases. Examining Figs. 3.12e and 3.12f, we see corresponding gains and losses in the Pearson correlation coefficients and MAEs, respectively. Surprisingly, the gains in accuracy are very slight as a function of number of bead types—with the correlation coefficient only increasing by 0.01 and the MAE decreasing by 0.2 kcal/mol—despite tripling the number of bead types. Even with the five-bead-type model, we achieve an MAE of 0.8 kcal/mol, within the standard for chemical accuracy. We deliberately chose not to include molecules that contained rings because this version of the Martini force field quantitatively fails in modelling molecular rings for many documented reasons. These reasons include lack of cross-parameterizations between normal-sized and the "small" sized beads used to model rings,[224] as well as the size disparity between the atomistic and coarse-grained ring structures [225]. For these reasons, and in anticipation of the new Martini version 3.0 that is currently being developed to address these flaws, we refrained from addressing ring molecules in this study.

3.4 Discussion *

Given the immense size of CCS, the creation of reduced models that efficiently subdivide the space is necessary for screening applications. Here we demonstrate the use of the water/octanol partitioning free energy as the parameter used to generate top-down chemically-transferable coarse-grained models of varying numbers of bead types. This choice of descriptor is inspired by the Martini force field, which prescribes the use of partitioning free energies as the tool to encode the natural hydrophobic/hydrophilic character of a molecule when determining the bead type to be used to represent a molecular fragment [89]. A major strength of Martini is its ability to model this important molecular property using simple Lennard-Jones potentials with varying attractive well depths. Here, we use the GDB as a proxy for CCS[208, 209] and apply the AUTO-MARTINI algorithm to compare the populations of the GDB molecules and their corresponding coarse-grained representations for four different force fields with varying numbers of bead types. This effectively amounts to a discretization of CCS projected onto $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ at multiple resolutions. Other oil/water partitioning free energies have also been proposed for the determination of bead-type assignment in Martini, such as hexadecane, chloroform, and ether [210]. We restrict ourselves to the water/octanol partitioning free energy because of the difficulty involved in obtaining either experimental partitioning free energy data or predictions for the wide variety of chemistries found in the GDB. The ALOGPS neural network allows us to obtain accurate predictions of $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ for new chemistries in this regard. While it is possible that the use of other water/oil partitioning free energies would change the resulting force-field parameterizations, previous studies have shown that many Martini partitioning free energies can be viewed as linear transformations of $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ [110, 208, 209]. Therefore, the use of a single type of partitioning free energy should be sufficient as a metric for parameterizing these types of models with respect to the overall hydrophobic/hydrophilic character of the bead types.

Fig. 3.7e quantifies the level of information loss using the JSD as the resolution is varied, allowing us to determine how effectively each of these force fields, including Martini, represents CCS projected onto $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$. The JSD decreases as the number of bead types increase. However, the information retention becomes negligibly greater, essentially plateauing after nine bead types. Remarkably, despite the fact that the Martini force field was parameterized using a small number of chemical compounds (relative to the large distribution of compounds used to parameterize the other models in this work), it shows only a minuscule increase in the JSD. This is mainly due to the lack of a bead type that is placed at 0.0 kcal/mol on the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ axis. The highly populated peak at this location is the major defining feature for the atomistic distribution of molecules mapping to unimers, and the placement of the Martini Nda and P1 bead types is insufficient

to fully capture this feature. Unfortunately, this increase in the JSD is amplified when comparing the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ distributions for dimer molecules, whereas for the nine and sixteen bead type models, the JSD seems to converge. The combinatorial explosion that results from doubling the size of molecules (i.e., going from unimer to dimer) is reflected in these histograms as a broadening of the total distribution, since more hydrophobic and hydrophilic values of $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ are possible as molecule size increases. Fig. 3.7e shows that the nine and sixteen bead type force fields match this combinatorial explosion.

On the other hand, Figs. 3.12e and f clearly demonstrate that a high level of accuracy is already achieved with respect to $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ using the five-bead-type force field. What, then, is the benefit to using a model with more than five bead types? As we show from Figs. 3.9 and 3.10, the main advantage is in back-mapping the coarse-grained representations to their likely atomistic counterparts [226]. Specifically, the nine bead force field is parameterized not by simply optimizing the JSD, but rather by clustering the GDB molecules into sub-distributions based on the type and number of heavy-atom substitutions on the carbon scaffold of each molecule as shown in Fig. 3.9. As expected, this clustering strategy also results in a minimal value of the JSD, while providing an added convenience. The distributions that were clustered to make this force field provide a method for predicting the chemistries that are most representative of a bead type. Since the standard deviations of these distributions are so large, such that some span across three different bead types, this provides only a rough idea of the probable chemistry accessible to a bead type. Moreover, knowledge of the presence of one or two heavy-atom substitutions on a carbon scaffold of up to 8 heavy atoms is insufficient for back-mapping given the number of ways in which they can be arranged on that scaffold resulting in wildly different chemical properties. Fig. 3.10 shows how different functional-group pairs will map clearly to specific bead types when the scaffold size is reduced to five heavy atoms. This extra constraint enables a clearer understanding of the range of unique chemistries that are accessible to a specific bead type. Decreasing the size of the scaffold from five to four heavy atoms yields correspondingly narrower distributions of $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, meaning that the same functional group pair can be found in fewer bead types. By no longer requiring functional-group pairs and increasing the scaffold size to eight heavy atoms, we begin obtaining distributions similar to those seen in Fig. 3.9.

Table 3.7 also demonstrates that the number of unique functional-group pairs that map to a given bead type decreases as the number of bead types increases, to the point where, for Martini as well as the sixteen bead type force field, there exist bead types that essentially back-map to a single functional group pair. Here, we see a clear parallel with structural coarse-graining methods: just as decreasing the size of the beads leads to a finer mapping of the configurational space, increasing

the number of bead types leads to a finer mapping of CCS. The efficiency of a coarse-grained model can be optimized by tuning the mapping function and bead size of a coarse-grained model such that the accuracy of the model is balanced with respect to the computational cost of simulating a greater number of particles. By fixing the geometric mapping method and bead size, and only varying the number of bead types possible, we instead balance between the accuracy of representing specific chemical features and the cost of parameterization and validation of the inter-particle potentials. We circumvent this cost by interpolating the Martini interaction matrix to obtain accurate parameters for all of the force fields presented in this work. However, this cost will be significant for models requiring a more rigorous parameterization scheme relying on other molecular descriptors. Separate from this trade-off between accuracy and parameterization cost, a "screening efficiency" can be defined as the average number of functional-group pairs that map to a single bead type, indicating a larger region of CCS being captured by a single bead type. Unsurprisingly, Table 3.7 shows that the five-bead-type force field clearly has the highest back-mapping efficiency.

This statistical analysis of functional-group pairs also suggests a Bayesian approach to computing the probability of a functional group pair, F, given a bead type, T, represented as $P(F|T)$ in equation 3.2. $P(F)$, the Bayesian prior, is the probability of finding the specific functional group pair in the set of molecular fragments (made up of five heavy atoms and containing two functional groups as defined by CHECKMOL) that mapped to single beads, and $P(T)$ is the probability of choosing the given bead type from that same data set. The likelihood, $P(T|F)$, shown in the left side of Fig. 3.10 prescribes the bead type or types to which a fragment could be assigned based on its chemistry—the equivalent of the Martini "bible" for assigning bead types. As shown in Table 3.7, the number of functional-group pairs with likelihoods greater than 0.99 (essentially localized to a single bead type) decreases as the number of bead types increases. The Martini force field deviates slightly from this trend, with two more functional-group pairs with high likelihoods as compared to the nine-bead-type force field. This may stem from the parameterization strategy used for Martini that relied on specific molecules and their functional groups rather than aiming to efficiently span chemical space by optimizing the JSD, as proposed in this work. The posterior probabilities, which provide a quantitative description of which chemistries are more representative of each bead type, increase as the number of bead types increases. This effect more easily facilitates the back-mapping of coarse-grained representations. These two quantities, the Bayesian likelihood and posterior, are essential for further exploring CCS covered by specific bead types and enabling both direct and inverse molecular design.

Interestingly, we immediately see a resolution limit with respect to the functional-

group pairs that map to specific bead types. Because there are certain length scales on the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ axis that correspond to the distribution of specific functional-group pairs, increasing the number of bead types will naturally split these distributions, such that one functional group pair is represented in multiple bead types. Fig. 3.10a shows that the majority of functional-group pairs are encompassed either by a single bead type or one of its neighbors on the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ axis. Increasing the number of bead types causes these splits to become more exacerbated, spanning multiple bead types for an increased number of functional-group pairs. This is the resolution limit of this type of top-down coarse-graining. The large bead sizes of these models leads to a high degree of variability in the chemistry, meaning that it is no longer obvious which functional group/functional group pair belongs to which bead type. The limit is most evident for the functional-group pairs mapping to the T3 and T13 bead types in Fig. 3.10d, indicating that they are placed too close to their neighbors on the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ axis. These functional-group pairs contain some combination of the following functional groups: alkene, alkyne, enamine, hydrazine, hydroxylamine, carboxylic acid derivatives, and fluorine substitution. The placement of these functional groups within a five-carbon scaffold will drastically shift the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ beyond the range of the next-nearest bead type on the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ axis and highlights the limitations of only using this single descriptor for the projection of CCS. While the addition of other partitioning free energies may further increase the accuracy of both the models themselves and the mapping of specific functional groups, these descriptors are encoding essentially the same information as $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$: the hydrophobicity of the underlying chemistry. However, determining a suitable orthogonal descriptor and then parameterizing a chemically-transferable coarse-grained force field to achieve a more direct relation with CCS is outside the scope of this work, and will be addressed subsequently.

3.5 Conclusion

In this work, we use the Jensen-Shannon divergence (JSD) to quantify the information loss in chemically-transferable top-down coarse-grained models with varying numbers of bead types, with the GDB as our proxy for chemical compound space (CCS). We find that Martini, while not designed to efficiently reduce CCS, performs remarkably well in this regard, closely matching the other force fields explicitly designed to minimize the JSD with only a small deviation. All force fields yield roughly the same level of accuracy with respect to $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, but vary greatly in their coverage of CCS. We used a Bayesian approach to calculate the probabilities of back-mapping given bead-types to fragments containing specific chemical substitutions. Here, we found it necessary to constrain the size of chemical fragments to five heavy atoms and the presence of two functional groups in

order to clearly differentiate between the chemical moieties mapping to each bead type. The results of this Bayesian analysis indicate that increasing the number of bead types decreases the range of accessible chemistry while increasing the corresponding posterior probabilities for each chemistry. However, there is a resolution limit when using this approach, as it does not take into account the specific positions of hetero-atom and bond substitutions within a fragment, causing different bead types to appear representative of the same chemistry. Martini, as well as other chemically-transferable coarse-grained models, can be used to quickly build structure–property relationships that span broad regions of CCS. Here we highlight the powerful combination of this method with Bayesian inference, providing an informed mapping of a coarse structure–property relationship to a higher resolution in chemical compound space and further enabling inverse molecular design.

This work also reinforces the conclusions of the previous chapter regarding the top-down approach to coarse-graining CCS. In the previous chapter, global unsupervised learning methods were applied to a data base of fragments that were mapped to Martini dimers. The results indicated that $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ correlated well with the number and type of functional groups found on a carbon scaffold. We therefore tested the extent to which the HTCG approach could be optimized by developing models that covered the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ axis at varying resolutions. We had further noted that most of the apolar compounds were grouped into two clusters only, which was at odds with the total number of apolar bead types used in Martini (C1 through C5). Indeed, we saw that the number of apolar bead types could be reduced to two while maintaining the overall accuracy of the model for apolar compounds, as was done for the five-bead-type model in this work (Fig. 3.12a). At the same time, the presence of the vertical series of points dominating the non-polar and polar regions of Fig. 3.12a also correlates with the results from the previous chapter, with far more than three clusters corresponding to these regions despite being represented by only three bead types in this model.

It was also evident from this work that, just accounting for the correlation between the number/type of heavy atom substitutions and $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ would be insufficient to easily identify specific functional groups for inverse molecular design without also drastically increasing the number of bead types. Even with sixteen neutral bead types, the number of functional-group pairs with significant backmapping probability was greater than twenty-five. This resolution limit stems from deliberately ignoring the structural information encoded in the low-dimensional maps in the previous chapter and determining how much chemical specificity could be preserved when only using this 1-D approach. Therefore, in the next chapter of this thesis, we tackle this problem of coarse-graining CCS from the opposite direction of the method used in this chapter: using a bottom-up approach that focuses only on clustering localized geometric environments while essentially ig-

noring any overarching global structure that may be present in our input data set of compounds.

4 Bottom-Up Chemically-Transferable Coarse-Grained Models that Preserve Structure

In Chapter 2, we saw that the high-throughput coarse-grained scheme can be used to quickly generate chemical structure–property relationships when using top-down models like Martini. We additionally demonstrated the means by which different molecular representations and unsupervised machine learning methods could be used to identify key features of CCS. We then made parallels between these unsupervised learning methods and the two prevailing philosophies of coarse-graining based on whether or not our input data set of compounds underwent a global (corresponding to top-down) or localized (corresponding to bottom-up) transformation. In Chapter 3, we explored this comparison between top-down coarse-graining and global transformations applied to CCS by constructing our own top-down coarse-grained models in the Martini mold that were optimized with respect to a single global descriptor. This allowed us to optimize the screening efficiency of our top-down models and relate this quantity to the number of bead types in each model. However, we found a resolution limit that restricted our ability to identify atomistic compounds that would likely map to our new coarse-grained representations—even when tripling the number of bead types used—because we relied only on a single descriptor, the water/octanol partition free energy.

We also saw in Chapter 2 that both the HDBSCAN clustering algorithm and the UMAP algorithm could be used to identify clusters in and reduce the dimensionality of complex representations of CCS. Both of these methods treated the input data set as a connected graph with each data point representing a node, and the edge weights were calculated according to the nearest neighbors surrounding each data point (i.e., a localized transformation rather than a global one, as is done for PCA). We now take the opposite approach from our work in Chapter 3, and transform the CCS using the SLATM vector, a high-dimensional molecular representation that decomposes each compound into a set of localized environments centered on

the heavy atoms making up each molecule (see Sec. 1.1.7). The SLATM vector encodes all 1-body, 2-body, and 3-body interactions within a cutoff radius for each heavy atom. Rather than project CCS onto a single descriptor/dimension, we instead project CCS into the space defined by these SLATM vectors, and use locally-defined distance metrics based on nearest-neighbor distances in this high-dimensional space to cluster local environments across different compounds. Representing CCS using single points from each cluster rather than taking all of the data points allows us to build a minimal coarse-grained model that effectively spans CCS. This approach is similar to traditional bottom-up coarse-graining approaches in which extraneous degrees of freedom are removed from compounds, retaining the essential physics of the higher-resolution system in the coarsened model. In the same vein, reducing CCS to a set of representative points obtained using unsupervised machine learning methods removes extraneous compounds that do not significantly increase the amount of information preserved. In our specific case, this is equivalent to forming a basis in the SLATM vector space defined by the representative SLATM vectors and removing any SLATM vectors that are nearly collinear with any of these basis vectors. In this chapter, we present a method that combines this bottom-up approach to coarse graining CCS with traditional bottom-up coarse graining techniques in order to develop a chemically-transferable coarse-grained model that accurately reproduces the bulk liquid-phase structure of small organic molecules and their mixtures. We show that the approach is successful overall, and identify specific cases for which our transferable coarse-grained model fails and discuss these cases in detail. Finally, we highlight several interesting questions and avenues of further research that stem from this project that will be addressed in future works. As this project is still a work in progress, the results and conclusions presented here are still being validated, and will soon be submitted for publication in a peer-reviewed journal.

4.1 Introduction

In order to facilitate molecular design for a wide variety of applications, there has recently been a growing interest in utilizing data-driven techniques to infer chemical structure–property relationships that span broad regions of chemical compound space [2, 5, 6, 11, 14, 15]. A common rate-limiting step in deriving these relationships is the acquiring of sufficient data so as to ensure the robustness and transferability of the resulting structure–property relationship. As such, a push for increasingly automated workflows for generating data via both experimental and computational methods has risen in tandem with these data-driven approaches. While experimental approaches are limited due to material cost and ease of chemical synthesis, computational methods, in which simulations of chemical compounds

are performed to obtain target properties, do not suffer from these restrictions. Instead, the prohibiting factor for this approach is the computational costs, usually requiring access to a high-performance computing cluster [26, 27]. These methods also require a degree of accuracy and transferability when modelling the chemical compounds of interest. Therefore, cheaper computational methods that can be successfully applied over large swaths of chemical compound space would be extremely beneficial for efficiently constructing structure–property relationships.

Many of the aforementioned data-driven approaches consist of supervised machine-learning techniques applied to organic small molecules in order to predict quantum-mechanical properties over a broad range of chemistries [16–18]. Key to the success of these methods is the use of molecular representations that contain high information content, making it easier to learn relationships for target properties [19, 118, 120]. The best performing representations that have been reported thus far all share certain characteristics. They encode the geometry of a given compound while preserving translational, rotational, and permutational symmetry with respect to atomic ordering. Importantly, they also include information pertaining to the many-body interactions that exist within each molecule. However, it is unclear whether these same representations, which require only a single conformation of a chemical compound, wield the same predictive power when used to estimate the thermodynamic properties of compounds in a bulk liquid phase. In general, thermal fluctuations play a significant role in determining these properties, requiring a computational method, like classical molecular dynamics or Monte Carlo, that enables the calculation of averages that account for multiple conformations in an ensemble [44]. However, in Chapter 2, we were able to use one of these representations, the Spectrum of London Axilrod-Teller-Muto (SLATM) vector, in combination with a kernel ridge-regression model to predict the water/octanol partition free energy despite representing each compound with a single conformation. Rauer and Bereau were also able to use the SLATM vector to predict hydration-free energies of organic small molecules in a similar fashion [187]. This implies that the SLATM vector (and other representations that encode the same information) can be successfully used to predict thermodynamic properties if enough data exists to train the model. However, relatively few training sets for these properties have been reported, primarily due to the high cost of acquiring data. This is especially true when targeting certain structural properties pertaining to bulk-phase phenomena for organic small molecules, which can only be "easily" obtained through computer simulations. Given that the size of CCS for small drug-like molecules is on the order of 10^{60}, the prospect of generating enough data to obtain reliable structure–property relationships for these target properties is daunting [4]. Again, this reinforces the demand for modeling tools that allow us to reduce the cost of simulations for screening purposes.

One such method that has been shown to drastically reduce the computational costs for these types of simulations is particle-based coarse-graining, in which groups of atoms are mapped to coarse-grained particles, known as beads [31]. The interactions that govern the behavior of these beads are assigned so as to retain the essential physics of the higher-resolution data. This results in simulations that are more computationally efficient due to the reduction in number of particles. Additionally, the parameterization of the coarse-grained interactions usually results in a smoothened free-energy landscape that is more easily traversed due to the removal of unimportant atomistic degrees of freedom [31]. In the previous two chapters, we showed that coarse-graining is a highly effective means for generating databases for thermodynamic properties, such that a structure–property relationship can be quickly inferred [110, 138, 227]. In those studies, we took advantage of the chemical transferability of the top-down Martini coarse-grained model, which accurately models the thermodynamic partitioning behavior of small molecules in different media [137]. However, Martini has been known to inaccurately reproduce structural features in many soft-matter systems. For example, certain cross-correlations between beads of different sizes are not included in the model [224]. Additionally, the Martini model fails to properly capture the phase behavior of certain ternary lipid mixtures [228, 229]. This makes it difficult to use the Martini model to screen compounds for applications where structural accuracy is important.

Bottom-up methods, on the other hand, yield high structural fidelity (in some cases by construction) [76–78]. Several studies have been conducted that focus only on a single system or a small number of systems, and much work has been done to elucidate how different bottom-up coarse-grained methods preserve the essential physics of the higher-resolution data [230–232]. Furthermore, certain bottom-up methods have been shown to ensure the transferability of a coarse-grained model across multiple thermodynamic state points, known as an extended ensemble. Dama et al. recently developed a method in which coarse-grained force fields parameterized at different densities are mixed in order to successfully model complex phase behaviors, such as the vapor-liquid equilibrium of a Lennard-Jones fluid [233]. Mullinax and Noid demonstrated a coarse-grained potential that was transferable across liquid-state binary mixtures of organic compounds [40]. Sanyal and Shell used local-density potentials to do the same for benzene-water solutions [41]. As far as we know, there has only been one reported study in which a chemically-transferable coarse-grained model was derived using a bottom-up approach. Sanyal et al. recently developed a new extended-ensemble relative-entropy method to develop a generalized coarse-grained protein-backbone model that could accurately reproduce the structures of over 200 different globular proteins [42]. Even in this case, however, the native contact information for each protein was

also required. Furthermore, there are practical considerations against employing these methods for high-throughput screening applications, as they require higher-resolution data to be initially generated for each compound of interest. Thus, there is demand for chemically-transferable coarse-grained potentials that preserve structural accuracy without requiring the generation of higher-resolution data in order for high-throughput coarse-grained screening applications to be feasible.

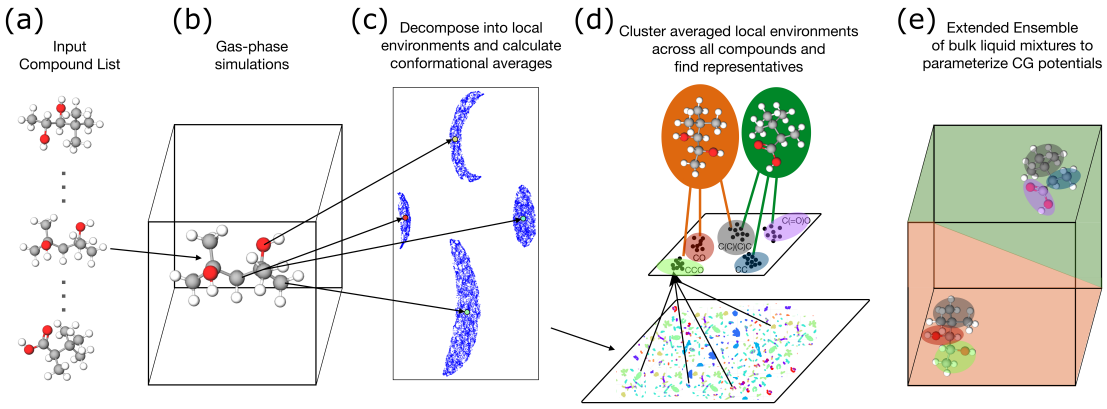

Figure 4.1: A schematic showing the protocol used to develop a bottom-up chemically-transferable coarse-grained model. (a)Given a set of compounds as input, (b) we first run gas-phase MD simulations of each compound. (c) We then decompose the gas-phase trajectories into a series of local environments and compute the conformational average of each unique local environment found in the trajectory. (d) We take these averaged local environments and cluster them across all compounds. (e) We use the clustering to find representative molecules, to which we apply extended-ensemble bottom-up coarse graining methods in order to obtain chemically-transferable coarse-grained potentials.

In this work, we present a new data-driven framework for creating chemically-transferable coarse-grained models with structural accuracy. Our workflow is shown in Fig. 4.1. Given a set of chemical compounds, this method applies unsupervised machine-learning methods in order to determine a subset of "representative" compounds that have the greatest probability of sharing features in configurational space with the most compounds in the remainder of the data set. This is accomplished by first running gas-phase simulations of each molecule in the input database. The resulting trajectories are decomposed into a set of local environments centered on each heavy atom, and averaged over the conformational space sampled by each molecule during the simulation. Density-based clustering methods are then applied to the conformationally-averaged local environments

of all of the compounds in the database. Compounds which contain the most promiscuous local environments—meaning the compound is found in many different clusters of local environments—are taken as "representative". We then apply the extended-ensemble MSCG method to this set of representative compounds. The resulting coarse-grained potentials are constructed to be transferable across the chemical space defined by the input data set using a limited number of bead types. We apply this protocol to the subset of C_7O_2 isomers in the Generated Database (GDB) and begin to quantify the extent to which the transferable potentials remain accurate [140, 161]. We quantify the accuracy of the transferable coarse-grained model by comparing the radial distribution functions (RDFs) of the coarse-grained system to its atomistic counterpart. We further parameterize coarse-grained force fields using the MSCG method but limited to specific state points, as would be traditionally done when developing a coarse-grained model for a single system of interest, and use these models as a benchmark for our transferable model. This is the first study in which the MSCG method has been applied to a large number of systems (703 in total) in an automated fashion. Our preliminary results show that applying this protocol does indeed result in transferable coarse-grained potentials with structural accuracy. Surprisingly, parameterizing the coarse-grained model by averaging over the entire extended ensemble led to drastic improvements in accuracy when compared to the coarse-grained model parameterized specifically for that state point. On the other hand, there were certain systems where the ensemble-averaged force field performed significantly worse than the state-point specific model. We examine each of these cases and put forward hypotheses explaining the observed trends. Specifically, we find that gains in accuracy are due to a "regularization-like" effect that effectively smoothens the average forces acting on specific coarse-grained bead types, reducing the effect of certain interactions that would otherwise dominate the system. In cases where the ensemble-averaged force field performs poorly compared to the state-point specific force field, we identify functional groups with complex interactions that result in highly diverse conformational states. Averaging over the whole ensemble for these functional groups results in a potential that is unable to reproduce any of these diverging conformational states. Overall, we provide a systematic means to perform a bottom-up coarse-graining of CCS, resulting in chemically-transferable coarse-grained potentials that retain structural accuracy in simulations of soft-matter systems. At the same time, we highlight the limitations of this approach and note key pitfalls to avoid when implementing this approach.

4.2 Methods

4.2.1 Database

We selected a subset of the Generated Database (GDB), a computer-generated set of drug-like organic compounds, to test our data-driven bottom-up approach [140, 161]. Specifically, we selected the set of GDB compounds which were made up of seven carbon atoms and two oxygen atoms only. We further filtered out any compounds containing triple bonds. After applying these filters, we were left with a database of 3441 C_7O_2 isomers, listed in their simplified molecular-input line-entry system (SMILES) format. Despite restricting the size of the molecules and only including three elements (C, O, and H), a large variety is still present in the resulting chemical structures. Furthermore, complex interactions, such as hydrogen-bonding and π-stacking interactions, are also present for many of the compounds in this database. Because the database was limited in terms of the chemical elements, but still contained compounds which we expected to display complex behavior in the bulk phase, we felt this choice of database would prove useful for determining which specific physical interactions would be (un)successfully captured by our chemically-transferable model.

4.2.2 Gas-phase simulations

For each compound in the database, we first ran single-molecule gas-phase molecular dynamics simulations. The initial structures were obtained by converting the molecules from their SMILES string representations to energy-minimized 3D conformations using the RDKIT package [234]. The force field parameters for each compound were generated using the CGENFF tool, included in the SILCSBIO 2018 package, which automatically assigns parameters from the CHARMM General Force Field based on the input chemistry [235]. The simulations were run at constant volume using a stochastic velocity-rescaling thermostat[67] to maintain a constant temperature, $T = 300$ K. The simulations were run using a 2 fs timestep for a total of 3 ns, with the LINCS algorithm used to constrain the hydrogen atoms [69]. A frame was output every 2 ps, yielding 1500 frames per simulation for each compound in the database. The GROMACS 16.1 package was used to run all of the systems simulated in this work at the atomistic resolution [236].

4.2.3 Defining local environments with SLATM

The Spectrum of London Axilrod-Teller-Muto (SLATM) vector describes a molecule as a sum of atomic environments that encode the 1-body, 2-body, and 3-body interactions within a cut-off distance [120, 121]. For a full description of this rep-

resentation, refer to Section 1.1.7, as we provide only a cursory explanation here. For each atom, its corresponding SLATM vector consists of the elemental atomic number (1-body), a spectrum of 2-body London interactions convoluted with a gaussian function (2-body), and a spectrum of 3-body Axilrod-Teller-Muto interactions also convoluted with a gaussian function (3-body). The two-body spectrum is computed over the distance as a London interaction between all pairs within a cut-off value with a specified step-size. Similarly, the three-body spectrum is computed as an Axilrod-Teller-Muto interaction over the angle for all triplets within the cut-off distance.

As seen in Fig. 1.7b, the atomic SLATM (aSLATM) vector is a concatenation of the 2-body and 3-body spectra with the atomic number of the atom. Note that this concatenation is performed over all possible many-body types found with respect to all atomic elements found in the input data set. For example, the 1-body component of a SLATM vector corresponding to a carbon atom in one of the compounds used in this database would be $[0.0, 6.0]$, whereas an oxygen atom would have a 1-body component of $[8.0, 0.0]$. This would then be followed by all possible two-body interactions (O-O,O-C,C-C), and subsequently all possible three-body interactions. If an atom does not have some of these interaction types (for example, a carbon atom will not have any O-O interactions), the vector is filled with zeros. This representation uses only the internal geometry of the molecule, making it translationally and rotationally invariant. Additionally, because the aSLATM vector carries the same format for each of the atom types, and the spectra do not depend on the permutational ordering of the atoms, the aSLATM vector is also permutationally invariant. Convoluting the interactions with Gaussian functions results in a continuous and differentiable metric. Furthermore, including the 3-body spectrum was shown to vastly improve the performance of statistical models that used the SLATM vector to predict quantum-mechanical properties [120]. Because of the different scaling applied to the different many-body types, a hierarchical structure is built in to this representation, guaranteeing that aSLATM vectors are separated first by the atom type, then by the positions of the nearest neighbors, and finally, the relative positions of the heavy atoms remaining within the cut-off radius. We applied the QML package made for PYTHON 2.7 to convert our database of compounds into aSLATM representations [179]. The default values, which were optimized for predicting quantum-mechanical properties, were used, with a cutoff value of 0.48 nm and a grid spacing of 0.003 nm and 0.03 radians for the 2-body and 3-body spectra, respectively.

Each frame of the gas-phase simulations is used to generate nine atomic SLATM vectors, with one vector per heavy atom. Hydrogen atoms were not included when calculating the SLATM vectors. Because the number of heavy atoms and chemical composition was constant across the entire database, the length and ordering of

the many-body types for each aSLATM vector was the same. Fig. 4.1c shows the aSLATM vectors of all 1500 frames that were output from a gas-phase simulation of the first molecule in our database projected into two dimensions using UMAP [106]. Note that there are only four large clusters due to the symmetry of the compound; both oxygen atoms, the terminal carbons, and the carbons bonded to the oxygen atoms are all symmetric with respect to the carbon in the middle of the compound (see Fig. 4.1b). Using HDBSCAN, we were able to easily identify the clusters shown above in an automated fashion (insensitive to the choice of HDBSCAN parameters), as was the case with all of the gas-phase results [98]. For a full description of how the HDBSCAN algorithm works, please refer to Chapter 1, as we provide only a cursory description here. The data is treated as a connected graph with data points representing nodes. The edges connecting these nodes are weighted according to a localized distance metric that depends on the nearest neighbor distances for each point in the data set. Rather than take a single cut-off length-scale or cut-off density as input, HDBSCAN requires the size of the smallest possible cluster to be defined in addition to the number of nearest neighbors accounted for when reweighting the graph edges. A dendrogram is then calculated that spans the entire data set, and the stability of clusters is determined by how "long-lived" they are as the furthest points from the cluster center are systematically removed until the minimum cluster size is reached. This "lifetime" metric essentially answers the following question: if the furthest assigned data point were removed, would the remaining data set still be considered a single cluster, or would it have to be split into separate clusters? The final clusters that are chosen are those that are the most stable under this criterion. After identifying clusters, we compute the cluster center as being the average of all aSLATM vectors that make up the cluster, and select the aSLATM vector that is closest to this average for each cluster. Therefore, the molecule shown in Fig. 4.1 is represented as four aSLATM vectors which are the conformational averages of the unique local environments which make up the molecule. We similarly apply this protocol to all the gas-phase trajectories and represent each of the 3441 molecules by their aSLATM cluster centers.

4.2.4 Selecting representative molecules

All of the aSLATM cluster centers obtained from the gas-phase trajectories were combined into a single data set and clustered using HDBSCAN. Here, we used the default HDBSCAN parameters, with both the minimum cluster size and number of nearest neighbors set to five points. Fig. 4.2 shows a UMAP of this data set colored by element type. Within this separation by atom type, separate clusters correspond to oxygen or carbon atoms that make up certain functional groups as well as their placement within a carbon scaffold (i.e., edge of the molecule versus middle of the molecule). Note that this UMAP is used only for visualization

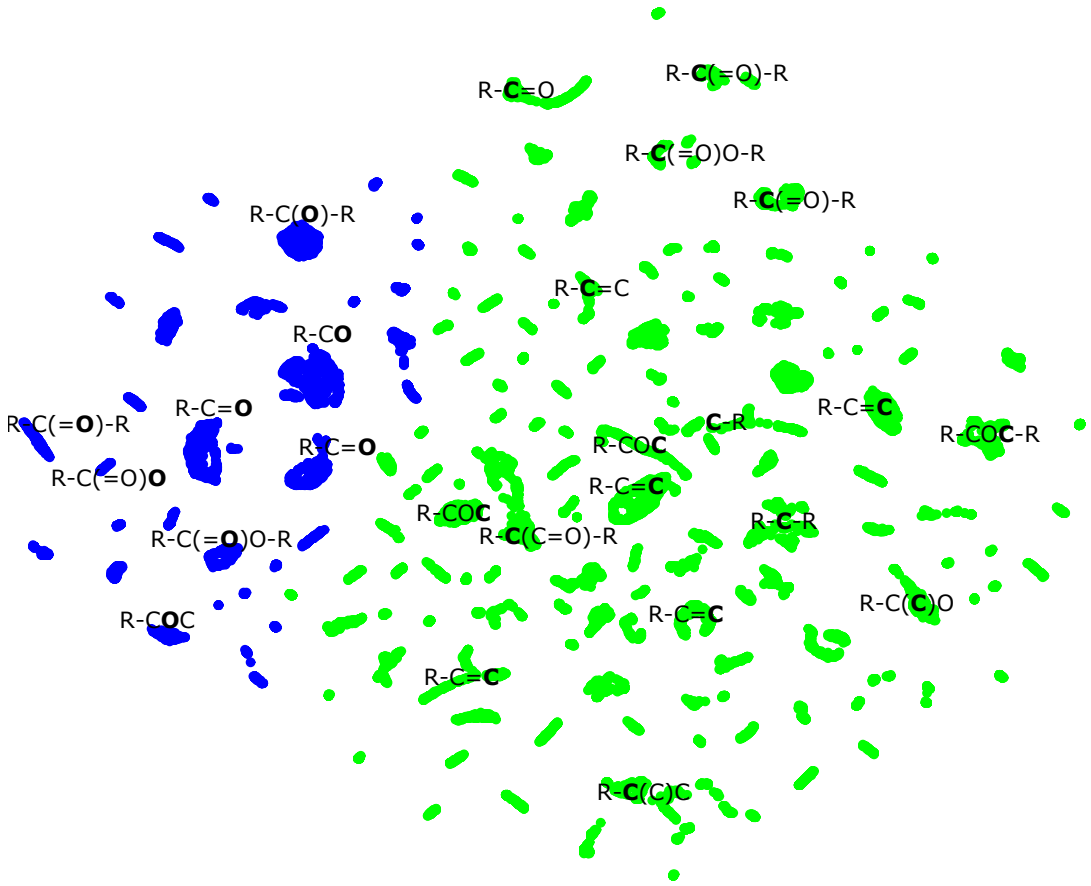

Figure 4.2: A UMAP projection of the averaged aSLATM vectors obtained from gas-phase trajectories of the 3441 C_7O_2 isomers used in this work. Points colored green represent aSLATM vectors corresponding to carbon atoms, whereas blue-colored points correspond to oxygen aSLATM vectors. Several large clusters are labeled as the functional group that is prominently featured in that cluster. The emboldened character corresponds to the specific atom that the aSLATM vector is centered on.

purposes, and the identification of clusters was performed in the high-dimensional space. We further note that, beyond the overall separation of aSLATM vectors based on atom type, no further global trends are seen across the various clusters defined. Although we only provide labels for a small fraction of the clusters identified in Fig. 4.2, we saw that most of the distinct clusters that are present in the UMAP are also labeled as distinct clusters according to our HDBSCAN results on the high-dimensional data. Because we were also able to identify the key chemical motifs that define these clusters via visual inspection, we are confident in the accuracy of the clustering results. We then chose representative molecules by first ranking them by the number of clusters "visited." We then included subsequent molecules if the number of new clusters visited by the molecule was greater than the number of clusters already visited by the other chosen molecules. By applying this simple algorithm, we found 19 molecules containing local environments that shared cluster assignments with over 92% of the assigned aSLATM vectors. These nineteen representative molecules, shown in Figs. 4.3–4.6, were then used as the foundation for our extended-ensemble approach.

4.2.5 Atomistic simulations of bulk liquid-phase binary mixtures

An extended ensemble consisting of bulk liquid-phase molecular dynamics simulations of each of the 19 representative molecules, as well as binary mixtures of the representative molecules, was constructed. Each system consisted of 400 molecules in total, with the concentrations for compounds in the binary mixtures ranging from 20% to 80% in 20% increments. Therefore, the number of state points simulated at the atomistic resolution was 19 (for each of the pure systems) plus every possible binary mixture of each of the compounds at the four different concentrations specified, yielding a total of 703 state points making up the extended ensemble.

Each of these 703 systems was simulated using the following protocol, which adapts many of the steps taken by Dunn and Noid [84]. 400 molecules were first randomly placed into an isotropic box with a volume of 1000 nm^3. The system was energy-minimized and then run in the NVT ensemble using a velocity-rescaling thermostat for 2 ns at a temperature of 1000 K [67]. The system was then cooled to 300 K over the course of the next 10 ns. At this point a Berendsen thermostat and barostat were used to reduce the size of the box and equilibrate the system in an NPT ensemble at 300 K and 1 bar [237]. The resulting densities ranged from 0.80 g/cm^3 to 1.0 g/cm^3. While no specific density data could be obtained for these 19 representative molecules, these densities roughly agree with those of 1,7-heptanediol (0.95 g/cm^3), heptanoic acid (0.92 g/cm^3), and pentyl acetate (0.87 g/cm^3), which also consist of 7 carbon and 2 oxygen atoms [238]. In a similar vein, we were unable to find previously-reported isothermal compressibilities for

these specific compounds, and used the isothermal compressibility of heptanoic acid, $7.4 \cdot 10^{-5}$ bar^{-1} for all systems [239]. Production runs were then carried out under these conditions in an NPT ensemble using a Nosé-Hoover thermostat and a Parinello-Rahman barostat with coupling constants of $\tau_T = 0.5$ ps and $\tau_P = 5.0$ ps, respectively [84]. The force field parameters used were the same as those used in the gas-phase simulations, with LINCS constraints applied to the hydrogen to heavy-atom bonds. The final trajectories consisted of 60 ns simulations of each system, of which the first 5 ns were discarded to allow for equilibration after applying the new thermostat and barostat. In machine learning parlance, this data is effectively the training data over which we will optimize our transferable coarse-grained model.

4.2.6 Applying the multi-scale coarse-graining technique

Bead Type	Fragment	Bead Type	Fragment	Bead Type	Fragment
B01	CC	B06	CCO	B11	C=CO
B02	CO	B07	COC	B12	OC=O
B03	C=C	B08	OCO	B13	C(C)(C)C
B04	C=O	B09	CC=C	B14	C(C)(C)O
B05	CCC	B10	CC=O		

Table 4.1: Bead types and their corresponding fragments in SMILES notation.

For a detailed description of the MSCG method, refer to Section 1.1.5, as we only provide an overview here. The first step in the coarse-graining process is to define a mapping function [31]. This function assigns atoms in the high resolution MD trajectory to pseudo-atoms called beads, and sets their configuration according to some assignment rule. Since the total degrees of freedom are reduced in the CG system, it is impossible to preserve all of the features of the high resolution system, regardless of the accuracy of the potentials assigned to the CG beads. This makes the choice of which atomistic fragments should be mapped to a single bead an important one, although, in practice, this is often based on chemical intuition alone. The analysis of the clusters shown in Fig. 4.2 naturally points to a mapping scheme in which specific functional groups are each assigned their own bead type. Therefore, we adopted a mapping scheme in which all combinations of two-heavy-atom and three-heavy-atom fragments consisting of carbon and oxygen are assigned to different bead types, as shown in Table 4.1. In order to ensure a "complete" mapping for all the compounds in our training set—meaning that all heavy atoms are assigned to a bead type and the topology of the fragments

are preserved—we also included two fully-branched bead types which mapped to four-heavy-atom fragments. This set of bead types allowed for many molecules in our training set to be mapped in multiple ways. The full set of training compounds as well as their coarse-grained mappings is shown in Figs. 4.3–4.6. Although the mappings shown in these figures range from circular to ellipsoid in shape, the potentials assigned to each bead type are radially symmetric.

The next step in the coarse-graining process is to determine the coarse-grained potential. When taking a bottom-up approach, the goal is to ensure that the probabilities of obtaining coarse-grained configurations are the same as the corresponding atomistic configurational probabilities [31, 230]. By equating these probabilities and solving for the coarse-grained potential yields a projection of the atomistic free energy surface onto the coarse-grained degrees of freedom, known as the many-body potential of mean force (MBPMF) [31].

The MBPMF is a projection of the atomistic free energy surface onto the coarse-grained degrees of freedom, as specified by the mapping function. Note that there is no restriction on the functional form of the MBPMF. The "many-body" aspect of the MBPMF refers to the fact that integrating over the residual atomistic degrees of freedom results in the generation of many-body interactions between CG beads. Therefore, conventional CG potentials, which are expressed as a sum of pair-wise contributions, will never fully approximate the MBPMF [31].

In this work, we use the Multiscale Coarse Graining (MSCG) approach, which aims to produce a coarse-grained potential that best approximates the MBPMF via a variational approach [77]. The functional to be optimized via the MSCG approach requires that the optimal CG potential will be the one that best reproduces the averaged net force acting on mapped CG sites from the atomistic trajectory. For this reason, the MSCG approach is also commonly referred to as the force-matching method for bottom-up coarse-graining. A major advantage of this variational approach is that it does not restrict the functional forms used to express the CG potential. Instead, the mean force is expressed in terms of force-field basis vectors. This enables the force-matching functional to be rewritten as a system of linear equations:

$$\sum_{D'} G_{DD'}\phi_{D'} = b_D, \tag{4.1}$$

where D denotes a single interaction type at a specified distance. In this equation, $G_{DD'}$ is a symmetric matrix called the metric tensor that measures the cross-correlations between all atomistic interactions when projected onto the force-field basis vectors defined. b_D is a vector obtained by projecting the MBPMF of the atomistic reference onto these force field basis vectors. Solving equation 1.78 yields the weights $\phi_{D'}$ corresponding to the optimal CG potential that minimizes the force-matching functional.

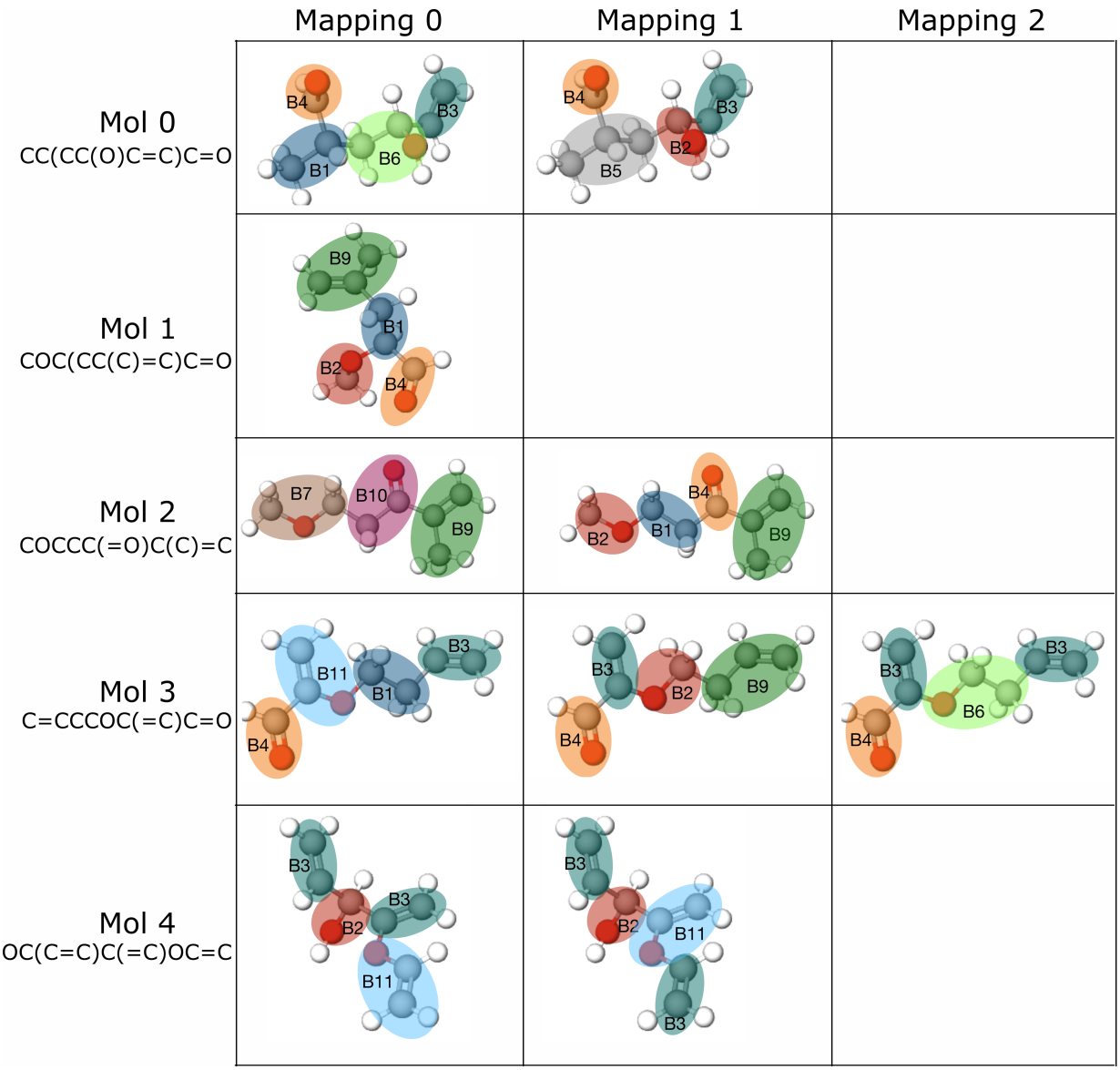

Figure 4.3: Representative molecules 0-4 used to make the extended ensemble, as well as each coarse-grained mapping applied to the molecule.

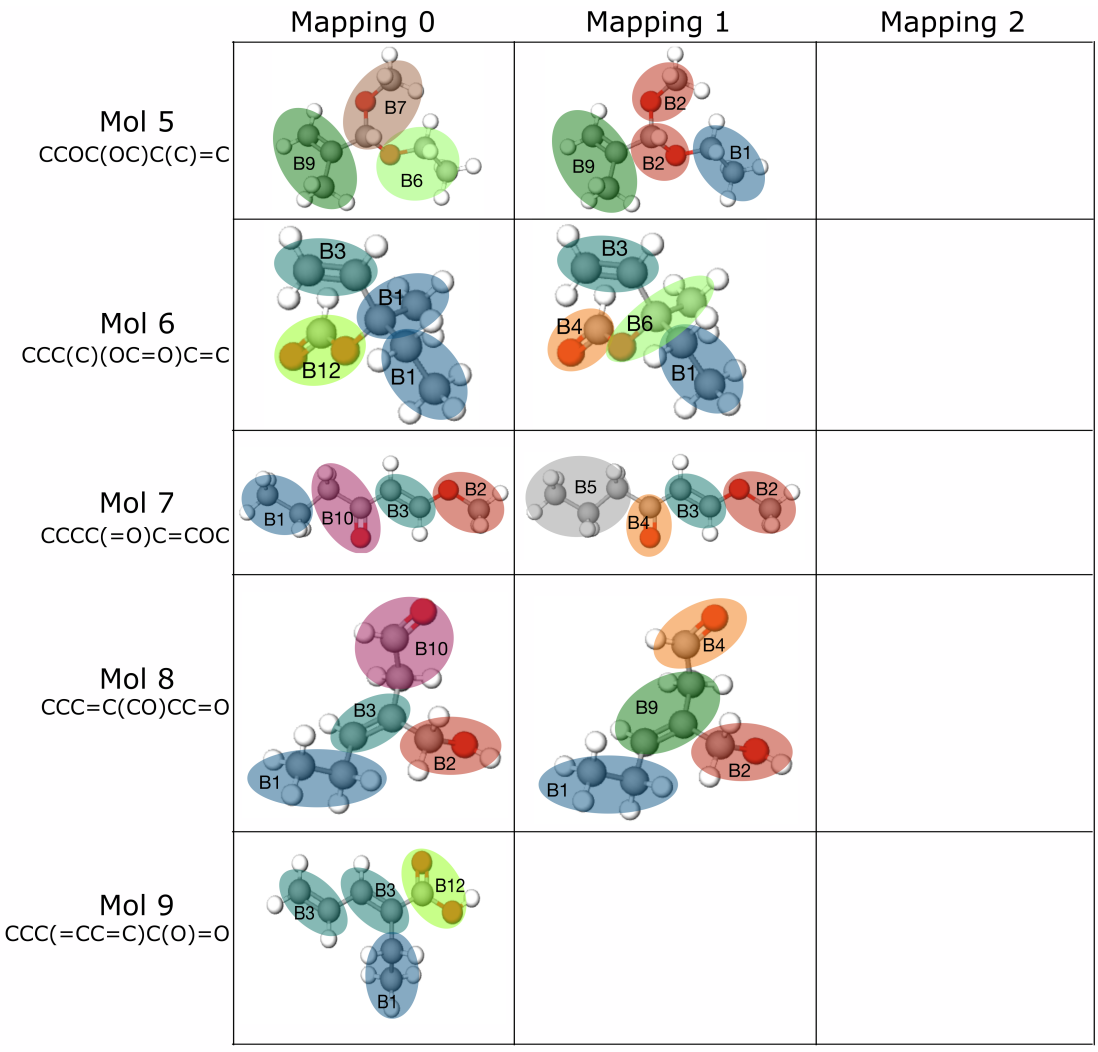

Figure 4.4: Representative molecules 5-9 used to make the extended ensemble, as well as each coarse-grained mapping applied to the molecule.

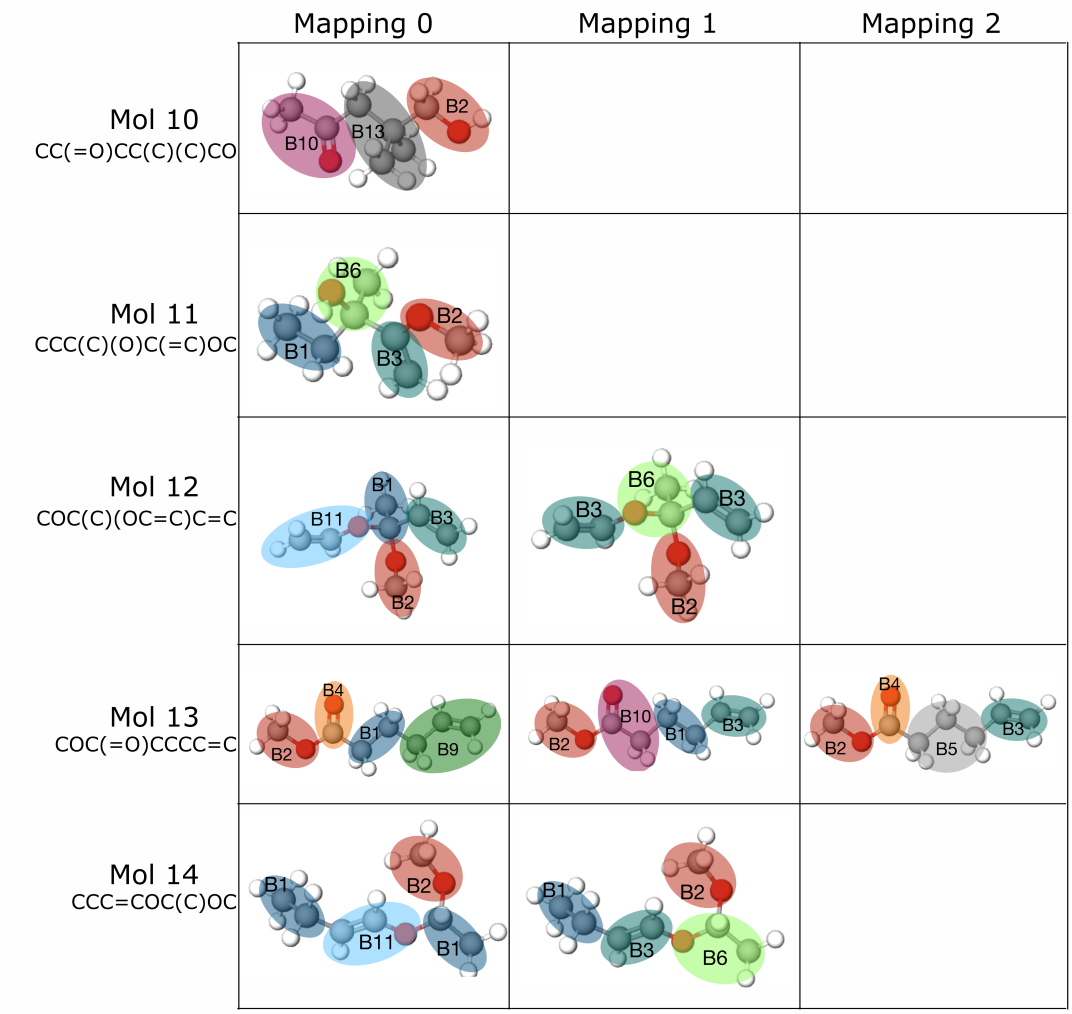

Figure 4.5: Representative molecules 10-14 used to make the extended ensemble, as well as each coarse-grained mapping applied to the molecule.

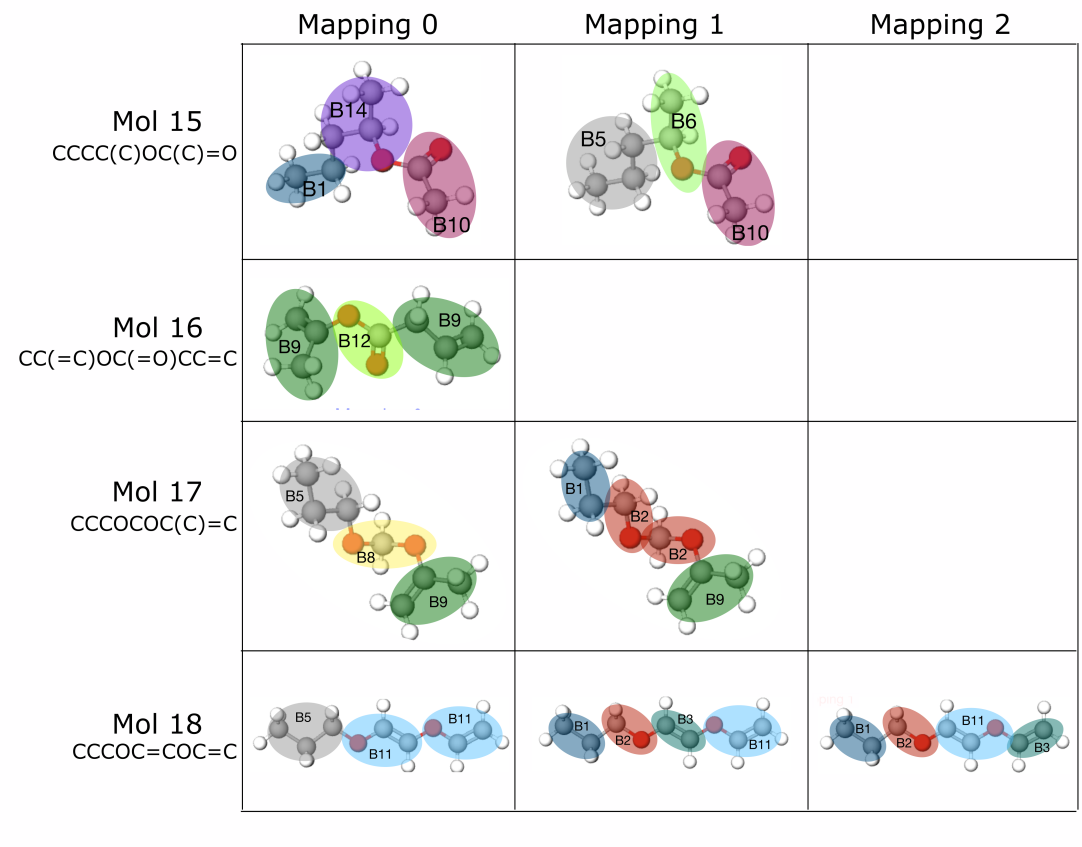

Figure 4.6: Representative molecules 15-18 used to make the extended ensemble, as well as each coarse-grained mapping applied to the molecule.

In practice, we used the BOCS software package developed by Dunn et al. to apply the MSCG method to each of the 703 state points in the extended ensemble [83]. For systems made up of compounds with multiple mappings, we systematically applied every possible mapping (or combination of mappings in the case of binary mixtures) and calculated the MSCG potential from each mapped atomistic trajectory. We first applied the direct Boltzmann inversion method (see 1.1.5 for details on this technique) in order to obtain intramolecular coarse-grained potentials (i.e., bonded, angle, and dihedral potentials) from each atomistic trajectory. In cases where the distribution of coarsened angles and dihedrals was not sampled for certain angle and dihedral values, we modified the resulting potential to include large barriers, effectively preventing the coarse-grained systems from sampling these values as well. We then used these potentials to calculate the contribution of the intramolecular interactions to the mean force, and subtracted them before solving equation 4.1, including only the nonbonded interactions and bonded interactions only. This ensured that the intramolecular contributions to the mean force were not incorrectly attributed to the pairwise nonbonded interactions. Bonded interactions are included in the calculation even after subtracting their contribution to the mean force for reasons of numerical stability [232, 240]. Including these interactions does not affect the results of the nonbonded calculations due to the large energy-scale separation between these interaction types, which also ensures that the force field basis vectors are decorrelated in the metric tensor. We represented these pairwise interactions as radially-isotropic fourth-order Basis splines with control points spaced every 0.01 nm ranging from 0.0 to 1.4 nm. Thus, a set of coarse-grained pairwise potentials was generated for each mapping at each state point. This protocol was applied using an automated framework, and, to the best of our knowledge, this is the first study in which such a large number of systems has been systematically coarsened using the MSCG method.

4.2.7 Averaging over the extended ensemble

Because a variational approach is used to find the potential that best approximates the MBPMF in the MSCG method, it is a simple matter to extend the variational principle over an extended ensemble consisting of multiple thermodynamic state points [40]. An average over the extended ensemble is defined in the following manner

$$\langle a_\gamma(\boldsymbol{r}_\gamma)\rangle = \sum_{\gamma}^{\Gamma} p_\gamma \langle a_\gamma(\boldsymbol{r}_\gamma)\rangle_\gamma, \tag{4.2}$$

where γ denotes the specific state point, and p_γ gives the probability of being in that ensemble, set to $1/\Gamma$, where Γ is the total number of state points in the extended ensemble. Each ensemble will have its own mapping, and a corresponding

MBPMF. In this case, solving the force-matching functional yields the potential that best approximates all of the MBPMFs over the extended ensemble, averaged as shown in the above equation. The derivation proceeds just as described in the previous section, while additionally taking the sums over all Γ corresponding to the different state points [40, 83].

In practice, we first create a metric tensor $G_{DD'}$ and mean force vector b_D that corresponds to all 105 pairwise interactions between the fourteen bead types that we have defined as well as all bonded interactions (to ensure numerical stability). We then iterate over all of the state points and mappings, adding each of the blocks of the metric tensor and segments of the mean force vector for a single state point to the corresponding block and segment in the extended ensemble metric tensor and mean force vector, respectively. We then compute the average for each row/column of the metric tensor and segment of the mean force vector by dividing by the number of state points that contained that specific interaction. This prevents interactions that are less represented in the extended ensemble from being neglected due to their lower contribution to the extended-ensemble mean force. We are currently deriving Equation 4.2 for the case where the extended ensemble consists of different systems that do not all have the same interaction types in order to prove that this averaging approach also obeys the variational principle. Using the BOCS software package, we solved Equation 4.1 with the extended ensemble metric tensor and mean force vector, yielding coarse-grained potentials that should be transferable across our input data set of 3441 C_7O_2 isomers.

4.2.8 Validation and quantifying structural accuracy

After obtaining the coarse-grained potentials by averaging over the extended ensemble and solving Equation 1.78, we validate our new model by first running coarse-grained simulations of the systems that make up the extended ensemble. This is done using both the state point specific (SP) potentials that would be used if applying the MSCG method to that specific state point only, as well as the extended-ensemble (EE) potentials. The intramolecular potentials used for both coarse-grained models are those obtained from the direct Boltzmann inversion of the intramolecular distributions calculated from the atomistic trajectories for each system. The coarse-grained simulations are run in the NVT ensemble using an isotropic box that has dimensions matching the average density calculated from the atomistic state point trajectory. A time step of $\delta t = 0.002\ \tau$ was used for all simulations, where τ is the natural time unit for the propagation of the model defined in terms of the units of energy $\mathcal{E}$, mass $\mathcal{M}$ and length $\mathcal{L}$ as $\tau = \mathcal{L}\sqrt{\mathcal{M}/\mathcal{E}}$. The simulations were run for 5 million time steps, with every 500th frame saved as output, and the first 500 output frames were discarded. The GROMACS 5.1 package

was used to run all coarse-grained simulations in this work [68]. We observed a speed-up factor of ~ 3.0 when comparing the coarse-grained to the atomistic simulations (with the coarse-grained simulations running at ~ 0.35 ns/CPU hour), but larger time steps can be used to simulate the coarse-grained systems, resulting in a further reduction of computational cost [84].

In order to assess the effectiveness of the EE potentials, we first calculate radial distribution functions ($g(r)$, RDFs) using the atomistic trajectories as well as the two coarse-grained trajectories. We then quantify the agreement between the coarse-grained and the atomistic RDFs by using the Jensen-Shannon divergence (JSD) [115]. The relative entropy framework has been previously established as a useful tool for evaluating the quality of coarse-grained models [78, 215]. The JSD was previously used in Chapter 3 to evaluate the agreement between distributions of partitioning free energies for atomistic compounds and their coarse-grained counterparts. Here, we again use this metric to quantify the agreement between our coarse-grained and atomistic RDFs. While the Kullback-Leibler divergence (D_{KL}) [216] directly relates two distributions, the JSD computes the relative entropy by comparing each of these distributions to a third distribution which is the average of the other two distributions, as shown in the following equations

$$D_{\mathrm{JS}} = \frac{1}{2} D_{\mathrm{KL}}\left(g(r)_{\mathrm{CG}} || g(r)_{\mathrm{avg}}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(g(r)_{\mathrm{AA}} || g(r)_{\mathrm{avg}}\right), \tag{4.3}$$

$$\text{where } D_{\mathrm{KL}}(g(r)_{\mathrm{A}} || g(r)_{\mathrm{A}}) = \sum_{r=0}^{r_{max}} a(r) \ln\left(\frac{a(r)}{b(r)}\right),$$

$$\text{and } g(r)_{\mathrm{avg}} = \frac{1}{2}(g(r)_{\mathrm{CG}} + g(r)_{\mathrm{AA}}).$$

In the above equations, we define D_{KL} in terms of two arbitrary RDFs, $g(r)_{\mathrm{A}}$ and $g(r)_{\mathrm{A}}$ ranging from $r = 0$ to r_{max} with values of $a(r)$ and $b(r)$ for the given radial distance values. For all RDFs, we used a grid spacing of 0.01 nm and an $r_{max} = 1.5$ nm. All RDFs were calculated using the GMX RDF package included in GROMACS 5.1. After computing these RDFs for all three trajectories, the JSDs for both the SP and EE are calculated by comparing their respective RDFs to the corresponding atomistic RDFs.

Fig. 4.7 provides examples of atomistic and coarse-grained RDFs along with the JSD value that quantifies the discrepancies between the two. Fig. 4.8 shows the JSD values averaged over all interaction types for a specific pure molecule system (pure signifies the data does not come from any of the binary-mixture state points) and mapping, whereas Fig. 4.9 shows the same results, but averaged with respect to specific interaction types over all of the pure systems. Thus, the JSD provides a convenient method for quantifying the overall accuracy of the transferable potentials compared to the state-point specific potentials.

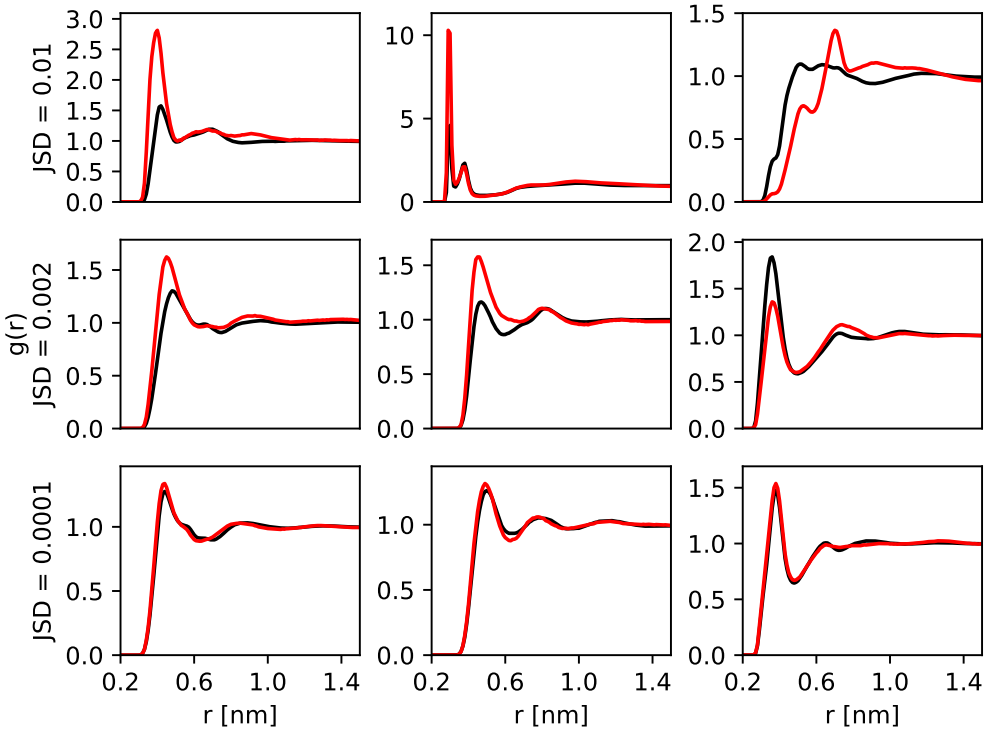

Figure 4.7: Examples of atomistic (black) and coarse-grained (red) RDFs with JSD values of 0.01 (top row), 0.002(middle row), and 0.0001 (bottom row). 0.002 is used in the rest of this work as the cut-off value for "good" agreement.

4.3 Results

We calculated atomistic and coarse-grained RDFs, using both the state-point specific (SP) model for each system as well as the model obtained from the extended-ensemble (EE), for all of the 19 single molecule systems. Because many molecules have multiple coarse-grained mappings (see Figs. 4.3-4.6), the total number of systems considered was 36. All of the intramolecular interactions (bonded, angle, dihedral) are the same for both SP and EE models for a given system, taken by performing direct Boltzmann inversion using intramolecular distributions obtained from the liquid-phase trajectories as detailed in the Methods section. The same comparisons for the rest of the ensemble (i.e., the remaining 684 binary mixture systems) are still underway, but we expect many of the errors seen in the pure systems to also occur in the mixed systems.

Despite only examining 36 mappings here, each mapping has six RDFs on average, meaning the total number of RDFs to compare is over 200. If one were to include the additional 684 binary mixture systems and all of their mappings, this would result in over 2,700 state points with an average of 12 RDFs per state point. Therefore, it is unfeasible to qualitatively scrutinize each and every RDF generated in this work (although all RDF data for the entire ensemble is currently being generated and will be accessible using a data-sharing service like ZENODO in the final version of this work). Instead, we use the JSD to quantify the accuracy of the SP and EE coarse-grained RDFs relative to the atomistic RDFs and average these JSD values over all the RDFs for each state point. Fig. 4.7 provides a useful reference for interpreting these JSD values in terms of the error when comparing atomistic and coarse-grained RDFs. Fig. 4.8 shows these averaged JSD values for each of the 36 state points (which includes all mappings) making up the pure systems. Also shown in this figure are the total mean and variance calculated using all the RDFs that make up these 36 state points for both the SP and EE models. One might expect the EE model to perform worse than the SP models because the EE model is obtained by averaging over many different state points, which include binary mixtures, rather than only using information from the single pure system that it is approximating. However, on average, the transferable EE model outperforms the SP models with an average JSD value of 0.0024 versus 0.0038, respectively. Indeed, we see several state points for which the EE model greatly outperforms the SP model (Molecule 3 mapping 0, Molecule 8 mapping 0, Molecule 1 mapping 0). However, some instances of the reverse case, in which the SP model shows better agreement with the atomistic structure compared to the EE model, is also seen (Mol6 Map0, Mol5). Interestingly, the variance of the EE model (shown in the legend of Fig. 4.8) is also significantly smaller than that of the SP models, indicating that there is more regularity in the quality of the EE model.

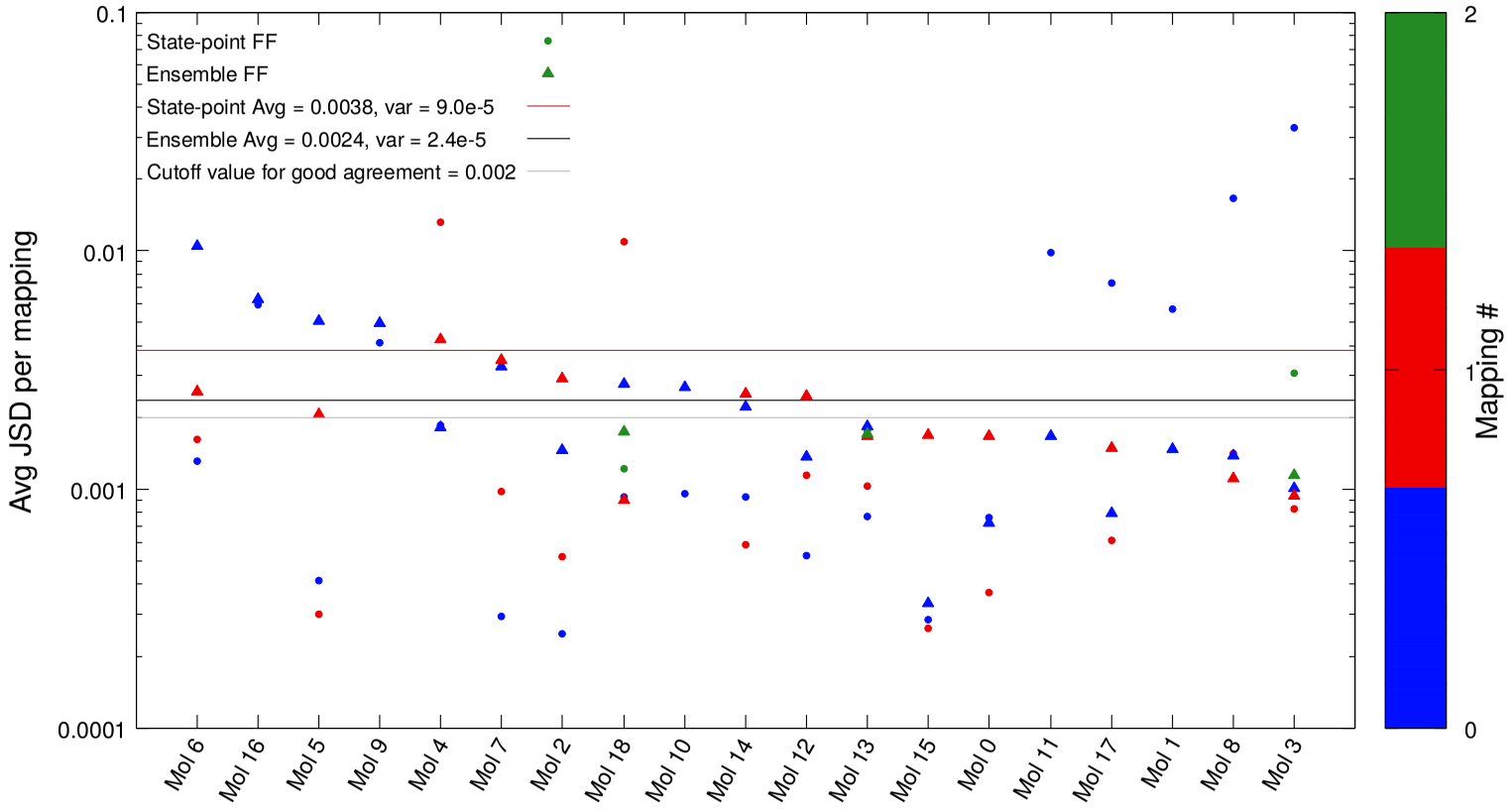

Figure 4.8: Average JSD values for each molecule and mapping, calculated for all pure systems (i.e., no binary mixtures) using both the state-point specific coarse-grained models (circles) as well as the transferable coarse-grained model obtained by averaging forces over the whole extended ensemble (triangles). Different colors correspond to different mappings of the same molecule, which match the labels shown in Figs. 4.3-4.6. The brown line corresponds to the average JSD over all systems simulated using the state point specific models, and the black line is the same but using the ensemble model. The grey line denotes the cutoff JSD value for "good" agreement with atomistic RDFs, 0.002. The molecules are ordered based on increasing agreement of the atomistic RDFs with the extended-ensemble coarse-grained RDFs, with the highest-JSD molecule on the left, and the lowest-JSD molecule at the right.

Next, we take the same data set and the same RDFs and average not according to molecule and mapping, but according to interaction type, as shown in Fig. 4.9. While there are many instances of the EE model performing worse than the SP models, the majority of the EE results remain under the 0.002 cutoff JSD value, whereas there is a much more even split when considering the SP results. Because the JSD quantifies accuracy on a logarithmic scale, the average JSD will be primarily influenced by the worst-performing RDFs. Since many of the systems observed in Fig. 4.8 do not fall under the cutoff value of 0.002, we conclude that the overall disagreement between the atomistic and coarse-grained structures for these systems stems from the presence of just one or two coarse-grained RDFs that fail to match their atomistic counterparts.

Molecule Num	SMILES	scaled SLATM distance from training set
19	CCC(CC)OC(C)=O	0.43
20	CC(C)=CC(=C)C(O)=O	0.48
21	C=COC(=C)C(=C)C=O	0.88
22	CC(C)(C)C(C=O)C=O	0.91
23	CC(C)C(C)(C=O)C=O	0.91

Table 4.2: Test molecules, their SMILES strings, and their SLATM distance to the training set scaled by the maximum possible distance.

Finally, we consider five molecules that were not included in the 19 molecules, which make up the "training" data set, but were part of the 3441 compounds that were used to select these 19. These "test" compounds were selected based on their molecular SLATM distance from the training compounds. The molecular SLATM vector consists of the summed aSLATM vectors belonging to a molecule. In order to quantify the similarity of compounds relative to our training compounds, we first construct a matrix of pairwise euclidean distances between molecular SLATM representations of each of the 3441 C_7O_2 isomers. The rows in the pairwise distance matrix corresponding to the 19 molecules in the training set were then summed together and then sorted from smallest to largest values, corresponding to the molecules closest and furthest from the training set compounds. The molecules as well as their Euclidean SLATM distance (scaled such that the maximum distance is 1.0) from the training set is given in Table 4.2.

These test molecules, as well as their mappings are also shown in Fig. 4.10. As done in Fig. 4.8, we average the JSDs of the coarse-grained SP and EE RDFs with respect to the atomistic RDFs for each system and plot the results. Again, for each simulation, the intramolecular interactions remained constant, whereas the intermolecular pairwise interactions came from either the SP or EE model.

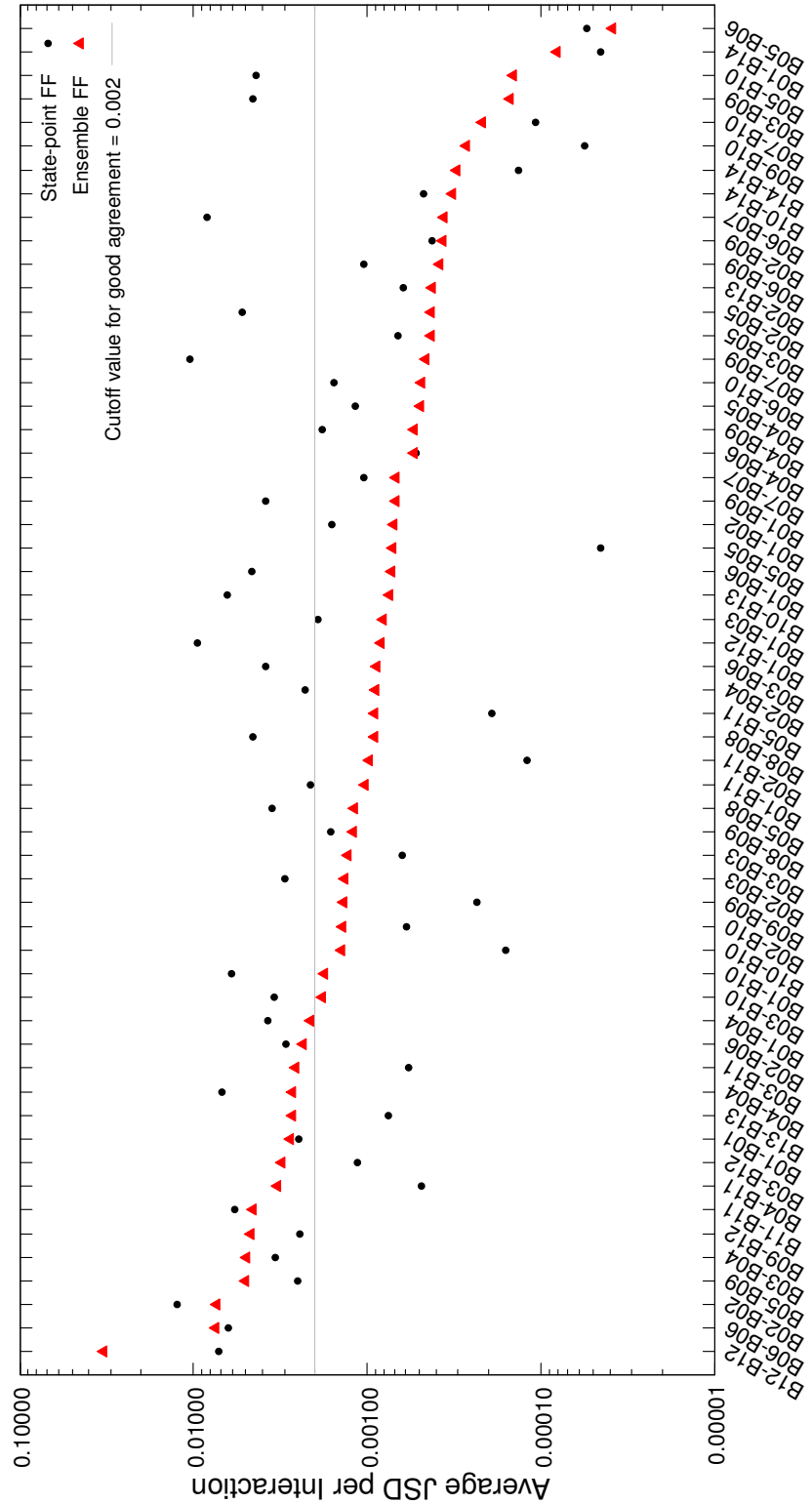

Figure 4.9: Average JSD values for each interaction used when simulating the pure systems shown in Fig. 4.8 using both the state-point specific coarse-grained models (circles) as well as the transferable coarse-grained model obtained by averaging forces over the whole extended ensemble (triangles). The grey line denotes the cutoff JSD value for "good" agreement with atomistic RDFs, 0.002. The interactions are ordered based on increasing agreement of the atomistic RDFs with the extended-ensemble coarse-grained RDFs, with the highest-JSD molecule on the left, and the lowest-JSD molecule at the right.

As expected, the structure of the closest molecule in terms of SLATM distance, Molecule 19, is successfully captured when using the EE model, and even shows improvements in structural accuracy compared to the SP model. Additionally, the EE model outperforms the SP model for Molecule 23 and shows the same level of accuracy as the SP models for Molecule 21, which are both relatively far from the training set in terms of the SLATM distance. On the other hand, the EE model underperforms compared to the SP model for the other "far" compound, Molecule 22. Surprisingly, the structure of Molecule 20 is poorly captured by the EE model despite being relatively close to the training set. Further analysis of Molecule 20 as well as other specific examples from the data shown here is given in the next section. Lastly, we note that performing this analysis on only five molecules is by no means statistically significant, and, as such, these preliminary results may not fully reflect the transferability of the EE model. Additional simulations of other test compounds are currently underway.

4.4 Discussion

Fig. 4.8 shows that both the overall mean and variance of the JSD for the EE model is lower than those of the SP models for each system, and, in some cases, the use of the EE model results in a drastic improvement over the SP model. Naively, one might expect that an extended ensemble force field derived by averaging over the net mean force of several state points would introduce some degree of error when trying to reproduce the structure at any single state point. These results demonstrate that the addition of extra information in the form of additional state points results in a force field that is not only more transferable, but also more structurally accurate than simply using the MSCG method to coarse-grain at a single state point. Furthermore, the reduced variance in the JSD for the EE model when compared to the SP model implies that the EE model provides more reliable expectations as to the quality of the coarse-grained force field and is less likely to produce a highly inaccurate result. However, Fig. 4.8 also reveals several cases for which the SP model greatly outperforms the EE model. In order to better understand why there is greater overall agreement with atomistic results when using the EE model as well as where the EE model fails, we further investigate specific systems for which the EE and SP models yield wildly different JSD values.

First, we consider the pure Molecule 3 system, where Fig. 4.8 shows that Mapping 0 shows the greatest improvement in structural accuracy when switching from the SP model to the EE model. The RDFs used to quantify this accuracy are shown in Fig. 4.11. The reason for this significant improvement is evident when comparing the SP RDFs (red curves) to the EE RDFs (green curves), which closely match the atomistic RDFs (black curves). Just by examining these results,

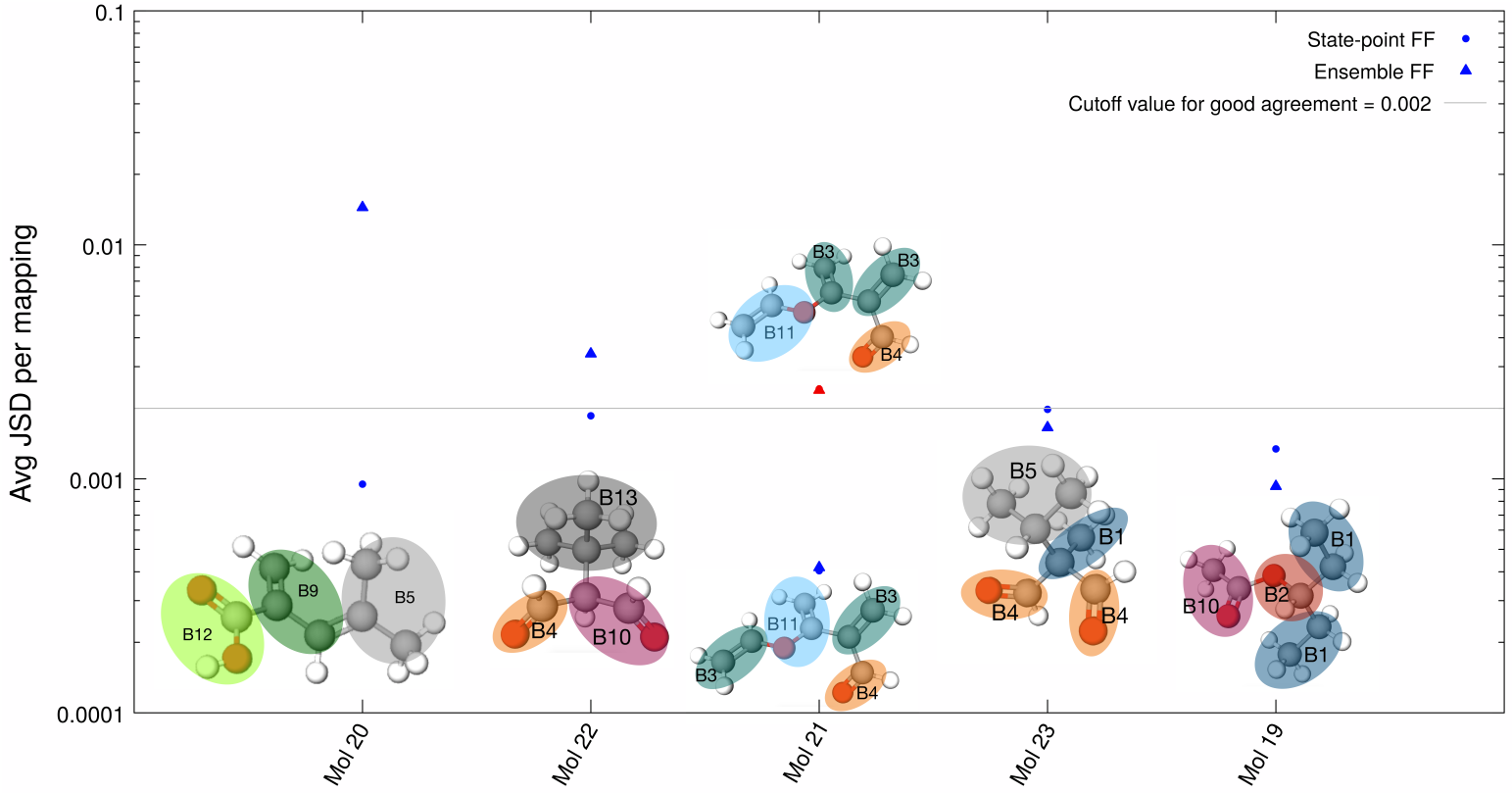

Figure 4.10: Average JSD values for each molecule and mapping, calculated for five test systems consisting of bulk liquid MD simulations using the molecules shown, along with their coarse-grained mappings. Both the state-point specific coarse-grained models (circles) as well as the transferable coarse-grained model obtained by averaging forces over the whole extended ensemble (triangles) are used. Only Molecule 21 has two mappings, with the second mapping shown above the corresponding average JSD, colored red. The grey line denotes the cutoff JSD value for "good" agreement with atomistic RDFs, 0.002. The molecules are ordered based on increasing agreement of the atomistic RDFs with the extended-ensemble coarse-grained RDFs, with the highest-JSD molecule on the left, and the lowest-JSD molecule at the right.

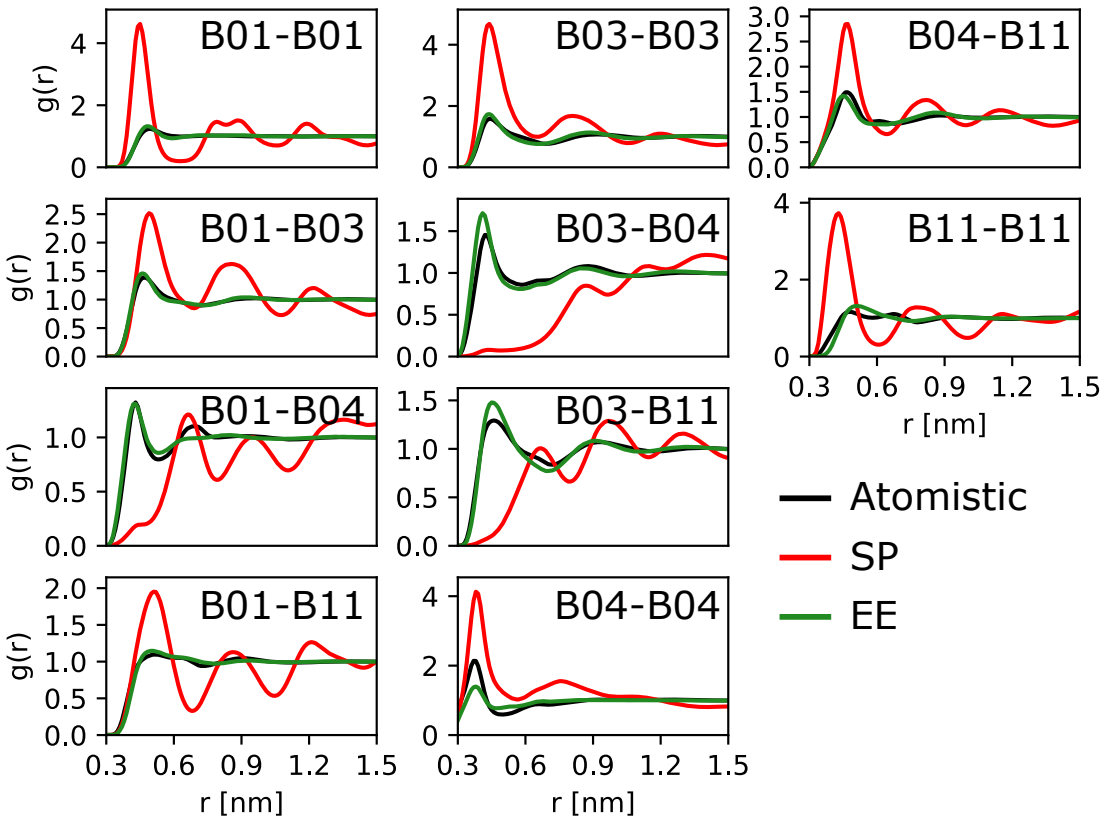

Figure 4.11: All RDFs calculated from atomistic and coarse-grained simulations of the pure Molecule 3, mapping 0 system. The black curves denote the atomistic RDF for the fragments which map to the bead types listed in the top-right of each plot. The RDFs colored red correspond to the state-point specific coarse-grained model, whereas the RDFs colored green correspond to the extended-ensemble model.

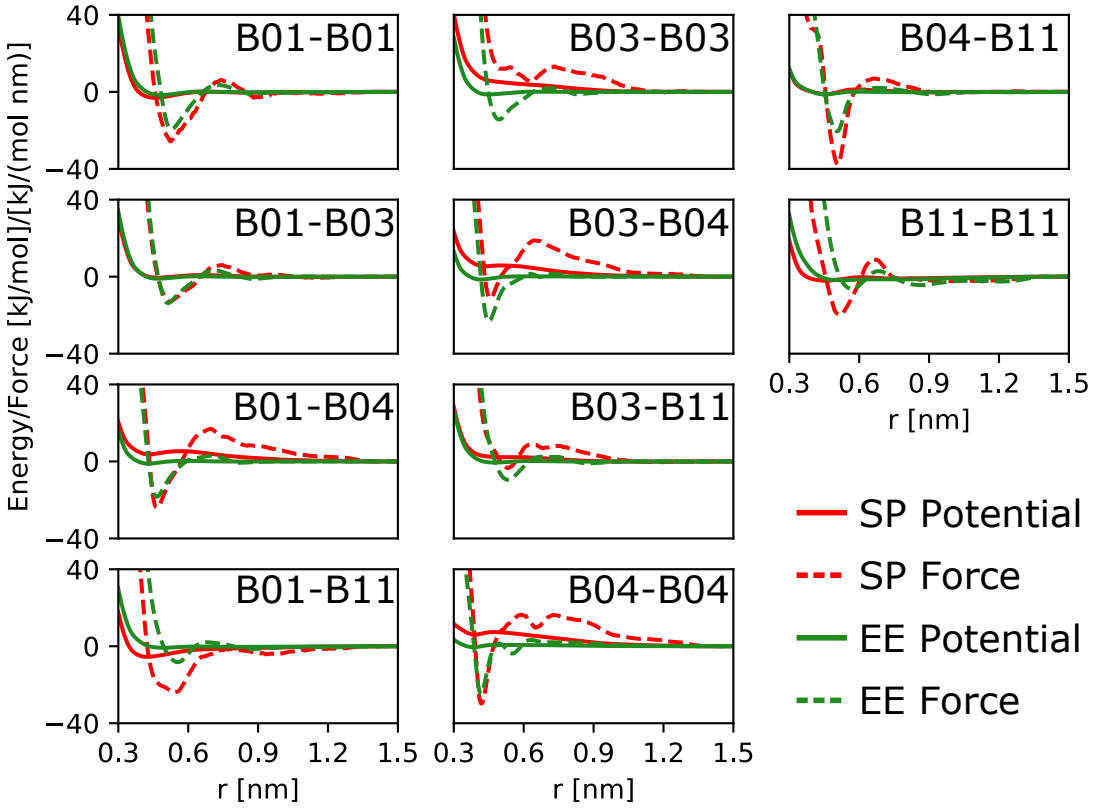

Figure 4.12: Potentials (solid lines) and forces (dashed lines) used in the state-point specific (red) and extended-ensemble (green) coarse-grained simulations of Molecule 3, mapping 0, which resulted in the RDFs shown in Fig. 4.11.

it is difficult to diagnose the SP model to determine which interaction(s) caused the poor agreement. Therefore, in Fig. 4.12, we compare the potentials and the corresponding instantaneous forces that are used when running the coarse-grained MD simulations for both models. This allows us to clearly see how the interactions change when accounting for a single state point versus the entire extended ensemble. In this case, we see that the forces (dashed lines) for the B01-B01 and B01-B03 interactions remain largely unchanged. The remaining interactions, however, seem to share the same qualitative features but show an overall reduction in the magnitude of the repulsive forces. This is consistent with several works which have shown that structure-based CG models tend to feature low cohesive energies in the liquid state, favoring repulsive or weakly attractive potentials [84, 241–243]. This is most evident when looking at any interactions that include a B04 bead type, which corresponds to a carbonyl group. The sole exception to this trend is the B01-B11 interaction, which instead shows a reduction in the magnitude of the attractive forces. Note that the EE potentials look qualitatively more similar to Lennard-Jones potentials than the SP potentials. Since the MSCG method does not explicitly aim to reproduce the structure, but rather aims to reproduce the mean force acting on a given bead type, it is likely that there is a high degree of degeneracy when solving Equation 1.78 for a single system. Essentially, there would be several pairwise potentials that could reproduce the mean force, but only a subset of these that would also result in an accurate structure. Altogether, this suggests that solving Equation 1.78 over the extended-ensemble promotes a regularization-like effect by accounting for correlations across conformational and chemical space for a given interaction. Averaging over these correlations has the net effect of smoothening sharp, localized features in the mean force while preserving the key features which remain across all systems in the ensemble. This smoothening of the mean force also reduces the degeneracy of solutions to Equation 1.78, resulting in a potential that minimizes the force matching functional and is more likely to provide structural accuracy. In order to validate this hypothesis, we are currently comparing sections of the metric tensor from the Molecule 3 state point to their corresponding sections in the metric tensor that is averaged over the whole ensemble. This then allows us to decompose the mean force into direct contributions, which come from the interactions shown in Fig. 4.12, and indirect contributions that result from other correlations found in the environment [230, 231]. Doing so will allow us to identify which specific contributions to the mean force are smoothened by averaging over the extended ensemble.

Next, we examine cases for which the EE model fails to reproduce structure as well as the SP models. Fig. 4.9 shows that that the B12-B12 interaction is significantly worse when compared to the SP model, with an average JSD value of 0.03 versus 0.007, respectively. This interaction is responsible for the performance

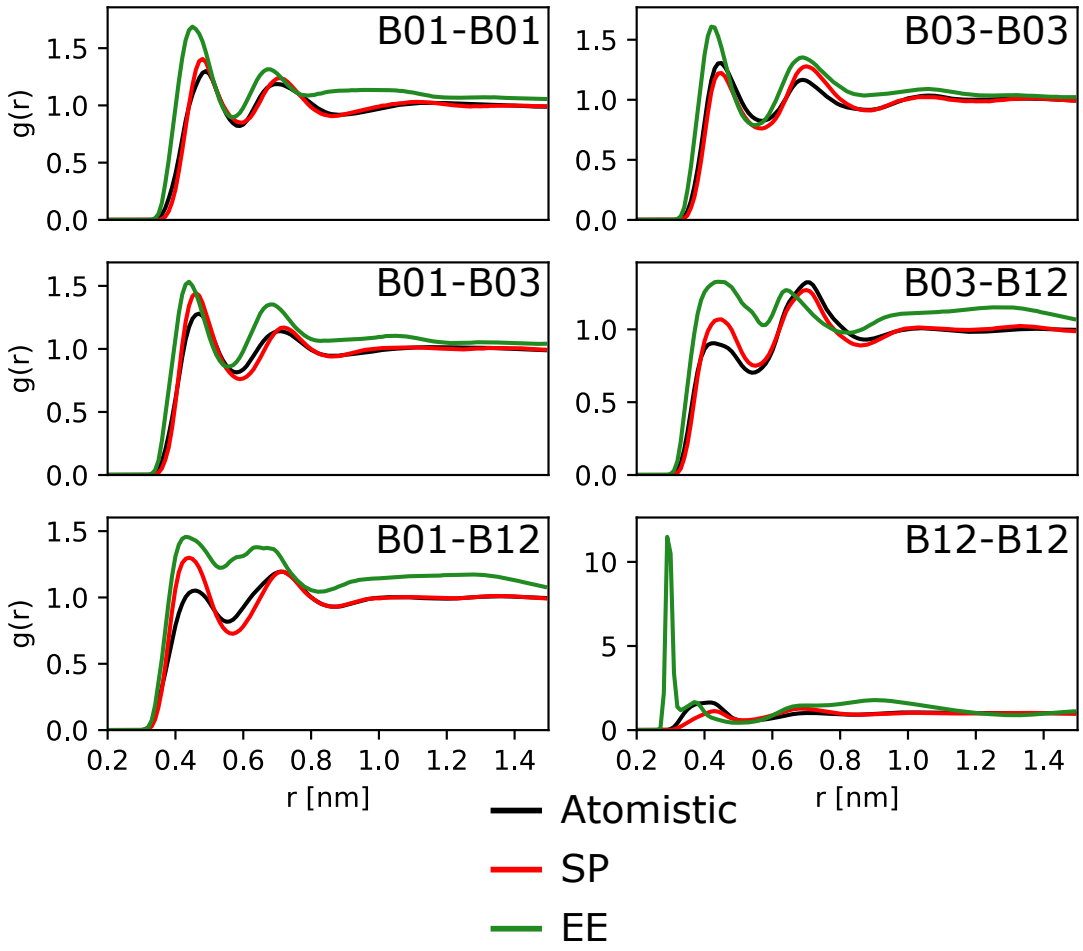

Figure 4.13: All RDFs calculated from atomistic and coarse-grained simulations of the pure Molecule 6, mapping 0 system. The black curves denote the atomistic RDF for the fragments which map to the bead types listed in the top-right of each plot. The RDFs colored red correspond to the state-point specific coarse-grained model, whereas the RDFs colored green correspond to the extended-ensemble model.

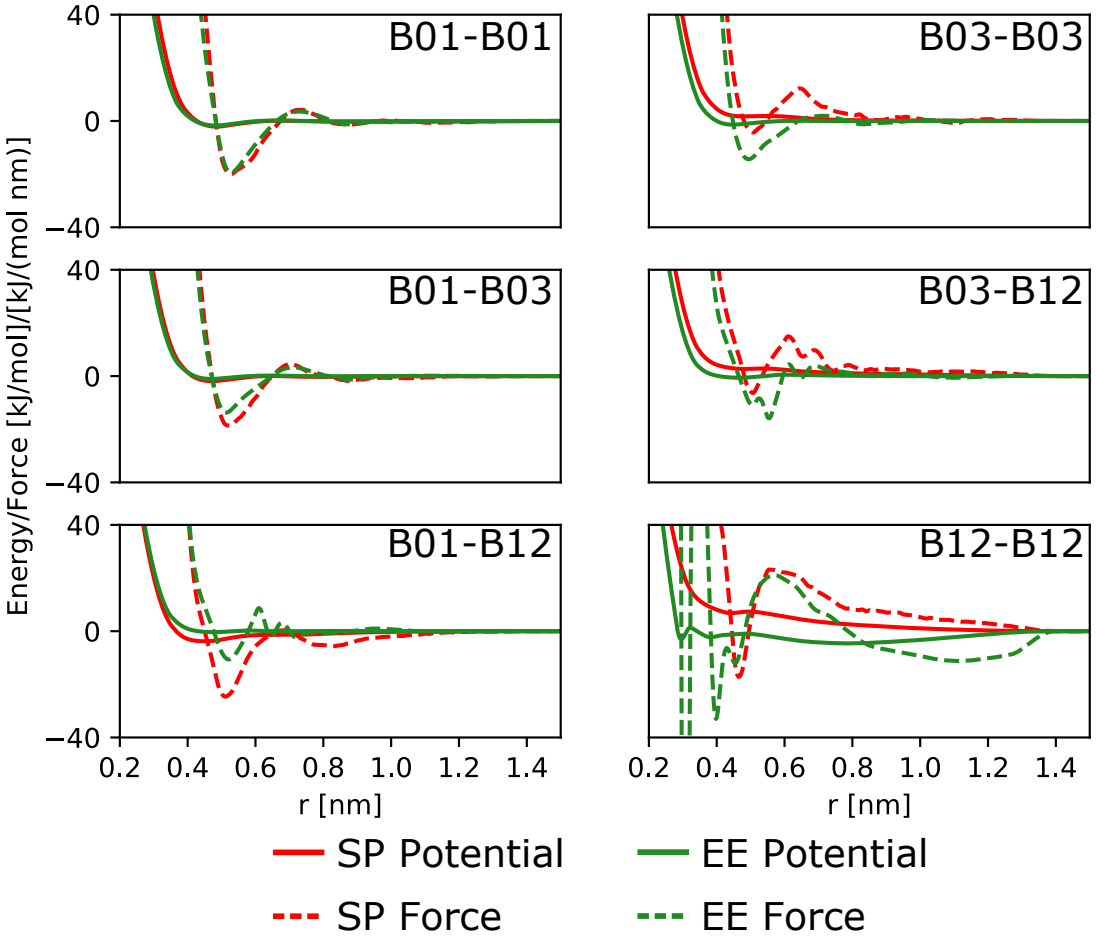

Figure 4.14: Potentials (solid lines) and forces (dashed lines) used in the state-point specific (red) and extended-ensemble (green) coarse-grained simulations of Molecule 6, mapping 0, which resulted in the RDFs shown in Fig. 4.13.

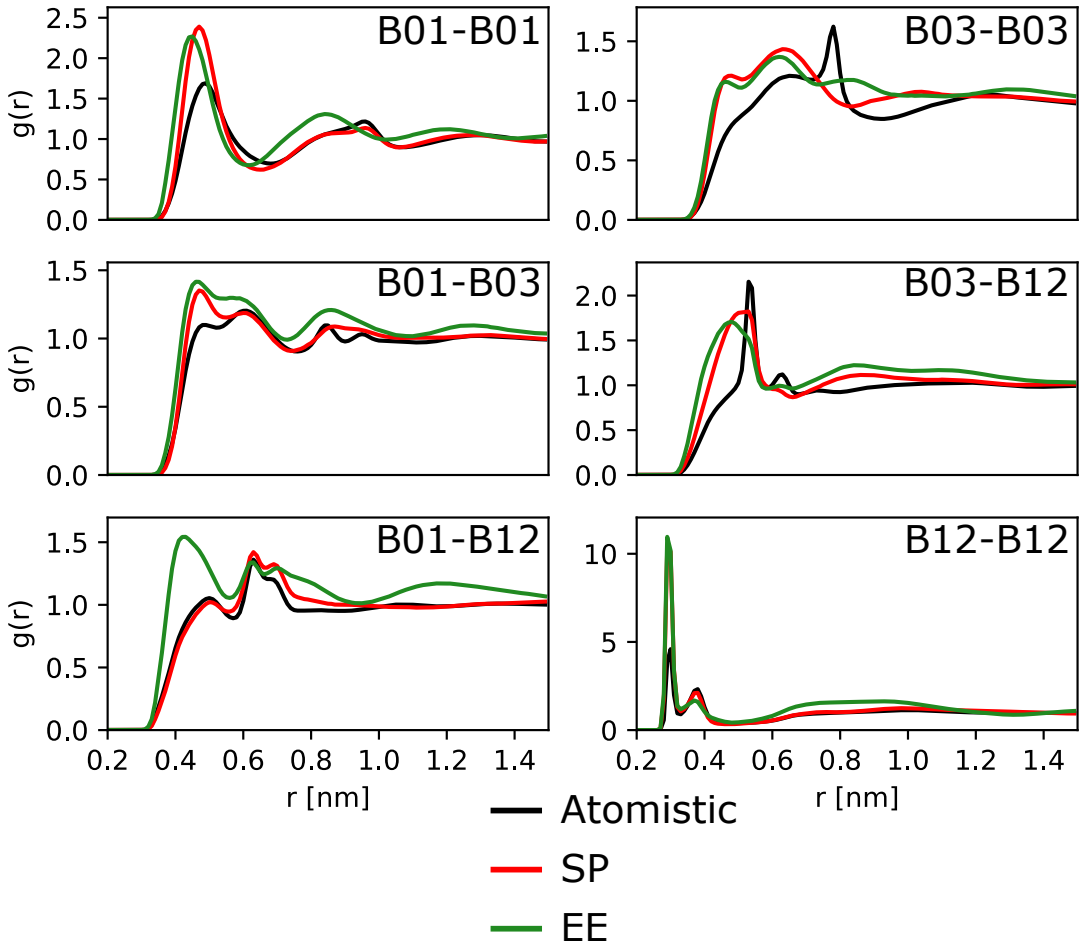

Figure 4.15: All RDFs calculated from atomistic and coarse-grained simulations of the pure Molecule 9 system. The black curves denote the atomistic RDF for the fragments which map to the bead types listed in the top-right of each plot. The RDFs colored red correspond to the state-point specific coarse-grained model, whereas the RDFs colored green correspond to the extended-ensemble model.

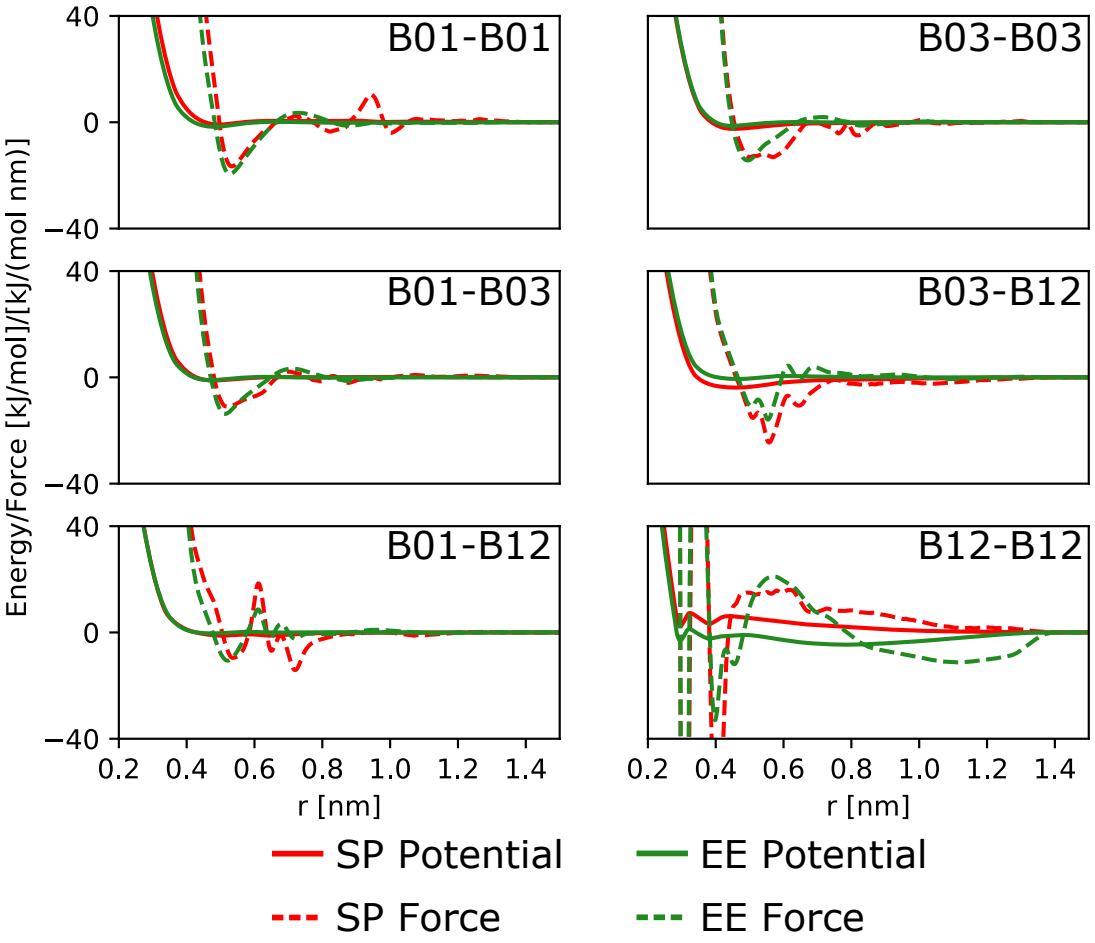

Figure 4.16: Potentials (solid lines) and forces (dashed lines) used in the state-point specific (red) and extended-ensemble (green) coarse-grained simulations of Molecule 9 which resulted in the RDFs shown in Fig. 4.15.

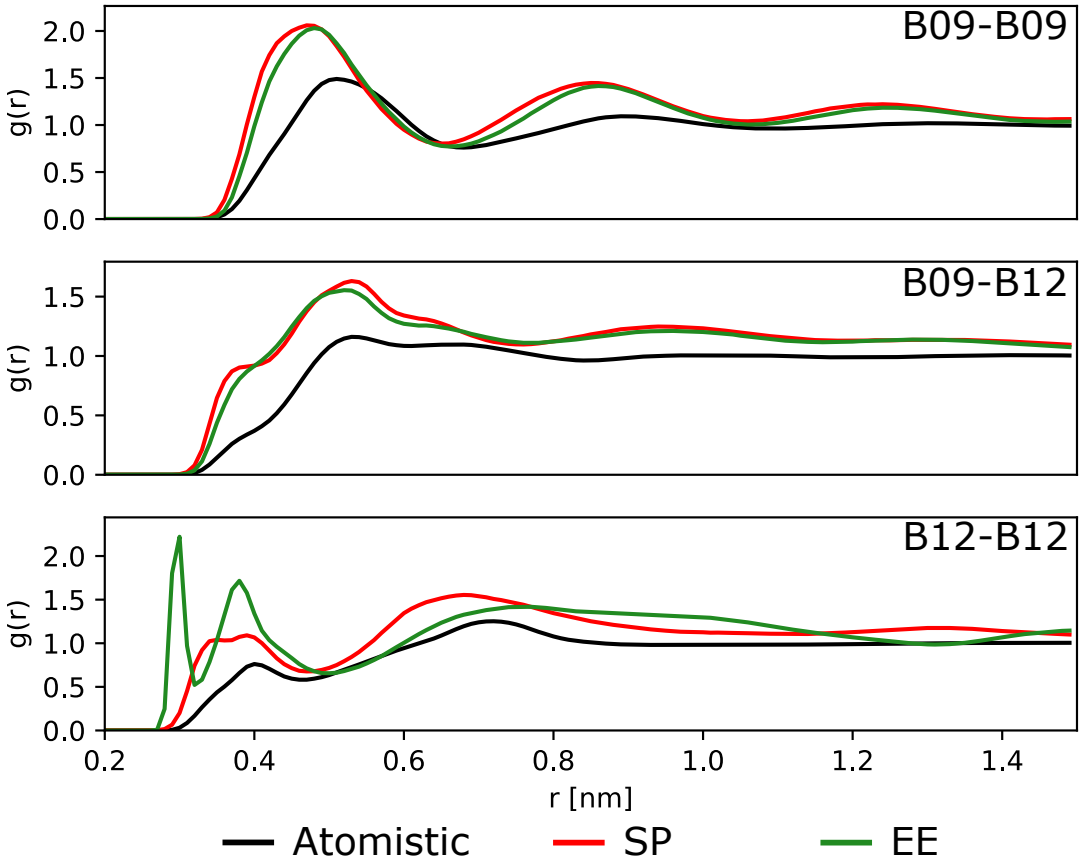

Figure 4.17: All RDFs calculated from atomistic and coarse-grained simulations of the pure Molecule 16 system. The black curves denote the atomistic RDF for the fragments which map to the bead types listed in the top-right of each plot. The RDFs colored red correspond to the state-point specific coarse-grained model, whereas the RDFs colored green correspond to the extended-ensemble model.

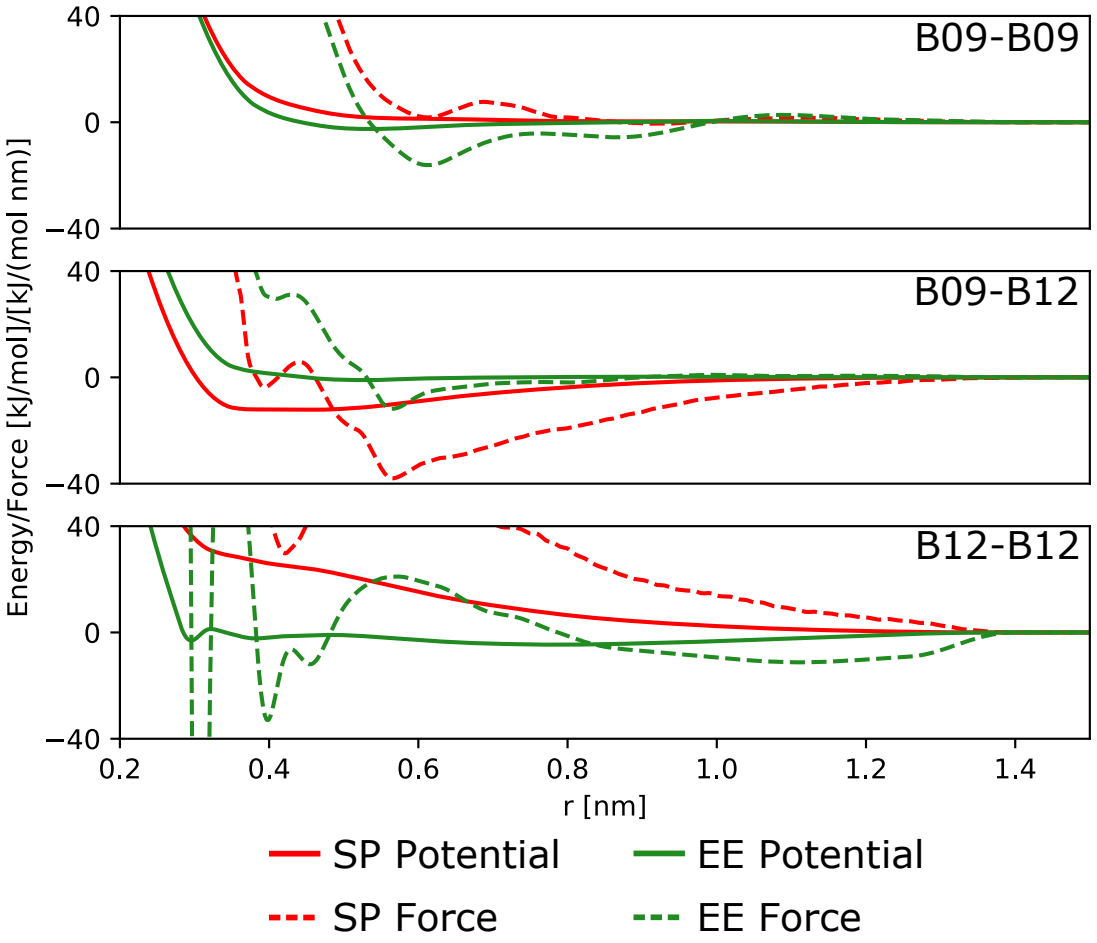

Figure 4.18: Potentials (solid lines) and forces (dashed lines) used in the state-point specific (red) and extended-ensemble (green) coarse-grained simulations of Molecule 16 which resulted in the RDFs shown in Fig. 4.17.

of the EE model for three out of the four worst performing molecules shown in Fig. 4.8: Molecules 6, 9, and 16. As we did for the Molecule 3 system, we again show the RDFs, potentials, and forces for each of these three systems in Figs. 4.13-4.18. It is evident from looking at these figures that the B12-B12 RDFs generated from the EE model all contain a sharp crystalline peaks, which are responsible for the increased JSD values seen in fig. 4.9. Furthermore, the B12-B12 potential contains very sharp kinks at approximately 0.3 nm, resulting in forces that dominate all other interactions in all three systems. Comparing the qualitative features of the B12-B12 potential with those of the SP model for each molecule allows us to identify which features of the B12-B12 potential from each molecule are preserved in the EE model. For example, the aforementioned kinks in the potential are also present in the SP model for Molecule 9, indicating that this feature of the EE B12-B12 interaction was "inherited" from Molecule 9. The next significant feature of the B12-B12 interaction results in the large repulsive forces seen at 0.6 nm, a feature that is qualitatively shared across all three of these systems. Finally, the attractive well at 0.8 nm, which results in large attractive forces at 1.1 nm is not seen in any of the SP models, which suggests that this feature comes from binary mixtures, rather than the pure systems shown here. Overall, it is clear that interactions involving the B12 bead type are obtained by averaging over significantly different configurational ensembles. This is made evident by examining the atomistic RDFs corresponding to Molecule 9 (Fig. 4.15) to those of Molecules 6 and 16 (Fig. 4.13 and Fig. 4.17). Molecule 9 shows liquid crystalline behavior, with sharp peaks seen in the B12-B12 RDF, whereas the other two compounds show bulk-liquid behavior. This is expected due to the chemical structure of Molecule 9 versus Molecules 6 and 16. Molecule 9 consists of alternating single and double bonds, implying that this compound experiences π-stacking interactions, which are known to result in the formation of liquid crystals for organic small molecules [244, 245]. Furthermore, the presence of the terminal carboxylic acid group means that hydrogen bonding is occurring in the bulk liquid phase, which also promotes ordering. On the other hand, the fragments that map to the B12 bead type for both Molecules 6 and 16 are esters that lack hydrogen bonding, and they also do not have any conjugated bonds that would promote π-stacking. Because the local environments for the B12 fragment for each of these cases is so different at the atomistic resolution, averaging over the correlations and the forces from these systems results in a potential that cannot reproduce either case. Both π-stacking and hydrogen-bonding interactions are highly anisotropic in nature, and it is difficult to determine whether both of these interactions can be reproduced using an automated approach using isotropic pairwise potentials without fine tuning, which is at odds with our automated approach [246, 247]. Therefore, an expanded force field basis that accounts for this anisotropy would need to be used when

performing the MSCG calculation over the extended ensemble, or specific modifications/corrections would have to be applied to the existing pairwise potentials in order to accurately model these interactions [248].

Finally, we turn our attention to the test systems used to determine whether or not the EE model is indeed chemically transferable. We chose molecules that were relatively close (Molecules 19 and 20) as well as far (Molecules 21-23) in terms of their molecular SLATM distance from the 19 compounds used to generate the extended ensemble. As shown by Fig. 4.10, molecules 19,21, and 23 all show either the same performance as the SP models or improvements in the structural agreement. On the other hand, the structure of the Molecule 20 and 22 systems is poorly reproduced by the EE model compared to the SP model. For Molecule 20, this disagreement again stems from the poorly modeled carboxylic acid group that maps to the B12 bead type. Indeed, Molecule 20 is quite structurally similar to Molecule 9 from the training set, as both have alternating single and double bonds as well as a terminal carboxylic acid group. The π-stacking interaction in combination with the presence of hydrogen bonding in the Molecule 20 system leads to the formation of liquid crystals (indicated by the sharp peaks in Fig. 4.19 just as was seen in the Molecule 9 RDFs. Despite these similarities, it is unclear as to why the SP model results in poor structural agreement for Molecule 9, whereas the same automated approach yields excellent agreement for Molecule 20. Interestingly, Fig. 4.20 shows that the greatest qualitative difference between the SP model and EE model stems not from the B12 bead type, but rather the B09 bead type, with a large peak appearing in the repulsive forces in the B09-B09 interaction for the SP model that is nonexistent in the EE model. This is also expected, as none of the molecules in the training set both had fragments mapping to a B09 bead and showed liquid crystal behavior in the bulk phase. This further reinforces the idea that chemical fragments that drastically alter the intermolecular behavior of the compound depending on their arrangement within the molecule (i.e., alternating double bonds, carboxylic acids versus esters) should not be mapped to a single bead type, as averaging over drastically different environments results in potentials unsuited for either environment. The fact that the SP models outperform the EE models for both this system and the Molecule 9 system further reinforces this recommendation.

The remaining molecule in the test set, Molecule 22, also poses a challenge for the EE model when compared to the SP model. Out of all the compounds chosen to test the transferability of the EE model, both Molecules 22 and 23 are the two furthest compounds from the training set compounds in terms of their SLATM distance. Both molecules are also quite similar to each other, as both are highly branched and symmetric with respect to the two carbonyl groups present in each compound. Fig. 4.10 shows, however, that while the coarse-grained mapping for

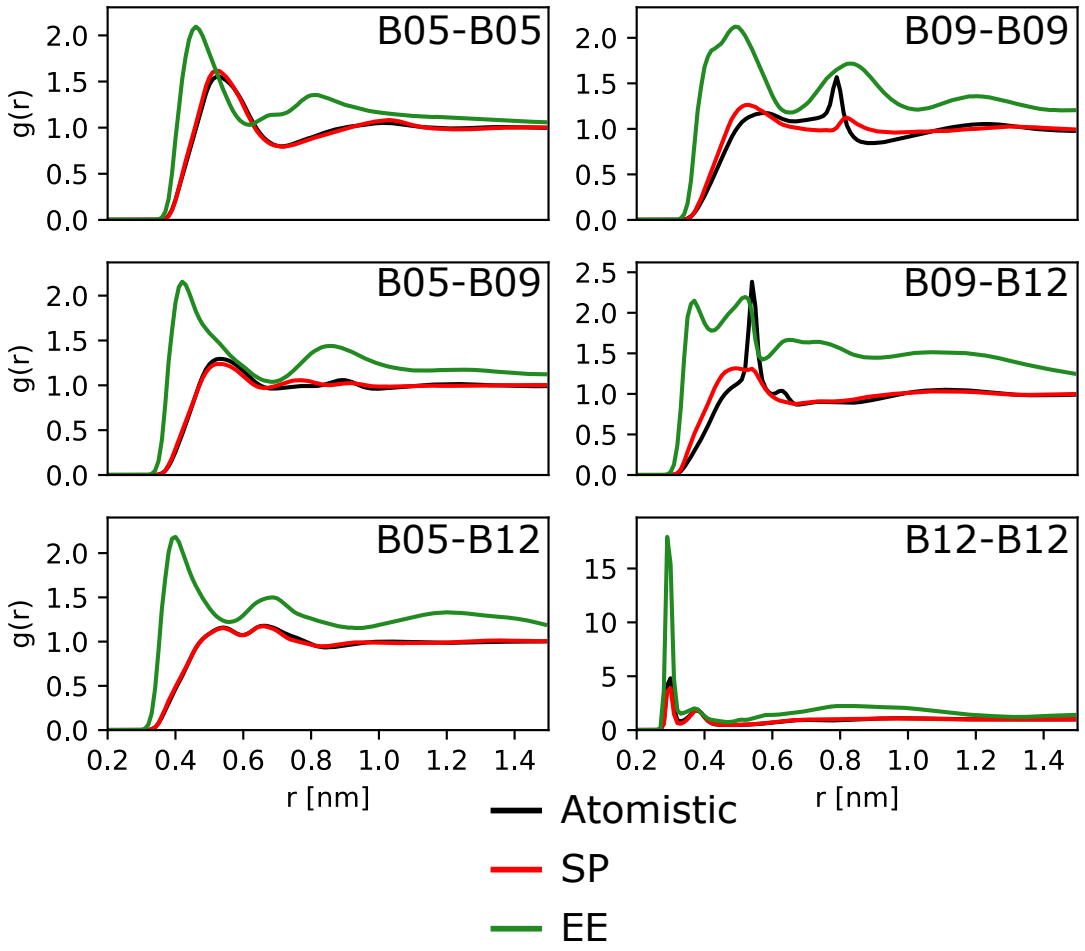

Figure 4.19: All RDFs calculated from atomistic and coarse-grained simulations of the pure Molecule 20 system. The black curves denote the atomistic RDF for the fragments which map to the bead types listed in the top-right of each plot. The RDFs colored red correspond to the state-point specific coarse-grained model, whereas the RDFs colored green correspond to the extended-ensemble model.

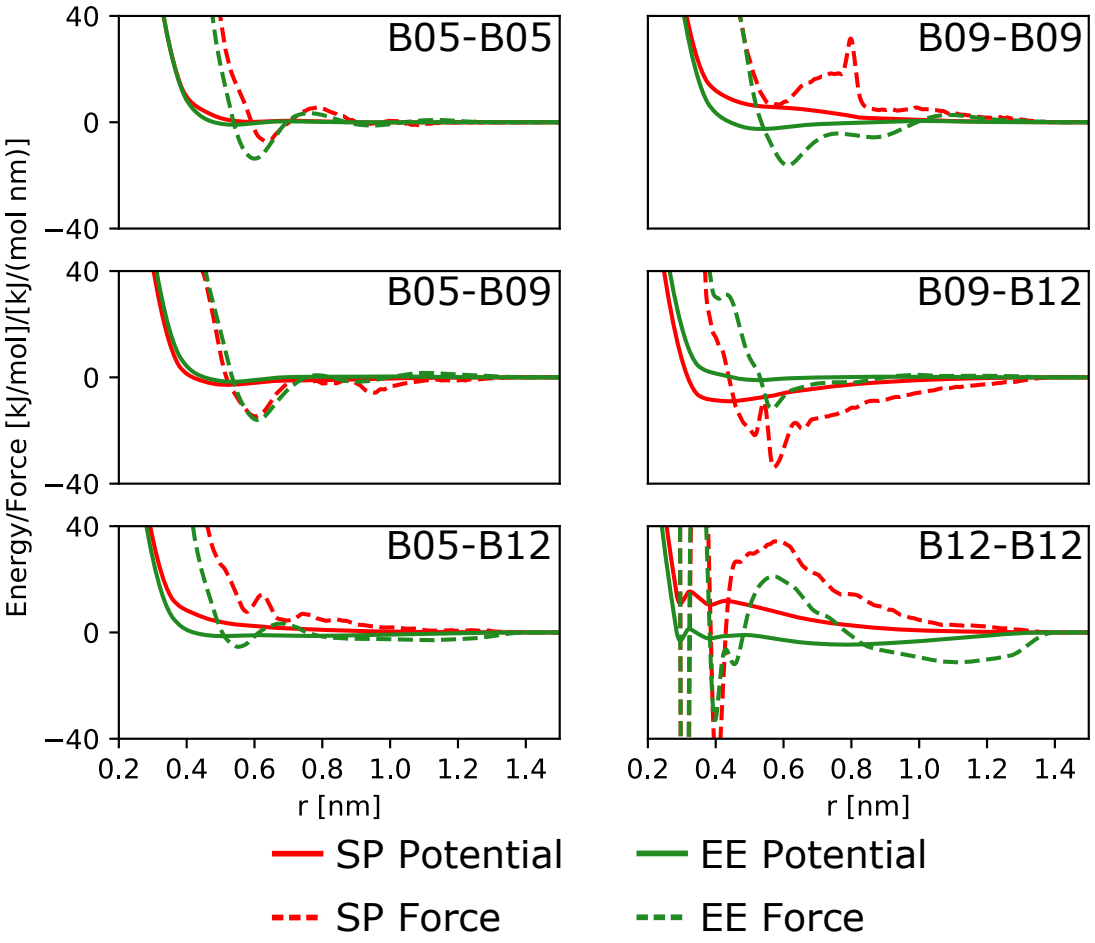

Figure 4.20: Potentials (solid lines) and forces (dashed lines) used in the state-point specific (red) and extended-ensemble (green) coarse-grained simulations of Molecule 20 which resulted in the RDFs shown in Fig. 4.19.

Molecule 23 is also symmetric, the mapping for Molecule 22 is asymmetric. The carbonyl groups in Molecule 23 are unevenly split into two-heavy-atom and three-heavy-atom fragments that are mapped to a B4 and a B10 bead type, respectively. The preservation of symmetry when choosing a coarse-grained mapping has also been shown by Chakraborty et al. to have a negligible effect on the structural accuracy of the coarse-grained model [249]. Indeed, when only using the Molecule 22 state point to calculate the potential, this asymmetry is irrelevant, as the SP model has an average JSD below the cutoff value for good agreement (Fig. 4.10). However, since none of the compounds in the training set have a symmetric atomistic structure that maps to an asymmetric coarse-grained representation, we hypothesize that these correlations, which would be present in the metric tensor used to calculate the SP model, are absent in the corresponding metric tensor for the EE model. Essentially, this means that the resulting EE potentials are likely to be less accurate for symmetric compounds with asymmetric coarse-grained mappings, since these types of symmetric atomistic correlations were not included in the training set. On the other hand, because the coarse-grained mapping of Molecule 23 is symmetric, the potentials are also symmetric, and the liquid-phase structure is accurately reproduced. For both the Molecule 22 system as well as the previously discussed examples (Molecules 6, 9, 16, and 20) in which the EE model is outperformed by the SP model, we are currently performing a mean force decomposition analysis so as to explicitly prove or disprove these hypotheses.

All in all, these initial results suggest that the EE model is indeed transferable and can be used to model most of the remaining compounds in our input database with structural accuracy without having to calculate new potentials for each new system. However, we did not account for specific intermolecular interactions (i.e., hydrogen bonding and π-stacking) that play a significant role in determining the structure of these systems in the bulk phase. The fact that an anomalous compound was introduced into the extended ensemble speaks to the strength of the clustering analysis and its ability to properly identify representative compounds from the gas-phase trajectories alone. While this clustering approach was successful in choosing representative compounds that would maximize the transferability of the resulting coarse-grained potentials, the mapping scheme used was unable to account for the emergent behaviors that occur in bulk liquid phases due to these intermolecular interactions. One strategy to overcome this issue could be to perform another clustering step after choosing the representative molecules from the gas-phase clustering results using the liquid-phase trajectories. This would reveal the extent to which the local environments for compounds containing the same fragments varies, and could be used to predict the number of bead types or other modifications to the force-field basis set that should be added in order to account for these differences. For example, adding a "B12-H" bead type for compounds

that have a carboxylic acid group rather than an ester could have mitigated the errors caused by using the B12 bead type in the EE model, although the intrinsically anisotropic nature of these interactions will limit the extent to which adding additional bead types improves the structural accuracy of the coarse-grained model. Investigating and implementing these possible improvements to the EE model will be the subject of a future work.

4.5 Conclusions and Future Work

In this work, we present a new workflow for obtaining chemically-transferable coarse-grained models that preserve the liquid-phase structure of organic small molecules. This method couples unsupervised learning methods with a traditional coarse-graining approach. We initially ran gas-phase MD simulations of all of the 3441 compounds in our input data set. Using the aSLATM molecular representation, we then represented each compound as the set of conformationally-averaged, unique local environments obtained from its gas-phase trajectory. we then used HDBSCAN, a graph-based clustering technique, to identify the clusters of local environments that were shared across all compounds in the input data set. The clusters were organized according to multiple hierarchies of increasing resolution, which corresponded to the many-body types encoded in the aSLATM representations. Furthermore, it was clear that the clusters were differentiated based on the types of functional groups that were shared across compounds, and we therefore used these functional groups in our coarse-grained mapping scheme. We then identified nineteen compounds whose local environments were found in the greatest number of clusters. We hypothesized that these representative compounds were therefore the most likely to share conformational similarities with the remaining compounds in the data set. We then performed extensive bulk liquid-phase simulations of these compounds as well as all possible binary mixtures between these compounds. These simulations formed the extended ensemble which was used to calculate the transferable coarse-grained potentials. Specifically, we applied the extended-ensemble MSCG method, using these liquid-phase trajectories as input, to obtain the transferable potentials. We note that, as far as we know, there has been no bottom-up coarse-graining study performed on such a large extended-ensemble up to this point. We then validated our results by running coarse-grained simulations of the pure (i.e., single-component) liquid systems using both the transferable coarse-grained potentials as well as coarse-grained potentials derived using only the atomistic trajectory of that specific thermodynamic state point, and comparing the RDFs using both models to the atomistic reference RDFs. Astonishingly, our transferable model outperformed the state-point specific models on average, even though the extended ensemble used to parameterize our

model mainly consisted of binary mixtures. We also tested the transferability of the model by simulating five test compounds and found that our model performed as well, if not better than, the state-point specific models for three out of five of these test compounds. By examining a specific system for which the transferable model showed significantly better structural agreement in greater detail, we have formed a plausible hypothesis as to why the transferable model yields better results. By averaging across the extended ensemble, certain sharp features in the mean force—which would otherwise dominate the coarse-grained potentials resulting from a single state point—are smoothened, while key features that persist across multiple state points are preserved. This results in a regularization-like effect that restricts the space of force fields that can optimize the force-matching functional to those that are more likely to also correctly reproduce the atomistic structure. However, we also found that certain chemical fragments that have significantly different interactions depending on where they are placed on a carbon scaffold. Specifically, neither the hydrogen-bonding behavior of carboxylic acid groups nor the delocalized π-orbital behavior that results from conjugated small molecules were accounted for. Because systems that included these interactions were grouped with systems that did not include these interactions, averaging the correlations and the mean forces for these bead types resulted in potentials that could not accurately capture either case. Furthermore, we hypothesized that compounds that are symmetric require a symmetric coarse-grained mapping in order for our transferable model to be fully applicable. We believe these two anomalies explain why our model was unable to reproduce the structures for the two remaining test compounds. We are currently investigating these cases in more detail so as to prove these hypotheses.

While we were able to demonstrate that our approach was successful, several questions remain. Aside from the unique cases for which we know our model fails (systems with hydrogen bonding, conjugation, symmetry), we should expect that each of the 3441 compounds can be accurately modeled, and we are currently running simulations of other compounds so as to make our results more statistically meaningful. We intend to test other compounds that consist only of carbon and oxygen heavy atoms but are outside the size restriction of nine heavy atoms total. Furthermore, the mapping scheme used in this work was not obtained in a rigorous manner; rather, we used the clustering of the aSLATM vectors to justify an encompassing approach in which all possible 2-heavy-atom and 3-heavy atom fragments were assigned a bead type, as well as two branched bead types which were added so as to ensure that every molecule in the database could be fully mapped. Recently, several metrics for evaluating the quality of a coarse-grained mapping have been proposed [81, 249, 250]. We are testing the viability of these metrics as a reference to tune the parameters of the aSLATM vector (sigma val-

ues, grid spacing, etc.) such that our approach yields clusters that correspond to coarse-grained mappings that preserve the most information from the higher resolution. In a similar vein, we are also remaking the transferable model using fewer state points to see if we can find the information threshold that must be reached in order to guarantee a certain level of transferability and accuracy. Additionally, we have always used direct Boltzmann inversion on the bulk liquid-phase trajectories to obtain intramolecular coarse-grained potentials. This is an unfeasible approach if we wish to implement this transferable model in a high-throughput screening scheme, similar to the methods used in Chapter 2 of this work. We are therefore testing the accuracy of our model when using intramolecular potentials obtained by performing direct Boltzmann inversion on the gas-phase trajectories. This also motivates the creation of a supervised machine-learning model that can predict coarse-grained intramolecular potentials in the bulk-liquid phase given an input molecular structure and its coarse-grained gas-phase intramolecular potentials. Another avenue that we are currently exploring is how to efficiently reduce the number of bead types while maintaining the same level of accuracy. Just as certain chemical fragments would require an expanded force field basis or an additional bead type in order to account for different interaction types, it is possible that certain chemistries (for example the aliphatic fragments) could all be treated as a single interaction type, which could greatly increase the screening efficiency of the model. In addition to the ongoing work mentioned above, we are also applying a pressure-matching method (see Chapter 1, Section 1.1.5) to ensure that our model correctly reproduces the atomistic pressure, and thus allow it to reproduce other thermodynamic properties. This will allow us to compare our coarse-grained model to the Martini model, which is known for its ability to accurately model thermodynamic partitioning.

Similar to UMAP, HDBSCAN is an unsupervised learning method that transforms the data in the high-dimensional space by modeling it as a graph and uses the nearest neighbors to re-weight the edges of the graph so as to identify stable clusters. In both cases, the re-weighting is not applied globally, but is unique to each point in the data set, and is dependant on the nearest neighbor distances for each point. Here, we demonstrate that this unsupervised learning approach provides a viable route to applying bottom-up coarse-graining to chemical compound space. Having investigated both top-down and bottom-up approaches to coarse-graining CCS, we summarize our results in the next and final chapter. We discuss the impact of the work and the questions raised due to our findings, which should be explored further.

5 Conclusions and Future Outlook

In this chapter, we summarize the content of each previous chapter, highlighting the importance of the results presented and the conclusions drawn. We begin with an overview, which is essentially a restatement of the introductory remarks and motivation described in Chapter 1. We then follow this with summaries of each of the subsequent chapters (which are modified versions of the conclusion sections of each chapter), and conclude with an outlook that highlights the questions raised by this work and discusses new avenues of research that could answer these questions.

5.1 Overview

Chemical structure–property relationships are essential for the development of new materials used in all facets of life. As implied by their name, these relationships connect chemical structures with properties of interest, which can include protein-ligand binding, atomization energies, macroscopic phase transitions, and countless others [9, 19, 251–253]. Specifically, a chemical structure–property relationship is constructed by first projecting the space of all chemical compounds, called chemical compound space (CCS), onto some number of descriptors that are related to the property of interest [1–3]. The hypersurface upon which the CCS lies after projection is the structure–property relationship. A good structure–property relationship not only provides a holistic and intuitive sense for the physical phenomena that give rise to properties of interest, but also enables quantitative predictions for new compounds. The challenge in constructing these hypersurfaces usually stems from a lack of data, as their accuracy and transferability will depend on how well-sampled CCS is with respect to the chosen descriptors. Therefore, a commonly used approach to developing chemical structure–property relationships is high-throughput screening, in which the properties of compounds are determined in an automated fashion. Both experimental and computational high-throughput schemes have been developed, each with their own advantages and disadvantages in addition to their cost in time and resources [11–13, 16–18]. This is compounded with the exorbitantly large size of CCS, estimated to be about 10^{60} for drug-like small molecules [4].

Recently, a number of methods that rely on machine-learning and other data-driven techniques have been developed to infer structure–property relationships for

the electronic properties of various organic and inorganic chemistries [15, 118, 121, 197, 200]. The high-throughput step for these schemes requires ab initio calculations, in which the electronic probability distribution is obtained by numerically solving Schrödinger's equation, to be run in vacuo for each compound screened. On the other hand, relatively few high-throughput computational methods have been proposed that build structure–property relationships for thermodynamic properties in the condensed phase, for which thermal fluctuations play an important role [24, 25, 254, 255]. For these methods, the corresponding technique used for screening is usually classical molecular dynamics (MD) simulations, which also present the main bottleneck towards their implementation in a high-throughput screening protocol due to high computational costs.

Coarse-grained (CG) models provide a means to circumvent these costs [31–34]. A CG model represents chemical compounds as particles in a similar fashion to all-atomistic (AA) MD simulations. However, each particle in a CG model represents groups of atoms rather than a single atom. The model can be constructed by projecting information from a high-resolution simulation (e.g. AA), by inferring microscopic behavior using macroscopic experimental results, or some combination of both. In all cases, the goal of the CG model is to reproduce certain properties of interest by projecting the information pertaining to the property onto a minimal set of parameters. This approach has interesting parallels with the construction of chemical structure–property relationships, as both involve relating chemical structure to a desired property using a reduced model. Furthermore, CG MD simulations require fewer particles compared to AA (usually by some multiplicative factor between two and ten). They also do not require as much sampling time due to the removal of unnecessary information that is irrelevant (i.e., has no correlation with) to the studied property. These two effects can significantly reduce the number of CPU hours required [31, 34, 39]. Therefore, CG modeling may provide a means to accelerate the computational high-throughput screening process for condensed phase thermodynamic properties.

The difficulty in utilizing CG models for high-throughput screening stems from the fact that many CG models are chemically specific, meaning that they are constructed for a single chemical compound or small set of chemical compounds, usually at a single thermodynamic state point [36–38]. Because they require an AA MD simulation or other, equally expensive experimental data for their parameterization, the transferability of these models is usually limited to the chemistry used in their construction. This means that for each compound, the high-resolution data would have to first be obtained, making the actual construction of the CG model unnecessary for a high-throughput approach. Several instances of extending the transferability of CG models have been demonstrated, but these are applied to the state point variables, allowing for CG models to be run at various temper-

atures, pressures, and concentrations given the same set of chemical compounds [38, 40, 41, 256]. On the other hand, relatively little work has been done that investigates the chemical transferability of CG models [42, 243, 257]. A chemically-transferable CG model would be highly beneficial in a high-throughput screening process because a single CG molecule would be representative of many different chemical compounds, thereby reducing the total number of simulations necessary to construct a chemical structure–property relationship.

The central theme of this work is to investigate the different ways in which CG modelling can be used to augment the computational high-throughput screening of CCS for condensed phase thermodynamic properties. We have shown that chemically-transferable CG models reduce the size of CCS and can be used to quickly construct broadly encompassing chemical structure–property relationships [60, 110, 138]. We further investigated how unsupervised machine learning (i.e., clustering and dimensionality reduction) allows us to coarse-grain CCS in both a top-down and bottom-up manner, and demonstrated approaches for parameterizing CG models that maximize their chemical transferability in both cases [227]. While further investigations are required, we are encouraged by our results and believe that the methods highlighted here mark a fundamental first step towards a new paradigm in efficiently constructing chemical-structure property relationships.

5.2 The High-Throughput Coarse-Grained Simulation Method

In Chapter 2, we first introduced the high-throughput coarse-grained (HTCG) approach as a means to quickly construct structure–property relationships that span CCS. By applying this method using the top-down Martini force field[79, 86, 88], we were able to identify linear relationships between key thermodynamic state points when modeling the behavior of small molecules in a lipid bilayer membrane environment. A single, easily-accessible parameter, $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ was the only required input in order to predict the transfer free energies between these state points [60]. We extended this structure–property relationship by introducing a second descriptor, the acidity, onto which we could then project the coarse-grained permeabilities for all Martini unimers and dimers [138]. We then demonstrated that further exploration of coarse-grained CCS, corresponding to Martini trimers and tetramers, was possible by implementing a Monte-Carlo scheme that used alchemical transformations to construct and optimize thermodynamic cycles that efficiently sampled the CG compound space [110]. A Kernel Ridge Regression (KRR) model was then trained on these results to further expand the transferability of these structure property relationships. In implementing the HTCG ap-

proach, we also demonstrated a drastic reduction of CCS when coarse-graining using Martini due to the degeneracy of molecules that were mapped to the same Martini representation. Approximately 1.8 million molecules from the Generated Database (GDB) were mapped to Martini unimers, dimers, and trimers, using the AUTO-MARTINI algorithm [90, 140, 161]. By performing a functional-group analysis on these compounds, we were able to provide a means to implement inverse molecular design when targeting a specific membrane permeability [138]. As far as we are aware, we are the first to apply a coarse-grained model in a high-throughput scheme that takes advantage of the chemical transferability of that model (which is traditionally seen as a negative attribute of a coarse-grained model) in order to dramatically increase the screening efficiency.

Next, we assessed three different molecular representations as well as three different dimensionality reduction techniques in order to determine whether unsupervised ML could provide a means for further screening in a hierarchical manner. We found that principal component analysis (PCA) and SKETCH-MAP preserve the global structure of the high-dimensional data while the UMAP visualizations consisted of well-separated clusters which were randomly placed in relation to each other [99, 104, 106]. Additionally, the SLATM and ASOAP representations were able to relate chemical structure to $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, whereas the Coulomb Matrix did not show a strong correlation to this property [118–120]. This insight led to the parameterization of a KRR model using the SLATM vector that could compete with the ALOGPS program, although it remains unclear as to why a single input configuration was sufficient to achieve such high accuracy when predicting a thermodynamic property [163, 175]. We showed that even relatively low-dimensional representations, like the modified Coulomb Matrix, could identify molecular scaffolds which could be used in a hierarchical screening approach. We also demonstrated that the clusters obtained from this unsupervised ML approach do not always correspond to specific Martini bead types, motivating our work in the next chapter. Importantly, we made parallels between different unsupervised learning techniques and different approaches to coarse-graining. PCA and SKETCH-MAP apply global transformations to high-dimensional data that is used to encode CCS in a similar fashion to how top-down coarse-graining methods globally map CCS to reproduce certain experimental data (in the Martini case this is $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$). On the other hand, the localized means by which UMAP and HDBSCAN transform the data is more reminiscent of bottom-up methods, which tend to be more chemically specific, and therefore localized in CCS. These (non-rigorous) analogies suggested strategies that would be useful for maximizing the chemical transferability of either top-down or bottom-up CG models, and each of these strategies was subsequently explored in the next two chapters.

5.3 Resolution limit of data-driven top-down coarse-grained models spanning chemical space

In Chapter 3, we used the Jensen-Shannon divergence (JSD) to quantify the information loss in chemically-transferable top-down coarse-grained models with varying numbers of bead types, with the GDB as our proxy for CCS [115]. We found that Martini, while not designed to efficiently reduce CCS, performed remarkably well in this regard, closely matching the other force fields explicitly designed to minimize the JSD with only a small deviation [227]. All force fields yielded roughly the same level of accuracy with respect to $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$, but varied greatly in their coverage of CCS. We used a Bayesian approach to calculate the probabilities of back-mapping given bead-types to fragments containing specific chemical substitutions. Here, we found it necessary to constrain the size of chemical fragments to five heavy atoms and the presence of two functional groups in order to clearly differentiate between the chemical moieties mapping to each bead type. The results of this Bayesian analysis indicated that increasing the number of bead types decreased the range of accessible chemistry while increasing the corresponding posterior probabilities for each chemistry. However, there was a resolution limit when using this approach, as it did not take into account the specific positions of hetero-atom and bond substitutions within a fragment, causing different bead types to appear representative of the same chemistry. Overall, we saw that Martini, as well as other chemically-transferable coarse-grained models, can be used to quickly build structure–property relationships that span broad regions of CCS. Here we highlighted the powerful combination of this method with Bayesian inference, providing an informed mapping of a coarse structure–property relationship to a higher resolution in chemical compound space and further enabling inverse molecular design.

This work also reinforced the conclusions of the previous chapter regarding the top-down approach to coarse-graining CCS. In the previous chapter, global unsupervised learning methods were applied to a data base of fragments that were mapped to Martini dimers. The results indicated that $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ correlated well with the number and type of functional groups found on a carbon scaffold. We therefore tested the extent to which the HTCG approach could be optimized by developing models that covered the $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ axis at varying resolutions. We had further noted that most of the apolar compounds were grouped into two clusters only, which was at odds with the total number of apolar bead types used in Martini (C1 through C5). Indeed, we saw that the number of apolar bead types could be reduced to two while maintaining the overall accuracy of the model for apolar compounds, as was done for the five-bead-type model in this work. At the same time, the non-polar and polar bead types mapped to a much wider range

of compounds, which also correlated with the results from the previous chapter, in which the unsupervised learning results indicated that these chemistries were prevalent in far more than three clusters.

It was also evident from this work that only accounting for the correlation between the number/type of heavy atom substitutions and $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ would be insufficient to easily identify specific functional groups for inverse molecular design without also drastically increasing the number of bead types. Even with sixteen neutral bead types, the number of functional-group pairs with significant backmapping probability was greater than twenty-five. This resolution limit stems from deliberately ignoring the structural information encoded in the low-dimensional maps in the previous chapter and determining how much chemical specificity could be preserved when only using this 1-D approach. While the relative entropy has been previously utilized for optimizing the quality of bottom-up CG models[258], this is the first study to use this metric to both optimize and quantitatively assess the chemical transferability of top-down CG models.

5.4 Bottom-Up Chemically-Transferable Coarse-Grained Models that Preserve Structure

In Chapter 4, we present a new workflow for obtaining chemically-transferable coarse-grained models that preserve the liquid-phase structure of organic small molecules. This method couples unsupervised learning techniques with a traditional coarse-graining approach. We initially ran gas-phase MD simulations of all of the 3441 compounds in our input data set. Using the aSLATM molecular representation, we then represented each compound as the set of conformationally-averaged, unique local environments obtained from its gas-phase trajectory. we then used HDBSCAN, a graph-based clustering technique, to identify the clusters of local environments that were shared across all compounds in the input data set [98]. The clusters were organized according to multiple hierarchies of increasing resolution, which corresponded to the many-body types encoded in the aSLATM representations. Furthermore, it was clear that the clusters were differentiated based on the types of functional groups that were shared across compounds, which invited using these functional groups in our coarse-grained mapping scheme. We then identified nineteen compounds whose local environments were found in the greatest number of clusters. We hypothesized that these representative compounds were therefore the most likely to share conformational similarities with the remaining compounds in the data set. We then performed extensive bulk liquid-phase simulations of these compounds as well as all possible binary mixtures between these compounds. These simulations formed the extended ensemble which was

used to calculate the transferable coarse-grained potentials. Specifically, we applied the extended-ensemble Multi-Scale Coarse Graining (MSCG) method, using these liquid-phase trajectories as input, to obtain the transferable potentials [40, 77]. We note that, as far as we know, there has been no bottom-up coarse-graining study performed on such a large extended-ensemble up to this point. We then validated our results by running coarse-grained simulations of the pure (i.e., single-component) liquid systems using both the transferable coarse-grained potentials as well as coarse-grained potentials derived using only the atomistic trajectory of that specific thermodynamic state point, and comparing the radial distribution functions (RDFs) using both models to the atomistic reference RDFs. Astonishingly, our transferable model outperformed the state-point specific models on average, even though the extended ensemble used to parameterize our model mainly consisted of binary mixtures. We also tested the transferability of the model by simulating five test compounds and found that our model performed as well, if not better than, the state-point specific models for three out of five of these test compounds. By examining a specific system for which the transferable model showed significantly better structural agreement in greater detail, we have formed a tentative hypothesis as to why the transferable model yields better results. By averaging across the extended ensemble, certain sharp features in the mean force—which would otherwise dominate the coarse-grained potentials resulting from a single state point—are smoothened, while key features that persist across multiple state points are preserved. This results in a regularization-like effect that restricts the space of force fields that can optimize the force-matching functional to those that are more likely to also correctly reproduce the atomistic structure. However, we also found that certain chemical fragments that have significantly different interactions depending on where they are placed on a carbon scaffold. Specifically, neither the hydrogen-bonding behavior of carboxylic acid groups nor the delocalized π-orbital behavior that results from conjugated small molecules were accounted for. Because systems that included these interactions were grouped with systems that did not include these interactions, averaging the correlations and the mean forces for these bead types resulted in potentials that could not accurately capture either case. Furthermore, we hypothesized that compounds that are symmetric require a symmetric coarse-grained mapping in order for our transferable model to be fully applicable. We believe these two anomalies explain why our model was unable to reproduce the structures for the two remaining test compounds. We are currently investigating these cases in more detail so as to prove these hypotheses, as well as running additional test systems so as to statistically validate our model.

Both UMAP and HDBSCAN are unsupervised learning methods that represent high-dimensional data as a graph and re-weight the edges of the graph so as to

either construct a low dimensional representation or to easily identify stable clusters. In both cases, the re-weighting is not applied globally, but is unique to each point in the data set, and is dependant on the nearest neighbor distances for each point. Here, we demonstrate that this unsupervised learning approach provides a viable route to developing bottom-up coarse-grained models with chemical transferability. As mentioned above, the transferability stems from two factors: the choice of representative molecules obtained through unsupervised learning methods and the smoothening of the force field basis correlations and mean forces that results from the extended ensemble approach. While other studies that examine chemical transferability of bottom-up coarse-grained models have been recently published, none of these studies have tested the limitations of their respective approaches [243, 257]. Furthermore, the number of compounds used in our study (3441) greatly exceeds the number studied in other works, and we expect to use the increased statistics provided by our large data set size to further validate our approach and its limitations.

5.5 Outlook

Future work pertaining to the coarse-graining of chemical compound space and investigating the chemical transferability of coarse-grained models will proceed along two main avenues of research. The first of these research goals is to obtain further understanding of the results reported here. For example, data-mining techniques are currently being applied to the permeability database that we created using the HTCG approach in order to further generalize this structure–property relationship. We are also working on better understanding the SLATM KRR model used to predict $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ by tuning the parameters of the model (Gaussian vs. Laplacian kernel, choice of Euclidean vs. Manhattan norm) as well as investigating the effect of increasing the number of configurations when training the model. Further comparisons with ALOGPS, in which we measure the accuracy of each when predicting $\Delta G_{\mathrm{W}\rightarrow\mathrm{Ol}}$ for larger and more-varied databases, are also underway. In addition to the ongoing work described in Chapter 4, it would be interesting to perform a sensitivity analysis with respect to the amount of "training" data used in the extended-ensemble MSCG method. This would allow us to better understand which of the 703 atomistic simulations were crucial in ensuring the transferability of the resulting coarse-grained model, allowing us to reduce the computational cost when building these types of models in the future. While we did not rigorously connect our clustering results to our coarse-grained mapping scheme, a recent study has linked the optimization of a coarse-grained mapping to spectral clustering techniques, which is the foundation for the HDBSCAN algorithm [250]. As this and other studies have been conducted only on single molecules or on small groups of

chemically similar compounds, it would be interesting to see how these approaches change when optimizing mappings across several different compounds [81, 259]. These approaches may also serve as a benchmark for modifying the parameters of the aSLATM vector (sigma values, grid spacing, cutoff) such that the resulting clusters could be directly linked to a coarse-grained mapping scheme.

The other avenue of research deals with extending the methods proposed here and applying them to new systems. For example, further work is currently underway to apply the HTCG approach in order to construct structure–property relationships for other target properties as well as screening for specific compounds or chemical moieties. One of these projects is using the five-bead coarse-grained model reported in Chapter 3 and extended to be compatible with the refined polarizable Martini force field [222]. The results from Chapter 4 also point to several studies that can extend the transferability and screening efficiency of the bottom-up coarse-grained model. For example, we have always used Direct Boltzmann Inversion on the bulk liquid-phase trajectories to obtain intramolecular coarse-grained potentials. This is an unfeasible approach if we wish to implement this transferable model in a high-throughput screening scheme. We are therefore testing the accuracy of our model when using intramolecular potentials obtained by performing Direct Boltzmann Inversion on the gas-phase trajectories. This also motivates the creation of a supervised machine-learning model that can predict coarse-grained intramolecular potentials in the bulk-liquid phase given an input molecular structure and its coarse-grained gas-phase intramolecular potentials. Successfully being able to translate gas-phase trajectories into liquid-phase intramolecular potentials would allow us to construct an efficient, generalized mapping algorithm like AUTO-MARTINI, but with our structurally-accurate model. As was done with Martini in Chapter 3, it would also be useful to determine how the accuracy of the bottom-up transferable model changes as a function of the number of bead types. For example, it is possible that the certain chemistries (for example the aliphatic fragments) could all be treated as a single interaction type, which could greatly increase the screening efficiency of the model, with an HTCG implementation in mind. We are also applying a pressure-matching method to ensure that our model also has thermodynamic consistency with the atomistic references. This will allow us to compare our coarse-grained model to the Martini model, and further increase the range of problems to which our model can be applied. Additionally, the large error due to the emergence of π-stacking or hydrogen bonding in our training set can potentially be mitigated by introducing additional bead types or by implementing a force-field surface-hopping scheme recently proposed by Rudzinski and Bereau, in which a liquid-crystal coarse-grained force field can be "hopped" onto when certain conditions are met [243]. Future avenues of research may involve testing the transferability of this model for molecules larger or smaller

than the nine-heavy-atom compounds used here, or expanding the training set to include other chemistries. In conclusion, we firmly believe that this first examination of the chemical transferability of coarse-grained models and their ability to reduce chemical compound space will eventually lead to significant advances in computational high-throughput screening and the discovery and design of new, better-performing materials.

Contributions

The first half of Chapter 2, as well as the entirety of Chapter 3 have been previously published as articles in peer-reviewed scientific journals. Chapter 4 is currently in preparation to be submitted as an article. All of the work specified in these publications was carried out in the Max Planck Institute for Polymer Research in Mainz. We now state the individual contributions for each Chapter in detail.

Chapter 2:
Roberto Menichetti, Kiran H. Kanekal, Kurt Kremer, and Tristan Bereau
In silico screening of drug-membrane thermodynamics reveals linear relations between bulk partitioning and the potential of mean force
The Journal of Chemical Physics 147(12):125101, 2017.
DOI: 10.1063/1.4987012

The original idea was developed by Tristan Bereau, Roberto Menichetti, and Kurt Kremer. The simulation setups were conceived by Tristan Bereau and Roberto Menichetti. Roberto Menichetti ran and analyzed all simulations. Kiran Kanekal ran the AUTO-MARTINI algorithm to construct the transfer free energy databases and quantified the reduction of chemical space. The paper was written by Roberto Menichetti, Kiran Kanekal, and Tristan Bereau, incorporating critical comments from Kurt Kremer.

Roberto Menichetti, Kiran H. Kanekal, and Tristan Bereau
Drug–membrane permeability across chemical space
ACS Central Science 5(2):290, 2019.
DOI: 10.1021/acscentsci.8b00718

The original idea was developed by Roberto Menichetti, Kiran Kanekal, and Tristan Bereau. The simulation setups were conceived by Roberto Menichetti and Tristan Bereau. The simulations were run and analyzed by Roberto Menichetti. The database construction and the functional group analysis was done by Kiran Kanekal. The paper was written by Roberto Menichetti, Kiran Kanekal, and Tristan Bereau.

Christian Hoffmann, Roberto Menichetti, Kiran H. Kanekal, and Tristan Bereau
Controlled exploration of chemical space by machine learning of coarse-grained representations
Physical Review E 100(3):033302, 2019.
DOI: 10.1103/PhysRevE.100.033302

The original idea was conceived by Roberto Menichetti, Kiran Kanekal, and Tristan Bereau. The simulation setups were conceived by Roberto Menichetti and Tristan Bereau. The simulations were run by Christian Hoffmann and Roberto Menichetti. The kernel ridge regression models were constructed and optimized by Christian Hoffmann. The database construction was done by Kiran Kanekal.

The ideas for the remaining machine learning and hierarchical screening sections were conceived by Kiran Kanekal and Tristan Bereau, and implemented by Kiran Kanekal.

Chapter 3:
Kiran H. Kanekal and Tristan Bereau
Resolution limit of data-driven coarse-grained models spanning chemical space
The Journal of Chemical Physics 151:164106, 2019.
DOI: 10.1063/1.5119101

The original idea was developed by Kiran Kanekal and Tristan Bereau. The implementation and data analysis were carried out by Kiran Kanekal. The paper was written by Kiran Kanekal, with critical commentary from Tristan Bereau.

Chapter 4:
Kiran H. Kanekal, Joseph Rudzinski, and Tristan Bereau
Bottom-Up Chemically-Transferable Coarse-Grained Models that Preserve Structure
In Preparation.

The original idea was developed by Kiran Kanekal, Joseph Rudzinski, and Tristan Bereau. The simulation setups and coarse-graining protocols were developed by Kiran Kanekal and Joseph Rudzinski. The implementation and data analysis were carried out by Kiran Kanekal. The paper was written by Kiran Kanekal, with critical commentary from Joseph Rudzinski and Tristan Bereau.

Bibliography

[1] Mati Karelson, Victor S Lobanov, and Alan R Katritzky. Quantum-chemical descriptors in QSAR/QSPR studies. *Chemical reviews*, 96(3):1027–1044, 1996.

[2] Tristan Bereau, Denis Andrienko, and Kurt Kremer. Research Update: Computational materials discovery in soft matter. *APL Materials*, 4(5):053101, 2016.

[3] Kunal Roy, Supratik Kar, and Rudra Narayan Das. *A primer on QSAR/QSPR modeling: fundamental concepts*. Springer, 2015.

[4] Christopher M Dobson. Chemical space and biology. *Nature*, 432(7019):824–828, 2004.

[5] Christoph Kuhn and David N Beratan. Inverse strategies for molecular design. *The Journal of Physical Chemistry*, 100(25):10595–10599, 1996.

[6] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[7] David Rogers and Anton J Hopfinger. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *Journal of Chemical Information and Computer Sciences*, 34(4):854–866, 1994.

[8] O Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. Exploring chemical compound space with quantum-based machine learning. *Nature Reviews Chemistry*, pages 1–12, 2020.

[9] Jon Paul Janet and Heather J Kulik. Resolving transition metal chemical space: Feature selection for machine learning and structure–property relationships. *The Journal of Physical Chemistry A*, 121(46):8939–8954, 2017.

[10] Christopher Lipinski and Andrew Hopkins. Navigating chemical space for biology and medicine. *Nature*, 432(7019):855–861, 2004.

[11] Lingjun Zhang, Yuriy Fedorov, Drew Adams, and Feng Lin. Identification of complement inhibitory activities of two chemotherapeutic agents using a high-throughput cell imaging-based screening assay. *Molecular immunology*, 101:86–91, 2018.

[12] Michael Shevlin, Max R Friedfeld, Huaming Sheng, Nicholas A Pierson, Jordan M Hoyt, Louis-Charles Campeau, and Paul J Chirik. Nickel-catalyzed asymmetric alkene hydrogenation of α, β-unsaturated esters: high-throughput experimentation-enabled reaction discovery, optimization, and mechanistic elucidation. *Journal of the American Chemical Society*, 138(10):3562–3569, 2016.

[13] Matthias Wambach, Robin Stern, Sandip Bhattacharya, Pawel Ziolkowski, Eckhard Müller, Georg KH Madsen, and Alfred Ludwig. Unraveling self-doping effects in thermoelectric TiNiSn half-Heusler compounds by combined theory and high-throughput experiments. *Advanced Electronic Materials*, 2(2):1500208, 2016.

[14] Jianzhuang Yao, Xia Wang, Haixia Luo, and Pengfei Gu. Understanding the Catalytic Mechanism and the Nature of the Transition State of an Attractive Drug-Target Enzyme (Shikimate Kinase) by Quantum Mechanical/Molecular Mechanical (QM/MM) Studies. *Chemistry–A European Journal*, 23(64):16380–16387, 2017.

[15] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[16] O Anatole Von Lilienfeld. Quantum machine learning in chemical compound space. *Angewandte Chemie International Edition*, 57(16):4164–4169, 2018.

[17] Albert P Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R Kermode, Gábor Csányi, and Michele Ceriotti. Machine learning unifies the modeling of materials and molecules. *Science advances*, 3(12):e1701816, 2017.

[18] Mie Andersen, Sergey V Levchenko, Matthias Scheffler, and Karsten Reuter. Beyond scaling relations for the description of catalytic materials. *ACS Catalysis*, 9(4):2752–2759, 2019.

[19] Felix A Faber, Anders S Christensen, Bing Huang, and O Anatole Von Lilienfeld. Alchemical and structural distribution based representation for universal quantum machine learning. *The Journal of chemical physics*, 148(24):241717, 2018.

[20] Yuping He, Ekin D Cubuk, Mark D Allendorf, and Evan J Reed. Metallic metal–organic frameworks predicted by the combination of machine learning methods and ab initio calculations. *The journal of physical chemistry letters*, 9(16):4562–4569, 2018.

[21] Sabine Körbel, Miguel AL Marques, and Silvana Botti. Stability and electronic properties of new inorganic perovskites from high-throughput ab initio calculations. *Journal of Materials Chemistry C*, 4(15):3157–3167, 2016.

[22] Thomas A Witten. Insights from soft condensed matter. *Reviews of Modern Physics*, 71(2):S367, 1999.

[23] Christine Peter and Kurt Kremer. Multiscale simulation of soft matter systems. *Faraday discussions*, 144:9–24, 2010.

[24] Kai Yang, Xinyi Xu, Benjamin Yang, Brian Cook, Herbert Ramos, NM Anoop Krishnan, Morten M Smedskjaer, Christian Hoover, and Mathieu Bauchy. predicting the Young's Modulus of silicate Glasses using High-throughput Molecular Dynamics simulations and Machine Learning. *Scientific reports*, 9(1):1–11, 2019.

[25] Min Xu, Andrea Unzue, Jing Dong, Dimitrios Spiliotopoulos, Cristina Nevado, and Amedeo Caflisch. Discovery of CREBBP bromodomain inhibitors by high-throughput docking and hit optimization guided by molecular dynamics. *Journal of medicinal chemistry*, 59(4):1340–1349, 2016.

[26] Timothy S Carpenter, Daniel A Kirshner, Edmond Y Lau, Sergio E Wong, Jerome P Nilmeier, and Felice C Lightstone. A method to predict blood-brain barrier permeability of drug-like compounds using molecular dynamics simulations. *Biophysical journal*, 107(3):630–641, 2014.

[27] G Giupponi, MJ Harvey, and G De Fabritiis. The impact of accelerator processors for high-throughput molecular modeling and simulation. *Drug discovery today*, 13(23-24):1052–1058, 2008.

[28] Jiří Průša and Michal Cifra. Dependence of amino-acid dielectric relaxation on solute-water interaction: Molecular dynamics study. *Journal of Molecular Liquids*, 303:112613, 2020.

[29] Ravindra W Tejwani, Malcolm E Davis, Bradley D Anderson, and Terry R Stouch. Functional group dependence of solute partitioning to various locations within a DOPC bilayer: a comparison of molecular dynamics simulations with experiment. *Journal of pharmaceutical sciences*, 100(6):2136–2146, 2011.

[30] Chi Hoon Park, Tae-Hyun Kim, Deuk Ju Kim, and Sang Yong Nam. Molecular dynamics simulation of the functional group effect in hydrocarbon anionic exchange membranes. *International Journal of Hydrogen Energy*, 42(32):20895–20903, 2017.

[31] William George Noid. Perspective: Coarse-grained models for biomolecular systems. *The Journal of chemical physics*, 139(9):09B201_1, 2013.

[32] Christine Peter and Kurt Kremer. Multiscale simulation of soft matter systems–from the atomistic to the coarse-grained level and back. *Soft Matter*, 5(22):4357–4366, 2009.

[33] Jörg Baschnagel, Kurt Binder, Pemra Doruker, Andrei A Gusev, Oliver Hahn, Kurt Kremer, Wayne L Mattice, Florian Müller-Plathe, Michael Murat, Wolfgang Paul, et al. Bridging the gap between atomistic and coarse-grained models of polymers: Status and perspectives. In *Viscoelasticity, atomistic models, statistical chemistry*, pages 41–156. Springer, 2000.

[34] Gregory A Voth. *Coarse-graining of condensed phase and biomolecular systems*. CRC press: Boca Raton, FL, 2008.

[35] Joseph F Rudzinski. Recent progress towards chemically-specific coarse-grained simulation models with consistent dynamical properties. *Computation*, 7(3):42, 2019.

[36] Tsuyoshi Terakawa and Shoji Takada. Multiscale ensemble modeling of intrinsically disordered proteins: p53 N-terminal domain. *Biophysical journal*, 101(6):1450–1458, 2011.

[37] Ali Ghavami, Erik van der Giessen, and Patrick R Onck. Coarse-grained potentials for local interactions in unfolded proteins. *Journal of Chemical Theory and Computation*, 9(1):432–440, 2013.

[38] Pritam Ganguly and Nico FA van der Vegt. Representability and Transferability of Kirkwood–Buff Iterative Boltzmann Inversion Models for Multicomponent Aqueous Systems. *Journal of chemical theory and computation*, 9(12):5247–5256, 2013.

[39] Emiliano Brini, Elena A Algaer, Pritam Ganguly, Chunli Li, Francisco Rodriguez-Ropero, and Nico FA van der Vegt. Systematic coarse-graining methods for soft matter simulations–a review. *Soft Matter*, 9(7):2108–2119, 2013.

[40] JW Mullinax and William George Noid. Extended ensemble approach for deriving transferable coarse-grained potentials. *The Journal of Chemical Physics*, 131(10):104110, 2009.

[41] Tanmoy Sanyal and M Scott Shell. Transferable Coarse-Grained Models of Liquid–Liquid Equilibrium Using Local Density Potentials Optimized with the Relative Entropy. *The Journal of Physical Chemistry B*, 122(21):5678–5693, 2018.

[42] Tanmoy Sanyal, Jeetain Mittal, and M Scott Shell. A hybrid, bottom-up, structurally accurate, G o^{-}-like coarse-grained protein model. *The Journal of chemical physics*, 151(4):044111, 2019.

[43] Yaxin An, Karteek K Bejagam, and Sanket A Deshmukh. Development of New transferable coarse-grained models of hydrocarbons. *The Journal of Physical Chemistry B*, 122(28):7143–7153, 2018.

[44] M Scott Shell. *Thermodynamics and statistical mechanics: an integrated approach*. Cambridge University Press, 2015.

[45] M Scott Shell. Lecture notes in Principles of modern molecular simulation methods, September 2019.

[46] Daan Frenkel and Berend Smit. Understanding Molecular Simulation. Computational Science Series. *Academic Press, San Diego Adcock SA, McCammon JA (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev*, 106:1589–1615, 2002.

[47] Michele M Miller, Stanley P Wasik, Guo Lan Huang, Wan Ying Shiu, and Donald Mackay. Relationships between octanol-water partition coefficient and aqueous solubility. *Environmental science & technology*, 19(6):522–529, 1985.

[48] Jack De Bruijn, Frans Busser, Willem Seinen, and Joop Hermens. Determination of octanol/water partition coefficients for hydrophobic organic chemicals with the "slow-stirring" method. *Environmental Toxicology and Chemistry: An International Journal*, 8(6):499–512, 1989.

[49] Ana Sánchez-Iglesias, Marek Grzelczak, Thomas Altantzis, Bart Goris, Jorge Perez-Juste, Sara Bals, Gustaaf Van Tendeloo, Stephen H Donaldson Jr, Bradley F Chmelka, Jacob N Israelachvili, et al. Hydrophobic interactions modulate self-assembly of nanoparticles. *ACS nano*, 6(12):11059–11065, 2012.

[50] Rohan Patil, Suranjana Das, Ashley Stanley, Lumbani Yadav, Akulapalli Sudhakar, and Ashok K Varma. Optimized hydrophobic interactions and hydrogen bonding at the target-ligand interface leads the pathways of drug-designing. *PloS one*, 5(8), 2010.

[51] Wataru Shinoda. Permeability across lipid membranes. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1858(10):2254–2265, 2016.

[52] Timon Idema. Membranes and vesicles. In Gerhard Gompper, Jan Dhont, Jens Elgeti, Christoph Fahlke, Dmitry Fedosov, Stephan Forster, Pavlik Lettinga, and Andreas Offenhausser, editors, *Physics of Life*, chapter C1. Forschungszentrum Juelich GmbH, 52425 Juelich, 2018.

[53] Jacob N Israelachvili. *Intermolecular and surface forces*. Academic press, 2011.

[54] Roberto Menichetti, Kiran H Kanekal, and Tristan Bereau. Drug–membrane permeability across chemical space. *ACS central science*, 5(2):290–298, 2019.

[55] Andrea Kopp Lugli, Charles Spencer Yost, and Christoph H Kindler. Anaesthetic mechanisms: update on the challenge of unravelling the mystery of anaesthesia. *European journal of anaesthesiology*, 26(10):807, 2009.

[56] Siewert-Jan Marrink and Herman JC Berendsen. Simulation of water transport through a lipid membrane. *The Journal of Physical Chemistry*, 98(15):4155–4168, 1994.

[57] Bernard Faller. Artificial membrane assays to assess permeability. *Current drug metabolism*, 9(9):886–892, 2008.

[58] J M Reis, B Sinko, and C HR Serra. Parallel artificial membrane permeability assay (PAMPA)-Is it better than Caco-2 for human passive permeability prediction? *Mini reviews in medicinal chemistry*, 10(11):1071–1076, 2010.

[59] Justin L MacCallum, WF Drew Bennett, and D Peter Tieleman. Distribution of amino acids in a lipid bilayer from computer simulations. *Biophysical journal*, 94(9):3393–3404, 2008.

[60] Roberto Menichetti, Kiran H Kanekal, Kurt Kremer, and Tristan Bereau. In silico screening of drug-membrane thermodynamics reveals linear relations between bulk partitioning and the potential of mean force. *The Journal of chemical physics*, 147(12):125101, 2017.

[61] R Byron Bird, Warren E Stewart, Edwin N Lightfoot, and Daniel J Klingenberg. *Introductory transport phenomena*. Wiley Global Education, 2015.

[62] Paula Yurkanis Bruice. Organic Chemistry. International Edition, 2004.

[63] Dominik Marx and Jurg Hutter. Ab initio molecular dynamics: Theory and implementation. *Modern methods and algorithms of quantum chemistry*, 1(301-449):141, 2000.

[64] Justin A Lemkul, Jing Huang, Benoît Roux, and Alexander D MacKerell Jr. An empirical polarizable force field based on the classical drude oscillator model: development history and recent applications. *Chemical reviews*, 116(9):4983–5013, 2016.

[65] Nuria Plattner, Stefan Doerr, Gianni De Fabritiis, and Frank Noé. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nature chemistry*, 9(10):1005, 2017.

[66] Caitlyn M Wolf, Kiran H Kanekal, Yeneneh Y Yimer, Madhusudan Tyagi, Souleymane Omar-Diallo, Viktoria Pakhnyuk, Christine K Luscombe, Jim Pfaendtner, and Lilo D Pozzo. Assessment of molecular dynamics simulations for amorphous poly (3-hexylthiophene) using neutron and X-ray scattering experiments. *Soft matter*, 15(25):5067–5083, 2019.

[67] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[68] MJ Abraham, D Van Der Spoel, E Lindahl, and B Hess. The GROMACS development team GROMACS user manual version 5.0. 4. *J. Mol. Model*, 2014.

[69] Berk Hess, Henk Bekker, Herman JC Berendsen, and Johannes GEM Fraaije. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry*, 18(12):1463–1472, 1997.

[70] Alchemistry.org. `http://www.alchemistry.org/wiki/Main_Page`. Accessed: 2020-04-28.

[71] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23(2):187–199, 1977.

[72] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[73] Tobias Fink and Jean-Louis Reymond. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *Journal of chemical information and modeling*, 47(2):342–353, 2007.

[74] Andrey A Toropov, Alla P Toropova, Dilya V Mukhamedzhanoval, and Ivan Gutman. Simplified molecular input line entry system (SMILES) as an alternative for constructing quantitative structure-property relationships (QSPR). *Indian Journal of Chemistry - Section A Inorganic, Physical, Theoretical and Analytical Chemistry*, 2005.

[75] Norbert Haider. Functionality pattern matching as an efficient complementary structure/reaction search tool: an open-source approach. *Molecules*, 15(8):5079–5092, 2010.

[76] W Tschöp, Kurt Kremer, J Batoulis, Ta Bürger, and O Hahn. Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates. *Acta Polymerica*, 49(2-3):61–74, 1998.

[77] Sergei Izvekov and Gregory A Voth. A multiscale coarse-graining method for biomolecular systems. *The Journal of Physical Chemistry B*, 109(7):2469–2473, 2005.

[78] Aviel Chaimovich and M Scott Shell. Coarse-graining errors and numerical optimization using a relative entropy framework. *The Journal of chemical physics*, 134(9):094112, 2011.

[79] Siewert J Marrink, H Jelger Risselada, Serge Yefimov, D Peter Tieleman, and Alex H De Vries. The MARTINI force field: coarse grained model for biomolecular simulations. *The journal of physical chemistry B*, 111(27):7812–7824, 2007.

[80] Maghesree Chakraborty, Chenliang Xu, and Andrew D White. Encoding and selecting coarse-grain mapping operators with hierarchical graphs. *The Journal of Chemical Physics*, 149(13):134106, 2018.

[81] Marco Giulini, Roberto Menichetti, M Scott Shell, and Raffaello Potestio. An information theory-based approach for optimal model reduction of biomolecules. *arXiv preprint arXiv:2004.03988*, 2020.

[82] Dirk Reith, Mathias Pütz, and Florian Müller-Plathe. Deriving effective mesoscale potentials from atomistic simulations. *Journal of computational chemistry*, 24(13):1624–1636, 2003.

[83] Nicholas JH Dunn, Kathryn M Lebold, Michael R DeLyser, Joseph F Rudzinski, and William George Noid. BOCS: Bottom-up open-source coarse-graining software. *The Journal of Physical Chemistry B*, 122(13):3363–3377, 2017.

[84] Nicholas JH Dunn and William George Noid. Bottom-up coarse-grained models that accurately describe the structure, pressure, and compressibility of molecular liquids. *The Journal of chemical physics*, 143(24):243148, 2015.

[85] Avisek Das and Hans C Andersen. The multiscale coarse-graining method. V. Isothermal-isobaric ensemble. *The Journal of chemical physics*, 132(16):164106, 2010.

[86] Siewert J Marrink and D Peter Tieleman. Perspective on the Martini model. *Chemical Society Reviews*, 42(16):6801–6822, 2013.

[87] Luca Monticelli, Senthil K Kandasamy, Xavier Periole, Ronald G Larson, D Peter Tieleman, and Siewert-Jan Marrink. The MARTINI coarse-grained force field: extension to proteins. *Journal of chemical theory and computation*, 4(5):819–834, 2008.

[88] Siewert J Marrink, Alex H De Vries, and Alan E Mark. Coarse grained model for semiquantitative lipid simulations. *The Journal of Physical Chemistry B*, 108(2):750–760, 2004.

[89] Bart MH Bruininks, Paulo CT Souza, and Siewert J Marrink. A practical view of the martini force field. In *Biomolecular Simulations*, pages 105–127. Springer, 2019.

[90] Tristan Bereau and Kurt Kremer. Automated parametrization of the coarse-grained Martini force field for small organic molecules. *Journal of chemical theory and computation*, 11(6):2783–2791, 2015.

[91] Igor V Tetko and Vsevolod Yu Tanchuk. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *Journal of chemical information and computer sciences*, 42(5):1136–1145, 2002.

[92] Igor V Tetko and Pierre Bruneau. Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database. *Journal of pharmaceutical sciences*, 93(12):3103–3110, 2004.

[93] Bradley Efron. Bayes' theorem in the 21st century. *Science*, 340(6137):1177–1178, 2013.

[94] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[95] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[96] Boaz Nadler and Meirav Galun. Fundamental limitations of spectral clustering. In *Advances in neural information processing systems*, pages 1017–1024, 2007.

[97] Ricardo JGB Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1):1–51, 2015.

[98] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

[99] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[100] Michael AA Cox and Trevor F Cox. Multidimensional scaling. In *Handbook of data visualization*, pages 315–347. Springer, 2008.

[101] Mukund Balasubramanian and Eric L Schwartz. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002.

[102] Payel Das, Mark Moll, Hernan Stamati, Lydia E Kavraki, and Cecilia Clementi. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proceedings of the National Academy of Sciences*, 103(26):9885–9890, 2006.

[103] Mary A Rohrdanz, Wenwei Zheng, Mauro Maggioni, and Cecilia Clementi. Determination of reaction coordinates via locally scaled diffusion map. *The Journal of chemical physics*, 134(12):03B624, 2011.

[104] Michele Ceriotti, Gareth A Tribello, and Michele Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences*, 108(32):13023–13028, 2011.

[105] Gareth A Tribello and Piero Gasparotto. Using Data-Reduction Techniques to Analyze Biomolecular Trajectories. In *Biomolecular Simulations*, pages 453–502. Springer, 2019.

[106] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[107] Vladimir Vovk. Kernel ridge regression. In *Empirical inference*, pages 105–116. Springer, 2013.

[108] Raghunathan Ramakrishnan and O Anatole von Lilienfeld. Machine learning, quantum mechanics, and chemical compound space. *arXiv preprint arXiv:1510.07512*, 2015.

[109] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[110] Christian Hoffmann, Roberto Menichetti, Kiran H Kanekal, and Tristan Bereau. Controlled exploration of chemical space by machine learning of coarse-grained representations. *Physical Review E*, 100(3):033302, 2019.

[111] Bing Huang and O. Anatole von Lilienfeld. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *The Journal of Chemical Physics*, 145(16):161102, 2016.

[112] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[113] David J Wales and Jonathan PK Doye. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, 1997.

[114] David J Wales and Harold A Scheraga. Global optimization of clusters, crystals, and biomolecules. *Science*, 285(5432):1368–1372, 1999.

[115] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[116] Jop Briët and Peter Harremoës. Properties of classical and quantum Jensen-Shannon divergence. *Physical review A*, 79(5):052311, 2009.

[117] O Anatole Von Lilienfeld, Raghunathan Ramakrishnan, Matthias Rupp, and Aaron Knoll. Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *International Journal of Quantum Chemistry*, 115(16):1084–1093, 2015.

[118] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87(18):1–16, 2013.

[119] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.

[120] Bing Huang and O. Anatole von Lilienfeld. Efficient accurate scalable and transferable quantum machine learning with am-ons, 2017.

[121] Bing Huang, Nadine O. Symonds, and O. Anatole von Lilienfeld. *Quantum machine learning in chemistry and materials*, pages 1883–1909. Springer International Publishing, Cham, 2020.

[122] Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769, 2016.

[123] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 46(1-3):3–26, 2001.

[124] Kiyohiko Sugano, Manfred Kansy, Per Artursson, Alex Avdeef, Stefanie Bendels, Li Di, Gerhard F Ecker, Bernard Faller, Holger Fischer, Grégori Gerebtzoff, et al. Coexistence of passive and carrier-mediated processes in drug transport. *Nature reviews Drug discovery*, 9(8):597–614, 2010.

[125] CY Lin. Uptake of anaesthetic gases and vapours. *Anaesthesia and intensive care*, 22(4):363–373, 1994.

[126] P Thanikaivelan, V Subramanian, J Raghava Rao, and Balachandran Unni Nair. Application of quantum chemical descriptor in quantitative structure activity and structure property relationship. *Chemical Physics Letters*, 323(1-2):59–70, 2000.

[127] Charles Pidgeon, Shaowei Ong, Hanlan Liu, Xiaoxing Qiu, Mary Pidgeon, Anne H Dantzig, John Munroe, William J Hornback, and Jeffery S Kasher. Iam chromatography: an in vitro screen for predicting drug membrane permeability. *Journal of medicinal chemistry*, 38(4):590–594, 1995.

[128] Mehran Yazdanian, Susan L Glynn, James L Wright, and Amale Hawi. Correlating partitioning and caco-2 cell permeability of structurally diverse small molecular weight compounds. *Pharmaceutical research*, 15(9):1490–1494, 1998.

[129] Mario Orsi and Jonathan W Essex. Passive permeation across lipid bilayers: a literature review. *Molecular Simulations and Biomembranes*, pages 76–90, 2010.

[130] Robert V Swift and Rommie E Amaro. Back to the future: can physical models of passive membrane permeability help reduce drug candidate attrition and move us beyond QSPR? *Chemical biology & drug design*, 81(1):61–71, 2013.

[131] Jared M Diamond and Yehuda Katz. Interpretation of nonelectrolyte partition coefficients between dimyristoyl lecithin and water. *J. Membrane Biol.*, 17(1):121–154, 1974.

[132] Lane W Votapka, Christopher T Lee, and Rommie E Amaro. Two relations to estimate membrane permeability using milestoning. *The Journal of Physical Chemistry B*, 120(33):8606–8616, 2016.

[133] Christopher T Lee, Jeffrey Comer, Nelson Leung Conner Herndon, Anna Pavlova, Robert V Swift, Chris Tung, Christopher N Rowley, Rommie E Amaro, Christophe Chipot, Yi Wang, and James C Gumbart. Simulation-based approaches for determining membrane permeability of small compounds. *J. Chem. Inf. Model.*, 56(4):721, 2016.

[134] Brian J Bennion, Nicholas A Be, Margaret Windy McNerney, Victoria Lao, Emma M Carlson, Carlos A Valdez, Michael A Malfatti, Heather A Enright, Tuan H Nguyen, Felice C Lightstone, and Timothy S Carpenter. Predicting a drug's membrane permeability: A computational model validated with in vitro permeability assay data. *The Journal of Physical Chemistry B*, 121(20):5228–5237, 2017.

[135] Chi Hang Tse, Jeffrey Comer, Yi Wang, and Christophe Chipot. The link between membrane composition and permeability to drugs. *Journal of chemical theory and computation*, 14:2895–2909, 2018.

[136] W. G. Noid. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.*, 139(9), 2013.

[137] Xavier Periole and Siewert-Jan Marrink. The Martini coarse-grained force field. *Biomolecular Simulations: Methods and Protocols*, pages 533–565, 2013.

[138] Roberto Menichetti, Kiran H Kanekal, and Tristan Bereau. Drug–membrane permeability across chemical space. *ACS Central Science*, 5(2):290–298, 2019.

[139] Tristan Bereau and Kurt Kremer. Automated Parametrization of the Coarse-Grained Martini Force Field for Small Organic Molecules. *J. Chem. Theory Comput.*, 11(6):2783–2791, 2015.

[140] Tobias Fink and Jean-louis Reymond. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F. *J. Chem. Inf. Model.*, 47(2):342–353, 2007.

[141] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable mol. simulat. *J. Chem. Theory Comput.*, 4:435−447, 2008.

[142] Siewert J. Marrink, Alex H. de Vries, and Alan E. Mark. Coarse grained model for semiquantitative lipid simulations. *J. Phys. Chem. B*, 108:750–760, 2004.

[143] Siewert J. Marrink, H. Jelger Risselada, Serge Yefimov, D. Peter Tieleman, and Alex H. de Vries. The martini force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111:7812–7824, 2007.

[144] Luca Monticelli, Senthil K. Kandasamy, Xavier Periole, Ronald G. Larson, D. Peter Tieleman, and Siewert-Jan Marrink. The martini coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.*, 4:819–834, 2008.

[145] D. H. De Jong, G. Singh, W. D. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schafer, X. Periole, D. P. Tieleman, and S. J. Marrink. Improved parameters for the martini coarse-grained protein force field. *J. Chem. Theory Comput.*, 9:687–697, 2013.

[146] Djurre H De Jong, Svetlana Baoukina, Helgi I Ingólfsson, and Siewert J Marrink. Martini straight: Boosting performance using a shorter cutoff and gpus. *Comput. Phys. Commun.*, 199:1–7, 2016.

[147] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52(12):7182–7190, 1981.

[148] Tsjerk A Wassenaar, Helgi I Ingólfsson, Rainer A Böckmann, D Peter Tieleman, and Siewert J Marrink. Computational lipidomics with insane: a versatile tool for generating custom membranes for molecular simulations. *Journal of chemical theory and computation*, 11(5):2144–2155, 2015.

[149] Tristan Bereau, Zun-Jing Wang, and Markus Deserno. More than the sum of its parts: Coarse-grained peptide-lipid interactions from a simple cross-parametrization. *J. Chem. Phys.*, 140(11):03B615_1–11220, 2014.

[150] S Jakobtorweihen, A Chaides Zuniga, T Ingram, T Gerlach, FJ Keil, and I Smirnova. Predicting solute partitioning in lipid bilayers: Free energies and partition coefficients from molecular dynamics simulations and cosmomic. *J. Chem. Phys.*, 141(4):07B622_1, 2014.

[151] Tristan Bereau and Robert H Swendsen. Optimized convergence for multiple histogram analysis. *J. Comput. Phys.*, 228(17):6119–6129, 2009.

[152] Jochen S Hub, Bert L De Groot, and David Van Der Spoel. g_wham: A free weighted histogram analysis implementation including robust error and autocorrelation estimates. *J. Chem. Theory Comput.*, 6(12):3713–3720, 2010.

[153] Christopher Z Mooney and Robert D Duvall. *Bootstrapping: A nonparametric approach to statistical inference*. Sage, 1993.

[154] Christophe Chipot and Andrew Pohorille. *Free energy calculations*. Springer, 2007.

[155] Ana Nikolic, Stéphanie Baud, Sarah Rauscher, and Régis Pomès. Molecular mechanism of β-sheet self-organization at water-hydrophobic interfaces. *Proteins: Structure, Function, and Bioinformatics*, 79(1):1–22, 2011.

[156] Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129(12):124105, 2008.

[157] Charles H Bennett. Efficient estimation of free energy differences from monte carlo data. *J. Comput. Phys.*, 22(2):245–268, 1976.

[158] Anita Toulmin, J Matthew Wood, and Peter W Kenny. Toward prediction of alkane/water partition coefficients. *J. Med. Chem.*, 51(13):3720–3730, 2008.

[159] Joseph F Rudzinski, Kurt Kremer, and Tristan Bereau. Communication: Consistent interpretation of molecular simulation kinetics using markov state models biased with external information. *The Journal of Chemical Physics*, 144(5):051102, 2016.

[160] Gerhard Hummer. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New Journal of Physics*, 7(1):34, 2005.

[161] Tobias Fink, Heinz Bruggesser, and Jean-Louis Reymond. Virtual exploration of the small-molecule chemical universe below 160 daltons. *Angewandte Chemie International Edition*, 44(10):1504–1508, 2005.

[162] A Patrícia Bento, Anna Gaulton, Anne Hersey, Louisa J Bellis, Jon Chambers, Mark Davies, Felix A Krüger, Yvonne Light, Lora Mak, Shaun McGlinchey, Michal Nowotka, George Papadatos, Rita Santos, and John P Overington. The chembl bioactivity database: an update. *Nucleic acids research*, 42(D1):D1083–D1090, 2014.

[163] Igor V. Tetko and Vsevolod Yu Tanchuk. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *J. Chem. Inf. Comput. Sci.*, 42(5):1136–1145, 2002.

[164] Calculator Plugin of ChemAxon Marvin 17.28.0, 2017.

[165] Chenzhong Liao and Marc C Nicklaus. Comparison of nine programs predicting p k a values of pharmaceutical substances. *Journal of chemical information and modeling*, 49(12):2801–2812, 2009.

[166] Norbert Haider. Functionality pattern matching as an efficient complementary structure/reaction search tool: An open-source approach. *Molecules*, 15(8):5079–5092, 2010.

[167] Pavel V Klimovich, Michael R Shirts, and David L Mobley. Guidelines for the analysis of free energy calculations. *Journal of computer-aided molecular design*, 29(5):397–411, 2015.

[168] Keith Paton. An algorithm for finding a fundamental set of cycles of a graph. *Communications of the ACM*, 12(9):514–518, 1969.

[169] Mordecai Avriel. *Nonlinear programming: analysis and methods*. Courier Corporation, 2003.

[170] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[171] Felix A Faber, Alexander Lindmaa, O Anatole Von Lilienfeld, and Rickard Armiento. Machine learning energies of 2 million elpasolite (a b c 2 d 6) crystals. *Physical review letters*, 117(13):135502, 2016.

[172] Lu Zhang, Jianjun Tan, Dan Han, and Hao Zhu. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug discovery today*, 22(11):1680–1685, 2017.

[173] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1:140022, 2014.

[174] Felix A Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S Schoenholz, George E Dahl, Oriol Vinyals, Steven Kearnes, Patrick F Riley, and O Anatole Von Lilienfeld. Prediction errors of molecular machine learning models lower than hybrid dft error. *Journal of chemical theory and computation*, 13(11):5255–5264, 2017.

[175] I V Tetko, V Y Tanchuk, and a E Villa. Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *J. Chem. Inf. Comput. Sci.*, 41(5):1407–1421, 2001.

[176] P.Y. Bruice. *Organic Chemistry*. Always Learning. Pearson, 2016.

[177] George Lambrinidis, Fotios Tsopelas, Costas Giaginis, and Anna Tsantili-Kakoulidou. *QSAR/QSPR modeling in the design of drug candidates with balanced pharmacodynamic and pharmacokinetic properties*, pages 339–384. Springer International Publishing, Cham, 2017.

[178] Ning-Ning Wang, Jie Dong, Yin-Hua Deng, Min-Feng Zhu, Ming Wen, Zhi-Jiang Yao, Ai-Ping Lu, Jian-Bing Wang, and Dong-Sheng Cao. Adme properties evaluation in drug discovery: Prediction of caco-2 cell permeability using a combination of nsga-ii and boosting. *Journal of Chemical Information and Modeling*, 56(4):763–773, 2016.

[179] AS Christensen, FA Faber, B Huang, LA Bratholm, A Tkatchenko, KR Muller, and OA von Lilienfeld. QML: A Python toolkit for quantum machine learning. *URL https://github. com/qmlcode/qml*, 2017.

[180] Bellman Richard. Dynamic programming. *Princeton University Press*, 89:92, 1957.

[181] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

[182] Kristin P Bennett, Usama Fayyad, and Dan Geiger. Density-based indexing for approximate nearest-neighbor queries. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–243, 1999.

[183] Michael E Houle, Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Can shared-neighbor distances defeat the curse of dimensionality? In *International conference on scientific and statistical database management*, pages 482–500. Springer, 2010.

[184] Johannes H Voigt, Bruno Bienfait, Shaomeng Wang, and Marc C Nicklaus. Comparison of the nci open database with seven large chemical structural databases. *Journal of chemical information and computer sciences*, 41(3):702–712, 2001.

[185] Lemont B Kier and Lowell H Hall. An electrotopological-state index for atoms in molecules. *Pharmaceutical research*, 7(8):801–807, 1990.

[186] PH Howard and William Meylan. Physprop database. *Syracuse Research Corp., Syracuse, NY*, 2000.

[187] Clemens Rauer and Tristan Bereau. Hydration free energies from kernel-based machine learning: Compound-database bias. *The Journal of Chemical Physics*, 153(1):014101, 2020.

[188] Marc Stieffenhofer, Michael Wand, and Tristan Bereau. Adversarial reverse mapping of equilibrated condensed-phase molecular structures. *arXiv preprint arXiv:2003.07753*, 2020.

[189] Kunal Roy, Supratik Kar, and Rudra Narayan Das. Statistical methods in qsar/qspr. In *A primer on QSAR/QSPR modeling*, pages 37–59. Springer, 2015.

[190] Lili Xi, Shanshan Pan, Xinran Li, Yonglin Xu, Jianyue Ni, Xin Sun, Jiong Yang, Jun Luo, Jinyang Xi, Wenhao Zhu, Xinran Li, Di Jiang, Richard Dronskowski, Xun Shi, G. Jeffrey Snyder, and Wenqing Zhang. Discovery of High Performance Thermoelectric Chalcogenides through Reliable High Throughput Material Screening. *Journal of the American Chemical Society*, 140(34):10785–10793, aug 2018.

[191] Asha K. Patel, Mark W. Tibbitt, Adam D. Celiz, Martyn C. Davies, Robert Langer, Chris Denning, Morgan R. Alexander, and Daniel G. Anderson. High throughput screening for discovery of materials that control stem cell fate. *Current Opinion in Solid State and Materials Science*, 20(4):202–211, aug 2016.

[192] Nicolas Mounet, Marco Gibertini, Philippe Schwaller, Davide Campi, Andrius Merkys, Antimo Marrazzo, Thibault Sohier, Ivano Eligio Castelli, Andrea Cepellotti, Giovanni Pizzi, et al. Two-dimensional materials from high-

throughput computational exfoliation of experimentally known compounds. *Nature nanotechnology*, 13(3):246, 2018.

[193] RL Greenaway, V Santolini, MJ Bennison, BM Alston, CJ Pugh, MA Little, M Miklitz, EGB Eden-Rump, R Clowes, A Shakil, et al. High-throughput discovery of organic cages and catenanes using computational screening fused with robotic synthesis. *Nature communications*, 9(1):2849, 2018.

[194] Michael C Burns, Jennifer E Howes, Qi Sun, Andrew J Little, DeMarco V Camper, Jason R Abbott, Jason Phan, Taekyu Lee, Alex G Waterson, Olivia W Rossanese, et al. High-throughput screening identifies small molecules that bind to the ras: Sos: Ras complex and perturb ras signaling. *Analytical biochemistry*, 548:44–52, 2018.

[195] Felix A. Faber, Anders S. Christensen, Bing Huang, and O. Anatole von Lilienfeld. Alchemical and structural distribution based representation for universal quantum machine learning. *The Journal of Chemical Physics*, 148(24):241717, jun 2018.

[196] Michael J. Willatt, Félix Musil, and Michele Ceriotti. Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Physical Chemistry Chemical Physics*, 20(47):29661–29668, dec 2018.

[197] Luca M Ghiringhelli, Jan Vybiral, Sergey V Levchenko, Claudia Draxl, and Matthias Scheffler. Big data of materials science: critical role of the descriptor. *Physical review letters*, 114(10):105503, 2015.

[198] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The δ-machine learning approach. *Journal of chemical theory and computation*, 11(5):2087–2096, 2015.

[199] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547, 2018.

[200] Tristan Bereau, Robert A DiStasio Jr, Alexandre Tkatchenko, and O Anatole Von Lilienfeld. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *The Journal of chemical physics*, 148(24):241706, 2018.

[201] Tristan Bereau, Denis Andrienko, and Kurt Kremer. Research Update: Computational materials discovery in soft matter. *APL Materials*, 4(5), 2016.

[202] Tristan Bereau. Data-driven methods in multiscale modeling of soft matter. *Handbook of Materials Modeling: Methods: Theory and Modeling*, pages 1459–1470, 2020.

[203] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. *Chemical reviews*, 116(14):7898–7936, 2016.

[204] W Schommers. Pair potentials in disordered many-particle systems: A study for liquid gallium. *Physical Review A*, 28(6):3599, 1983.

[205] F Ercolessi and J. B Adams. Interatomic Potentials from First-Principles Calculations: The Force-Matching Method. *Europhysics Letters (EPL)*, 26(8):583–588, jun 1994.

[206] Gary S Ayton, Will G Noid, and Gregory A Voth. Multiscale modeling of biomolecular systems: in serial and in parallel. *Current opinion in structural biology*, 17(2):192–198, 2007.

[207] Siewert J Marrink and D Peter Tieleman. Perspective on the Martini model. *Chemical Society reviews*, 42(16):6801–6822, 2013.

[208] Roberto Menichetti, Kiran H. Kanekal, Kurt Kremer, and Tristan Bereau. In silico screening of drug-membrane thermodynamics reveals linear relations between bulk partitioning and the potential of mean force. *The Journal of Chemical Physics*, 147(12):125101, sep 2017.

[209] Roberto Menichetti, Kiran H. Kanekal, and Tristan Bereau. Drug–Membrane Permeability across Chemical Space. *ACS Central Science*, 5(2):290–298, feb 2019.

[210] Siewert J Marrink, H Jelger Risselada, Serge Yefimov, D Peter Tieleman, and Alex H De Vries. The MARTINI Force Field : Coarse Grained Model for Biomolecular Simulations. *Journal of Physical Chemistry B*, 111(June):7812–7824, 2007.

[211] Kiran H. Kanekal and Tristan Bereau. Resolution limit of data-driven coarse-grained models spanning chemical space. `http://doi.org/10.5281/zenodo.3403594`, July 2019.

[212] Tristan Bereau. auto_martini master branch, commit: 2c0c095898860cbe086e64b973b13a45f80137b6. `https://github.com/tbereau/auto_martini`, 2019.

[213] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[214] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[215] Thomas T Foley, M Scott Shell, and William George Noid. The impact of resolution upon entropy and information in coarse-grained models. *The Journal of chemical physics*, 143(24):12B601_1, 2015.

[216] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, mar 1951.

[217] Eric Jones, Travis Oliphant, and Pearu Peterson. {SciPy}: Open source scientific tools for {Python}. *https://www.scipy.org*, 2014.

[218] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[219] Cesar A López, Andrzej J Rzepiela, Alex H De Vries, Lubbert Dijkhuizen, Philippe H Hünenberger, and Siewert J Marrink. Martini coarse-grained force field: extension to carbohydrates. *Journal of Chemical Theory and Computation*, 5(12):3195–3210, 2009.

[220] Semen O Yesylevskyy, Lars V Schäfer, Durba Sengupta, and Siewert J Marrink. Polarizable water model for the coarse-grained martini force field. *PLoS Comput Biol*, 6(6):e1000810, 2010.

[221] Julian Michalowsky, Lars V Schäfer, Christian Holm, and Jens Smiatek. A refined polarizable water model for the coarse-grained martini force field with long-range electrostatic interactions. *The Journal of chemical physics*, 146(5):054501, 2017.

[222] Julian Michalowsky, Johannes Zeman, Christian Holm, and Jens Smiatek. A polarizable martini model for monovalent ions in aqueous solution. *The Journal of chemical physics*, 149(16):163319, 2018.

[223] Hans C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393, feb 1980.

[224] Riccardo Alessandri, Paulo CT Souza, Sebastian Thallmair, Manuel N Melo, Alex H De Vries, and Siewert J Marrink. Pitfalls of the martini model. *Journal of chemical theory and computation*, 15(10):5448–5460, 2019.

[225] Riccardo Alessandri, Jaakko J Uusitalo, Alex H de Vries, Remco WA Havenith, and Siewert J Marrink. Bulk heterojunction morphologies with atomistic resolution from coarse-grain solvent evaporation simulations. *Journal of the American Chemical Society*, 139(10):3697–3705, 2017.

[226] Roberto Menichetti, Kurt Kremer, and Tristan Bereau. Efficient potential of mean force calculation from multiscale simulations: Solute insertion in a lipid membrane. *Biochemical and Biophysical Research Communications*, 498(2):282–287, mar 2018.

[227] Kiran H Kanekal and Tristan Bereau. Resolution limit of data-driven coarse-grained models spanning chemical space. *The Journal of chemical physics*, 151(16):164106, 2019.

[228] Timothy S Carpenter, Cesar A López, Chris Neale, Cameron Montour, Helgi I Ingólfsson, Francesco Di Natale, Felice C Lightstone, and Sandrasegaram Gnanakaran. Capturing phase behavior of ternary lipid mixtures with a refined martini coarse-grained force field. *Journal of chemical theory and computation*, 14(11):6050–6062, 2018.

[229] Matti Javanainen, Balazs Fabian, and Hector Martinez-Seara. Comment on" capturing phase behavior of ternary lipid mixtures with a refined martini coarse-grained force field". *arXiv preprint arXiv:2009.07767*, 2020.

[230] Christopher R Ellis, Joseph F Rudzinski, and William G Noid. Generalized-yvon–born–green model of toluene. *Macromolecular theory and simulations*, 20(7):478–495, 2011.

[231] Joseph F Rudzinski and William G Noid. The role of many-body correlations in determining potentials for coarse-grained models of equilibrium structure. *The Journal of Physical Chemistry B*, 116(29):8621–8635, 2012.

[232] Joseph F Rudzinski and William G Noid. Bottom-up coarse-graining of peptide ensembles and helix–coil transitions. *Journal of chemical theory and computation*, 11(3):1278–1291, 2015.

[233] James F Dama, Anton V Sinitskiy, Martin McCullagh, Jonathan Weare, Benoît Roux, Aaron R Dinner, and Gregory A Voth. The theory of ultra-coarse-graining. 1. general principles. *Journal of chemical theory and computation*, 9(5):2466–2480, 2013.

[234] Greg Landrum. Rdkit documentation. *Release*, 1:1–79, 2013.

[235] Kenno Vanommeslaeghe and Alexander D MacKerell Jr. Automation of the charmm general force field (cgenff) i: bond perception and atom typing. *Journal of chemical information and modeling*, 52(12):3144–3154, 2012.

[236] MJ Abraham, D Van Der Spoel, E Lindahl, and B Hess. The gromacs development team gromacs user manual version 2016.1. 0. *J. Mol. Model*, 2016.

[237] Philippe H Hünenberger. Thermostat algorithms for molecular dynamics simulations. In *Advanced computer simulation*, pages 105–149. Springer, 2005.

[238] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

[239] Wai-Ting Vong and Fuan-Nan Tsai. Densities, molar volumes, thermal expansion coefficients, and isothermal compressibilities of organic acids from 293.15 k to 323.15 k and at pressures up to 25 mpa. *Journal of Chemical & Engineering Data*, 42(6):1116–1120, 1997.

[240] JW Mullinax and William George Noid. Reference state for the generalized yvon–born–green theory: Application for coarse-grained model of hydrophobic hydration. *The Journal of chemical physics*, 133(12):124107, 2010.

[241] Han Wang, Christoph Junghans, and Kurt Kremer. Comparative atomistic and coarse-grained study of water: What do we lose by coarse-graining? *The European Physical Journal E*, 28(2):221–229, 2009.

[242] M Guenza. Thermodynamic consistency and other challenges in coarse-graining models. *The European Physical Journal Special Topics*, 224(12):2177–2191, 2015.

[243] Joseph F Rudzinski and Tristan Bereau. Coarse-grained conformational surface hopping: Methodology and transferability. *arXiv preprint arXiv:2009.08705*, 2020.

[244] Takashi Kato, Norihiro Mizoshita, and Kenji Kishimoto. Functional liquid-crystalline assemblies: self-organized soft materials. *Angewandte Chemie International Edition*, 45(1):38–68, 2006.

[245] Nicholas E Jackson, Brett M Savoie, Kevin L Kohlstedt, Monica Olvera de la Cruz, George C Schatz, Lin X Chen, and Mark A Ratner. Controlling conformations of conjugated polymers and small molecules: The role of nonbonding interactions. *Journal of the American Chemical Society*, 135(28):10475–10483, 2013.

[246] Cristina Greco, Anton Melnyk, Kurt Kremer, Denis Andrienko, and Kostas Ch Daoulas. Generic model for lamellar self-assembly in conjugated polymers: linking mesoscopic morphology and charge transport in p3ht. *Macromolecules*, 52(3):968–981, 2019.

[247] Frank H Stillinger and Thomas A Weber. Inherent structure in water. *The Journal of Physical Chemistry*, 87(15):2833–2840, 1983.

[248] Tristan Bereau and Joseph F Rudzinski. Accurate structure-based coarse graining leads to consistent barrier-crossing dynamics. *Physical Review Letters*, 121(25):256002, 2018.

[249] Maghesree Chakraborty, Jinyu Xu, and Andrew D White. Is preservation of symmetry necessary for coarse-graining? *Physical Chemistry Chemical Physics*, 22(26):14998–15005, 2020.

[250] Thomas T Foley, Katherine M Kidder, M Scott Shell, and WG Noid. Exploring the landscape of model representations. *Proceedings of the National Academy of Sciences*, 2020.

[251] Bingjie Hu, Xin Zhou, Michael A Mohutsky, and Prashant V Desai. Structure-property relationships and machine learning models for addressing cyp3a4-mediated victim drug-drug interaction risk in drug discovery. *Molecular Pharmaceutics*, 2020.

[252] Sk Abdul Amin, Nilanjan Adhikari, Shovanlal Gayen, and Tarun Jha. An integrated ligand-based modelling approach to explore the structure-property relationships of influenza endonuclease inhibitors. *Structural Chemistry*, 28(6):1663–1678, 2017.

[253] Fan-Chen Kong, Ye-Fei Li, Cheng Shang, and Zhi-Pan Liu. Stability and phase transition of cobalt oxide phases by machine learning global potential energy surface. *The Journal of Physical Chemistry C*, 123(28):17539–17547, 2019.

[254] Alan Cooper. Thermodynamic fluctuations in protein molecules. *Proceedings of the National Academy of Sciences*, 73(8):2740–2741, 1976.

[255] Ji-Hwan Lee, Sung-Min Choi, Changwoo Doe, Antonio Faraone, Philip A Pincus, and Steven R Kline. Thermal fluctuation and elasticity of lipid vesicles interacting with pore-forming peptides. *Physical review letters*, 105(3):038101, 2010.

[256] James F Dama, Jaehyeok Jin, and Gregory A Voth. The theory of ultra-coarse-graining. 3. coarse-grained sites with rapid local equilibrium of internal states. *Journal of chemical theory and computation*, 13(3):1010–1022, 2017.

[257] Brooke E Husic, Nicholas E Charron, Dominik Lemm, Jiang Wang, Adrià Pérez, Andreas Krämer, Yaoyi Chen, Simon Olsson, Gianni de Fabritiis, Frank Noé, et al. Coarse graining molecular dynamics with graph neural networks. *arXiv preprint arXiv:2007.11412*, 2020.

[258] M Scott Shell. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *The Journal of chemical physics*, 129(14):144108, 2008.

[259] Zhiheng Li, Geemi P Wellawatte, Maghesree Chakraborty, Heta A Gandhi, Chenliang Xu, and Andrew D White. Graph neural network based coarse-grained mapping prediction. *Chemical Science*, 2020.