A SCREENING STUDY OF TRIBOLOGICAL PROPERTIES FOR THIN FILMS

By

Co Dai Quach

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTERS OF SCIENCE

in

Chemical Engineering

March 31, 2023

Nashville, Tennessee

Approved:

Clare McCabe, Ph.D.

Peter Cummings, Ph.D.

To my friends and family whose support made this possible.

ACKNOWLEDGMENTS

Funding for this work is provided by the National Science Foundation (NSF) through Grants OAC-1835874 and DMR-1852157. This research used resources provided by the Office of Science of the Department of Energy at the Oak Ridge Leadership Computing Facility operated under Contract DE-AC05-00OR22725 via an award from the INCITE program and the National Energy Research Scientific Computing Center (NERSC) operated under Contract DE-AC02-05CH11231.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Microelectromechanical systems (MEMS) and nanoelectromechanical systems (NEMS) are devices on the scale of a few micrometers or even nanometers. These systems play an important role in a wide range of applications, including medical, automotive, aerospace, consumer electronics, *etc*. The micro/nano scale of MEMS and NEMS devices presents unique challenges to the design and operation of these devices. Due to their minuscule size, these devices are affected by interfacial forces and the atomic-level roughness of the surfaces, leading to stiction and wear issues.[1,2] The high surface-to-volume ratio also makes these devices more susceptible to contamination, which can further exacerbate the wear and stiction problems.[1,2]

Ongoing research and development efforts are addressing these issues and improving the performance of these devices in various applications. Unfortunately, traditional lubricants, such as long-chain alkanes, are ineffective at this scale. These lubricants undergo a phase transition in a nano-confined gap and become too viscous to reach all of the crevices of MEMS and NEMS.[3] Alternative materials and fabrication techniques that can reduce surface roughness and improve tribological properties are being explored and developed. For example, using nanoscale coatings and surface treatments, such as diamond-like carbon[4–6] and tetrafluoroethylene,[7] can help reduce degradation. Additionally, the development of novel lubricants, such as graphene-based[8,9] and self-assembled monolayer lubricants,[2,10,11] has shown promise in improving the tribological performance of these devices.

1.1 Thin Films Coatings as Lubricants

A lubricant aids in transmitting forces, particles, or energy between two surfaces, and lower frictional forces of surfaces in contacts. The effectiveness of the lubricants can be measured by their ability to minimize frictional forces, and can be measured by the force of adhesion (F_0) and coefficient of friction (COF) as depicted in Equation 2.1. Force of adhesion, also referred to as adhesive force, measures the force required to pull two surfaces in contact apart. While COF is a metric quantifying the amount of a resisting force divided by the force in the perpendicular direction pushing the surfaces together. Some examples of common COF's are static, kinetic, deformation, molecular, and rolling coefficients. In this work, we are measuring the kinetic COF, which quantifies the force restricting movement of one surface relative to its neighboring surface. Among the alternative lubrication schemes, thin film coatings are seen as a promising solution for reducing friction and wear in mechanical devices that have micro and nanoscale surface separations.[1,2] These films can be fine-tuned through changes in their terminal group chemistries, backbone chain length,

backbone chemistry, and film composition. All have impacted their lubricating effectiveness along with other properties, such as durability, solvent interactions, and thermal response.[11,12]

Particularly, the terminal group, as shown in various experimental and computational studies, has a significant impact on the tribological response of the thin film coating. For example, phenyl-terminated monolayer thin films have been shown to yield higher frictional forces than methyl-terminated films, explained by the twisting ability of the phenyl groups, hampering movement during shear.[13] Meanwhile, hydroxylated and carboxylated films had higher frictional and adhesive forces (F_0) compared to methyl-terminated films, attributed by their capability to form inter-monolayer hydrogen bonds. In addition to traditional experiments, computational studies have also reported similar trends.[12,14] Having more than one chemistry at an interface introduces additional cross-interactions that could alter the lubricating properties. Experiments by Brewer *et al.* demonstrated that a methyl-functionalized microscope tip in contact with either hydroxyl or carboxyl terminated monolayers results in a lower COF compared to the same tip in contact with a methyl terminated monolayer.[14] Furthermore, introducing heterogeneity to individual monolayer films, *i.e.*, having two or more terminal group chemistries within the same surface or layer, offers the potential to get better lubricating performance. Computational studies conducted by Lewis *et al.* for monolayers composed of methyl-terminated alkanes mixed with perfluoroalkanes demonstrate a regime where the COF is reduced compared to either pure component system.[15] The cross-interactions between the multitude of terminal groups in these systems are complex, as hinted by Le *et al.*[16] However, this presents an information-rich parameter space that could offer more insight into how to design a perfect lubricating thin film.

All of these studies, whether experimental or computational, confirm the prominent role of terminal groups in the lubricating ability of monolayer films. Only examining terminal group chemistry creates a vast realm of research to explore on its own where different chemical parameters can be used to optimize film properties. However, when combining this plethora of information with the monolayer film design, it will provide thousands of lubricating properties. By collecting this database of knowledge, we can better understand the quantitative structure-property relationships (QSPR) of these films.[12,17,18]

1.2 Molecular Dynamics and High-throughput Screening

Despite the many benefits, investigating a vast parameter space which could range from thousands to tens of thousands combinations, poses an overwhelming challenge for any researcher. The shear number of potential thin film design makes it practically impossible to be studied using traditional experimental methods. experimental methods can be incredibly expensive, time-consuming, and highly prone to error. Thus, in order to make a dent in the many combinations, we can apply computational techniques, which can be scaled up by the ever-increasing availability of computing resources.

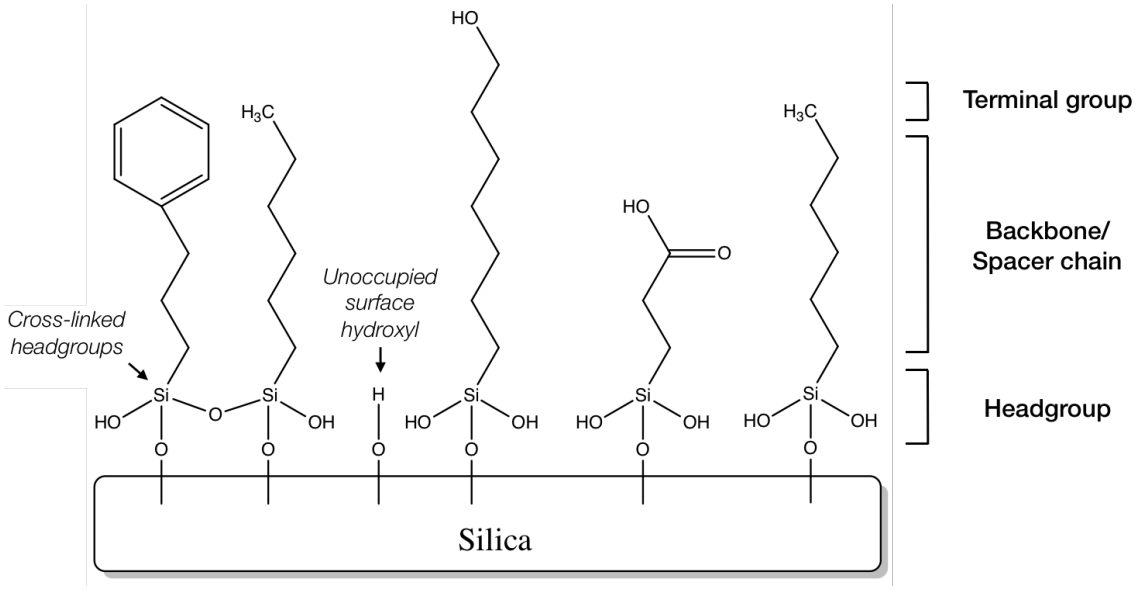

Figure 1.1: Visualization of monolayer thin films studied with optimizable components, *e.g.*, terminal group, backbone, backbone chain length, degree of crosslinking, and film composition.

As computing resources become more advanced and integrated into all science fields, we can simulate more systems in record times. MD simulations, a popular method of study, are increasingly being utilized as a powerful tool for studying complex chemical and physical systems. This avoids the need to develop experimental synthesis techniques, which may be non-trivial and time intensive. One of the key benefits of MD simulations is their ability to perform large-scale sweeps of the accessible parameter space.

To study these thin films on the nanoscale, MD simulations are an excellent candidate to quickly and effectively reveal the intrinsic properties associated with defect-free films on pristine contaminant-free surfaces. Such approach has been utilized by Summers *et al.* to study and optimize various parameters describing the monolayer, including backbone chain length, chain densities, and a small collection of terminal chemistries.[12] Recent development of the Molecular Simulation and Design Framework (MoSDeF)[19] and the Signac Framework[20] enables the systematic and automated large-scale screening of soft matter systems. This allows for reproducible initialization and parameterization of systems in order to manage these large data spaces. These tools have been used to design and perform large scale molecular dynamics screening studies in several recent studies[12,21,22] as well as used to fully capture the provenance of simulation workflows for increased reproducibility in other work.[23,24] The idea of maintaining repeatability in molecular simulations is an obvious and yet until recently, unregulated subject. Many simulations fail to fully disclose the parameters, necessary codes, data analysis, etc., in regard to their systems of interest. Even if the work is transparent and all the materials accessible to others, if the resources are not transferable from one computer to another,

this constitutes the work as not reproducible. Adopting open-source tools and standards, those that have undergone extensive review and testing, is essential to mitigate the reproducibility issue that is lurking in the computational community.

1.3 Machine Learning Integration

The optimization of monolayer films using merely brute force computational screening, at certain point, can become impractical. As the parameter space is further expanded, *i.e.*, when more variables are introduced, the number of system configurations to be considered could increase to near infinite, making it impossible to be fully surveyed. Instead, a more efficient approach can be achieved by combining computational screening with machine learning (ML) techniques, where data generated from MD simulations are used to train predictive models. These models help mitigate the workload for the high-throughput screening by extrapolating trends found from small subsets of data to predict properties for the entire parameter space. This information can be used to guide subsequent steps to either only focus on systems with a high likelihood of possessing desirable properties or strategically simulating systems that best improve the understanding of underlying QSPR trends. As a result, this saves time and resources that may have been wasted simulating configurations that offer little value.

ML is a technique with ever-increasing popularity, operating as a branch of artificial intelligence (AI), which utilizes data and algorithms to imitate how humans learn and apply thought processes. These techniques have been applied in conjunction with other traditional computational simulation techniques, *e.g.*, MD and Density Functional Theory (DFT), in various subjects, including protein folding,[25–27] polymers/-monomers optimization,[28–30] molecular properties prediction,[31,32] all with great success. Depending on the specific quantity of interest, different machine learning algorithms can be applied. Relevant to this work, Neural Network[30,32] and Random Forest,[12,33] are useful methods that exhibit the great benefits of applying ML to analyze and project *in silico* data. As the use of *in silico* data in predictive designs continues to increase, the ability to rapidly generate, screen, learn, and predict from this data will be valuable tools for both computational and experimental researchers.

1.4 Prior Work and Scope of Study

In prior work, we developed a screening framework to explore the role of terminal group chemistry on thin film frictional properties under shear. The study focused on uniform monolayers with 16 different terminal group chemistries, with each monolayer terminal group chemistry independently varied to study trends and combinations of chemistries that provided favorable tribological properties. The COF and F_0 were used to quantify the systems.[12] Data from 100 different monolayer combinations were analyzed with the random for-

est regression model, creating ML models that predict COF and F_0 solely from the chemistry of the terminal group. ML models need a certain amount of training data in order to be accurate. Too little training data and the model loses accuracy; too much training data and the algorithm over-predicts and wastes resources. Despite the limited amount of data available for training, the ML models still exhibit good predicting capability. The success of the models suggest the benefit of utilizing a similar approach to pre-screen parameter space and accelerate the discovery of monolayer combinations with desirable properties.

Expanding upon the previous results, this study[1] will consider a larger parameter space by introducing a heterogeneous monolayer, *i.e.*, having two terminal group chemistries in one surface. Specifically, we consider systems in which one monolayer consists of a single homogeneous or uniform terminal group (C in Figure 2.2), while the other monolayer is made up of two chemistries (A and B in Figure 2.2 a), with varying relative compositions. The study considers a pool of 19 different chemistries, creating 9747 unique dual-monolayer systems, when accounting for the mixing ratio of terminal groups in the mixed monolayer. With the generated data, the ML random forest algorithm will be employed to create a predictive model, allowing for the further projection of the screened data. In chapter 2 we provide an overview of the computational approach, focusing on the simulation workflow, analysis methods, and the ML model. In the results section of chapter 3, we present the data generated from the MD screening and identify key terminal groups and combinations associated with improved tribological performance. In chapter 4, we will investigate suitable strategies for utilizing ML models to guide future work. This includes creating predictive models and apply them toward screening of even larger data spaces (193,131 unique systems, created from 621 chemistries from the CheMBL library.[34,35]) Finally, we experiment with hypothetical scenarios where ML is further integrated with the high-throughput screening process to speed up similar high-throughput screening processes, reducing required time and computing resources. All relevant information to maintain reproducibility in generating this data and workflow is readily available and adaptable for others to use, following the principle of TRUE (Transferable, Reproducible, Usable by others and Extensible) simulations described by Thompson *et al.*.[23] The instructions to access the accompanied GitHub repository is described in Appendix A.

[1]This work is adapted and reproduced with permission from AIP Publishing from the following work (Quach, C. D., Gilmer, J. B., Pert, D. O., Mason-Hogans, A., Iacovella, C. R., Cummings, P. T. & McCabe, C. High-Throughput Screening of Tribological Properties of Monolayer Films Using Molecular Dynamics and Machine Learning. *The Journal of Chemical Physics,* 5.0080838. ISSN: 0021-9606, 1089-7690 [Feb. 2022])

CHAPTER 2

Methods

2.1 Molecular Dynamics Setup

MD is a computer simulation method used in computational chemistry to study the behavior and interactions of molecules. It uses numerical methods to integrate the equations of motion for a system of atoms or molecules, taking into account their interactions with each other and with their environment, over a period of time. The output from an MD simulation is a series of snapshots that describe the positions and motions of the atoms or molecules in the system, allowing researchers to study the molecular-scale behavior and properties of the system. This method can help provide insights into the behavior of complex molecular systems that are difficult to study experimentally, as well as predict properties and behaviors of a molecular system. MD simulations allow researcher to screen through a vast parameter space, studying the various properties whether thermodynamic or dynamic in operating under different conditions.

In this study, we are considering a system that is made up of two opposing amorphous silica surfaces, each coated with an alkylsilane monolayer, forming a dual-monolayer. Each surface, sized 5 *nm* x 5 *nm*, is created followed the procedure described by Summers *et al.*,[36] and is available in the form of a Python script, with the instructions to access available in the Appendix A. The silica surfaces have an average surface roughness of 0.11 *nm*, a desirable approximation based off of Black *et al.*, which utilized a more computationally intensive synthesis mimetic simulation.[37] 100 alkylsilane chains are chemisorbed to each surface, resulting in a surface density of 4 $chains/nm^2$. The surface density is consistent with prior computational[12,36,37] and experimental[2] studies that determine chain surface densities to be between 4.0-5.0 $chains/nm^2$. Each alkylsilane chain is a fully saturated 17 carbon backbone capped with a terminal group from Figure 2.2 b. The remaining uncoordinated oxygens were hydrogenated to mimic surface oxidation.

Each monolayer system was prepared using the MoSDeF software suite.[19,38,39] Specifically, the initialization of the monolayer structure is encapsulated as an mBuild recipe,[38,40] which preserves the entire process used to construct the monolayer structure. The foyer library[39,41] was used to atomtype and parameterize each system with the Optimized Potential for Liquid Simulation - All Atoms (OPLS-AA) forcefield.[42] A forcefield is where bonded and non-bonded information for each atom is stored. Bonded interactions include: bonds, angles, and torsions. Non-bonded parameters include electrostatics and van der Waals interactions. Parameters for the alkylsilane chains were taken from GROMACS 5.1[43,44] and those for the silica surface from Lorenz *et al.*[45] The force field details are provided in Appendix B and is also available in the accompanied

GitHub repository (see Appendix A). The particle-particle particle-mesh (PPPM) algorithm was used to calculate the long-range electrostatic interactions, using a force and pressure correction in the z-dimension to support slab geometries; systems are periodic in the monolayer plane.[44,46] The original script for the process described above is available in the accompanied GitHub repository (see Appendix A).

MD simulations were performed using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) and GROMACS simulation engines.[43,44,47,48] LAMMPS was solely used to relieve the system of initial high-energy configurations and possible atom overlaps. The more stable structure generated was then fed into GROMACS to perform the rest of the simulation workflow, starting with energy minimization following a steepest descent algorithm, followed by a 1 ns equilibration in the canonical (NVT) ensemble at 298 K using the Nóse-Hoover thermostat. An NVT simulation was then performed at 298 K in which the two surfaces were brought into contact by applying a constant normal force of 5 nN along the z direction to the bottom surface over 0.5 ns, allowing for the distance between the two surfaces to reach a steady state. After compression, shearing simulations (with surfaces moving at relative speed of 10 ms^{-1}) were performed at 3 different normal loads of 5 nN, 15 nN, and 25 nN. Specifically, the shearing process is simulated by pulling a ghost particle, which is coupled to the top surface via a harmonic spring with a spring constant of 10,000 $kJ/(mol.nm^2)$, in the x direction at 10 ms^{-1}. The shear is simulated for 10 ns, and the last 5 ns is used for analysis (production regime). An example of a dual-monolayer system is shown in Figure 2.1

2.2 High-Throughput Screening Workflow

High-throughput screening is a scientific technique used in drug discovery, chemical biology, and materials science to rapidly test large numbers of samples for specific biological or chemical activities. It involves the use of automated equipment and processes to perform multiple tests in parallel, enabling researchers to screen thousands of compounds or samples in a short amount of time. The goal of high-throughput screening is to identify potential leads for further study, with the ultimate aim of discovering new drugs, chemicals, or materials with desired properties. Depending on the type of experiments, high-throughput screening combines the use of high-speed robotic systems, microfluidic technologies, and high-throughput analytical instruments, or purely computational simulation tools to screen large libraries of compounds or biological samples. The results of screening are used to prioritize compounds or samples for further study, which can be validated through more detailed and time-consuming laboratory experiments.

For our application, the high-throughput screening with MD can be used to scan through large parameter space to determine thin film coatings that can provide beneficial lubrication properties for nano-scale surfaces in contact. Of the two surfaces in the dual monolayer systems, as detailed in the previous section, the bottom surface is homogeneous (singular terminal group C in Figure 2.2 a), and the top surface contains a mixture

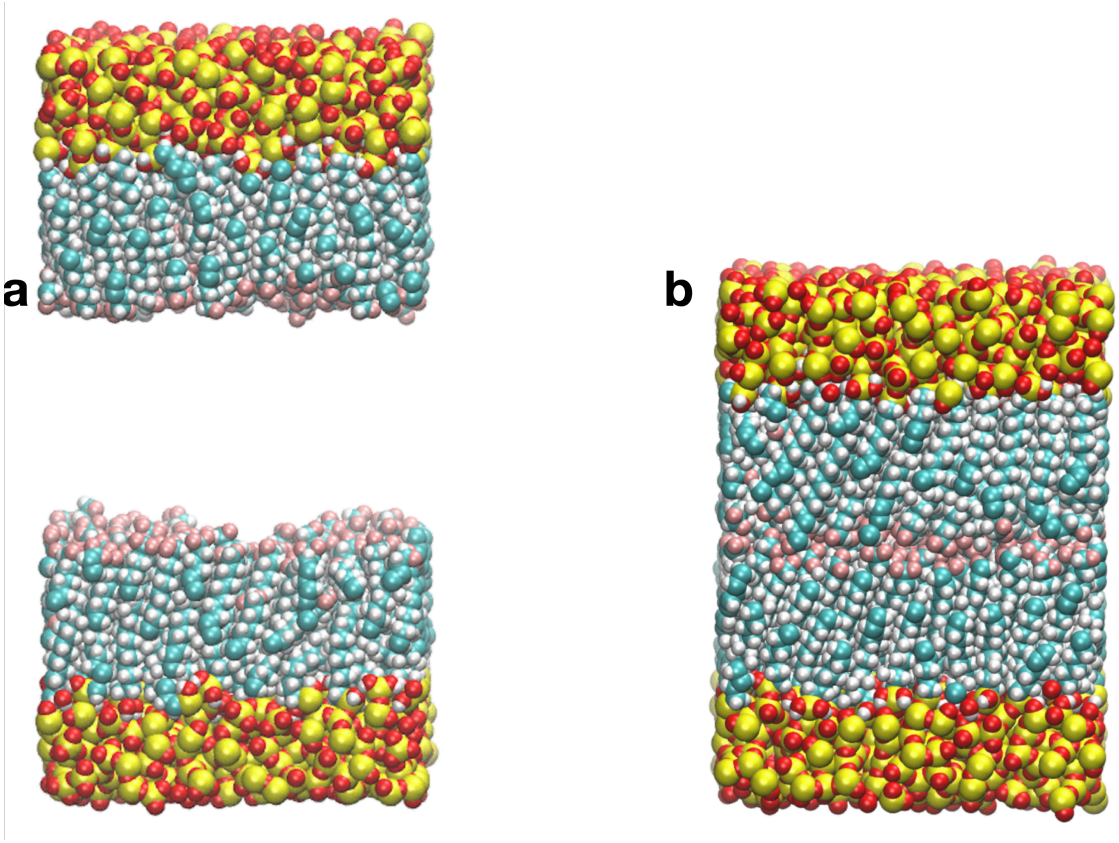

Figure 2.1: Snapshot of dual-monolayer system (difluoro-terminated on both top and bottom monolayer) during the energy minimization (a) and shearing (b) stage

of two types of alkylsilane chains (groups A and B in Figure 2.2 a), differing by their terminal groups. The mixing ratios for the top monolayers considered in this study include 0.5:0.5 and 0.25:0.75. The pool of 19 different terminal group chemistries investigated are shown in Figure 2.2 b; this adds 3 additional terminal group chemistries to those considered by Summers *et al.*[12] The uniform bottom monolayer and the mixed top monolayer can be composed of any combination of groups from Figure 2.2 b, with the constraint that the two groups in the mixed monolayer must be different. In total 12,996 combinations ([19 terminal groups in uniform layer] * [19 * 18 terminal group combinations in mixed layer] * [2 composition ratios]) were considered; this translates to a total of 116,964 simulations (12,996 * 3 * 3) when factoring in the composition ratios studied, the 3 normal loads, and 3 replicates considered for each system. Of the 12,996 systems considered, 3249 systems with the mixing ratio in the top monolayer of 0.5:0.5 were duplicated during the screening and thus such combinations had 3 additional replicates; in total 9747 unique combinations ($19*19*18$ of 0.25:0.75 systems + $1/2*19*19*18$ of 0.5:0.5 systems) were considered.

Managing and executing screening workflows, such as described above can be challenging. Major hurdles include the ability to automate each step of the simulation workflow, managing large workspaces, and mon-

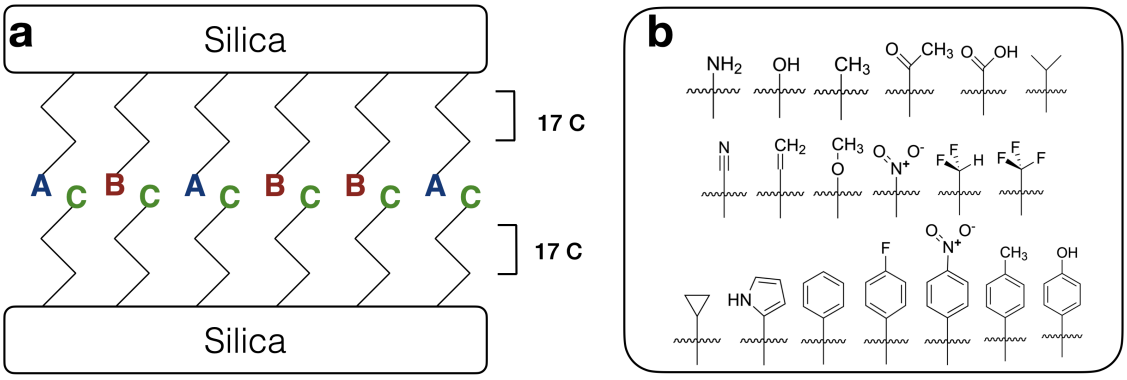

Figure 2.2: (a) Simplified schematic of the systems studied. The top monolayer is a mixture of two types of terminal groups chemistries (A and B), studied at two different mixing ratios (0.25:0.75, 0.5:0.5), while the bottom monolayer is homogeneous (chemistry C). (b) Depiction of the 19 different chemistries considered. From top to bottom, left to right, the terminal groups are amino, hydroxyl, methyl, acetyl, carboxyl, isopropyl, cyano, ethylene, methoxy, nitro, difluoromethyl, perfluoromethyl, cyclopropyl, pyrrole, phenyl, fluorophenyl, nitrophenyl, toluene, phenol.

itoring the simulation progress of the project as a whole. These obstacles are overcome by the utilization of several open-source software packages including the MoSDeF Framework[19] and the Signac Framework.[20,49] The use of the MoSDeF framework allows for the encapsulation of the system initialization step, including the construction of chemical systems based on provided variables and perform atomtyping and parameterization all in the same ecosystem. The Signac Framework is used to manage, monitor, and advance the progress of the automated project as a whole. The use of these open-source software ensures that all scripts and input parameters used to initialize the systems, submit the systems for simulation, and analyze the systems are captured and preserved, ensuring the simulations are TRUE (Transparent, Reproducible, Usable by Others, and Extensible).[23] A summary of the screening workflow is visualized in Figure 2.3. All scripts and parameter files are available in the associated GitHub repository (see Appendix A). We also note that a small subset of simulations (less than 1% of the total) failed to complete due to unstable initial configurations. However, in all cases, each unique system composition reported includes at least 3 replicates.

2.3 Calculation of Frictional Properties

The coefficient of friction (COF), μ, and adhesion force, F_0, were calculated from the last 5 ns of the simulation trajectory from each system under shear using the modified version of Amonton's law of friction given by Equation 2.1.

$$F_f = F_0 + \mu \times F_N \tag{2.1}$$

In the equation, F_f, F_0, μ and F_N represent the frictional force, the adhesive force, the coefficient of fric-

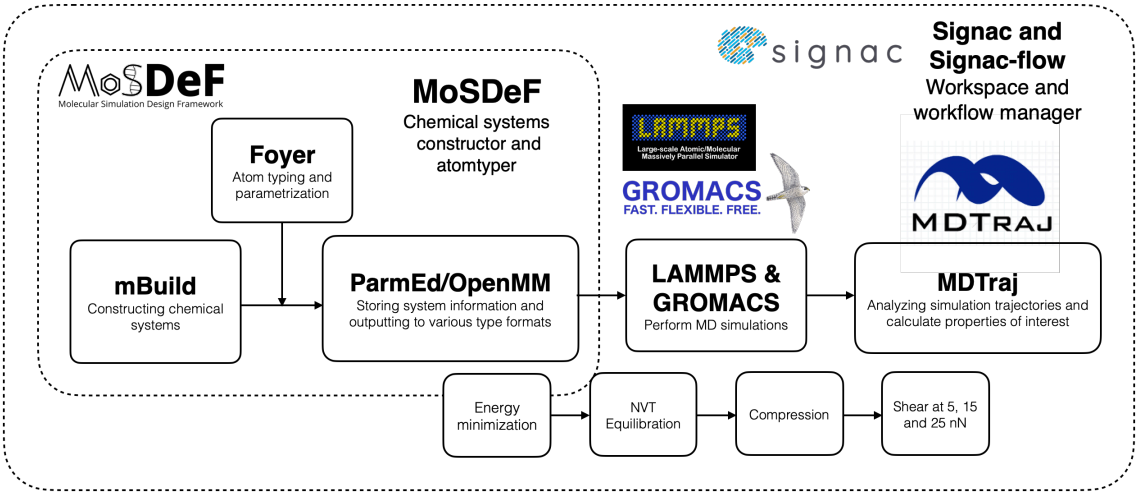

Figure 2.3: High-throughput screening workflow utilizing all open-source software. Specifically, the MoSDeF software suite is utilized to automatically perform system initialization, while GROMACS and LAMMPS are used to perform the molecular dynamics simulation, whose results are analyzed with MDTraj; the workspace and workflow as a whole is managed by signac and signac-flow.

tion, and the normal force, respectively. A linear regression of the average friction force (ordinate) versus normal load (abscissa) were used to calculate the COF from the slope and F_0 from the intercept of the regression line with the ordinate axis. The frictional force is calculated by summing all the forces in the direction of shear on the bottom monolayer every 1 *ps* and averaged over the last 5 *ns* of the simulation (the production regime) using MDTraj.[50]

2.4 Machine Learning Model

2.4.1 Random Forest Regressor Algorithm

Random forest regression is a machine learning algorithm used for regression tasks, that is, predicting a continuous target variable based on input features.[51,52] It is an ensemble method that combines multiple decision trees to make predictions, as opposed to a single decision tree. Once provided a training set of data composed of a set of input parameters and their expected outcomes, the random forest ensemble model will create a series of decision trees, each generated from a sub-sample of the training data. The predictions of each individual tree are combined to make the final prediction, either through averaging or a weighted average. The use of multiple decision trees helps to reduce over-fitting and increase the robustness of the model, as well as improve the accuracy of predictions. Random forest regression is also capable of handling non-linear relationships between input features and the target variable, as well as handling missing values in the data. Random forest regression is widely used in a variety of applications, including stock price prediction, weather forecasting, and medical diagnosis. Hence, this method can be deemed reasonable to be applied in our study

to predict the tribological properties of thin film coatings.

The MD-generated tribological data set is analyzed using the random forest regressor, as implemented in the scikit-learn library,[52] where the input is a composite molecular descriptors, or "fingerprint", of the terminal group combinations, and output being their corresponding tribological properties/lubricating efficacy. This setup is consistent with the prior work by Summers *et al.*[12] Each predictive model will rank the importance of each of the features in the "fingerprint" based on how each input affects the final prediction and unveils information regarding properties that play determinant roles in predicting the tribological properties of the monolayer. These features, in addition to the ability to provide prediction for novel systems, makes random forest regressor advantageous for screening/discovery research. All of the random forest models in this study have 1000 trees, ensuring the predictions converge in a reasonable amount of time.[33] Each decision tree in the forest is allowed to expand until all leaves are pure (choosing splits that decrease impurity defined by the Gini impurity). All models used mean squared error (MAE) as error criterion during training. Each random forest model, and its subsequent decision trees, are trained with 35 features, which are molecular descriptors calculated through RDKit.[53] This setup is consistent with previous study by Summers *et al.*[12] allowing for direct comparison between these studies, focusing on the accuracy of the models and feature importance ranking determined from the two sets of data.

2.4.2 Molecular Fingerprint

In our implementation, the training data for ML algorithm are pairs of inputs, *i.e.,* the "fingerprint" representing each system, and expected outputs, *i.e.,* the COF and F_0. Here, the "fingerprint" of each system is the combined chemical and physical attributes of the component terminal groups calculated using the RDKit cheminformatics library.[53] Chemoinformatics is a branch of computational chemistry that deals with the representation, storage, manipulation, and analysis of chemical and biological data. It involves the use of computer algorithms, data structures, and databases to study chemical and biological systems. In cheminformatics, chemical structures are modeled using mathematical algorithms and represented as numerical data, which makes them ideal input for ML models.

The procedure to generate a "molecular fingerprint" is adapted from our previous work.[12] In short, the "fingerprint" of each system is the weighted average of its component terminal groups' molecular descriptors. Specifically, each individual terminal group can be represented by two SMILES strings[54]: one of a hydrogen capped structure and one of a methyl capped structure. Each SMILES string, in turn, is provided to the RDKit cheminformatics library to determine molecular descriptors characterizing the chemical and physical properties of structure.[53] These descriptors are categorized into four groups: size (*e.g.*, number of heavy atoms), shape (*e.g.*, planarity), complexity (*e.g.*, connectivity), and charge distribution (*e.g.*, topologi-

cal polar surface area). Our previous work has determined that, while shape characteristics can be adequately represented by a hydrogen-capped structure, other features are more accurately described with structures resembling the terminal group when attached to alkylsilane backbone.[12] Hence, we used the hydrogen-capped SMILES string to determine descriptors relating to shape, and used the methyl-capped SMILES string to calculate the remaining descriptors. In total, each terminal group is described by 53 descriptors, further detailed in Appendix D.

From molecular descriptors of individual structures, we calculate the descriptors of the top and bottom monolayers. Descriptors for the mixed monolayer with two terminal groups, *i.e.*, the top monolayer, are the weighted average by the relative composition of its component terminal groups' descriptors. Descriptors for the bottom monolayer are the descriptors of its sole terminal group. We note this representation does not contain all of the information regarding connectivity of constituent chains and distribution pattern, and may not fully represent our dual-monolayer system.[55] However, since we are mainly interested in the interactions of different terminal group chemistries in the inter-monolayer regions/interfaces, our composite molecular descriptor "fingerprint" representation was found to be sufficient. This representation only encodes minimal information about the region of interest, with the assumption that the two terminal groups in the top monolayer are evenly distributed. Next, the descriptors of the top and bottom monolayer are combined, retaining only the average and minimum of each descriptor as the "molecular fingerprint" of the entire system. With this approach, each system will be represented by a total 106 descriptors ([53 metrics]*[2 corresponding to min and mean]), which have been shown to sufficiently summarize the most important features of these systems.[12]

Finally, these "molecular fingerprints" undergo a dimensionality reduction step developed by Summers *et al.* using the source code hosted in the accompanied GitHub repository.[12] In this step, descriptors whose values are at least 90% correlated will be reduced to only one attribute, and descriptors with variance values below 2% are also removed. This step reduced the number of descriptors of each system to be 34, which are then used as the input parameters to the ML models. The process of determining molecular "fingerprint" of each system is summarized in Figure 2.4.

2.4.3 Evaluation of the Model

From the entire available *in silico* data set of 9772 data points generated from the MD high-throughput screening, which includes data generated from Summers *et al.*[12] and excludes systems whose simulation failed to complete, 20% (1956 data points) is set aside for testing purpose while the rest (7816 data points) is utilized to train the ML models. This allows for an independent set of systems that the predictive models have not seen/been optimized for, and provide more reliable information regarding the accuracy and bias of the ML models. The accuracy of the predictive model can be determined by comparing the properties obtained from

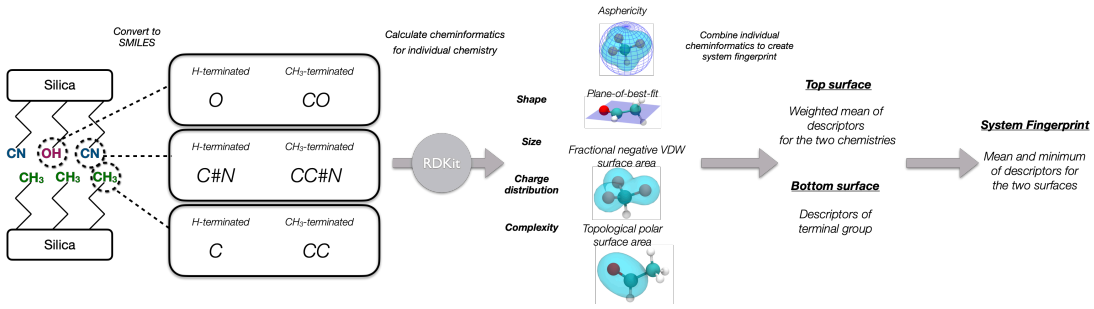

Figure 2.4: Process of generating the molecular descriptors (fingerprints) of the dual monolayer systems. Component terminal group chemistries of each top and bottom monolayer are represented by an H-terminated and methyl (CH3)-terminated SMILES string, which can be used by RDKit to calculate corresponding molecular descriptors. For this study, we consider a total of 53 descriptors (listed in Appendix D), which can be grouped into 4 categories, namely, shape, size, charge distribution, and complexity. The weighted averages of these descriptors are then calculated to represent their corresponding surface, which in turn, will be used to determine the fingerprint of each system. Figure adapted from Summers *et al.*[12]

the simulations, *i.e.*, our ground truth, to the values estimated by the ML model, for all systems in the test set. The comparison can be visualized as Predicted vs Simulated plots (see Figure 4.1). From this comparison, the coefficient of determination (R^2), and mean absolute percentage error (MAPE) are used quantify the accuracy of the ML models. The R^2 describes the correlation between the simulated and predicted values, and MAPE provides the scaled error metrics.[56]

2.5 Integration of ML with High-Throughput Screening

A potential application to assist the high-throughput screening process would be using the ML model as a guide to focus on only simulating a subset of systems in the parameter space that best supports the goal of the study. This allow researchers to only focus on designs that would yield desirable properties, instead of having to survey and sort the parameter landscape in its entirety. One way to do so would be to integrate a ML technique early on in the screening process. In our application specifically, MD can be used to generate a small set of data to train baseline ML models, which can then be utilized to quickly evaluate possible candidates and determine the next set of systems to be simulated. The results from such baseline models, can be utilized for various efforts, such as focusing on only simulating systems with the most favorable properties, which will improve the accuracy of the predictive models only in a certain region of the parameter space, or simulating those that could improve the overall robustness of the subsequent models.

To test this hypothesis, we begin by training a baseline ML model initialized with only 100 data points randomly sampled from the whole data set. The ML model can then be utilized to provide estimations on the remainder of the data set to determine the next set of systems to be considered. From the predictions, we will "simulate" only the top 100 systems (ranking either the COF or F_0) and append them to the previous

training set to create the next set of predictive models. The ML-MD high-throughput screening method can then be carried out in an iterative manner, depending on the complexity of the systems and the availability of computational resources. For our demonstration, 25 iterations were conducted, generating approximately 2500 systems, or a quarter of the entire proposed terminal group combinations. The process were repeated three times, differed only by the initial 100 data points used to train the baseline model, for statistical purpose. The success of this approach can be measured by the number of systems ”simulated” and compare their COF or F_0 to how many of the top 500 systems were generated through each iteration.

CHAPTER 3

Results and Discussion: High-throughput Screening

3.1 Results Overview

With the amount of data generated through the high-throughput screening process, there are many ways we can view and analyze the data. Considering first the results of the high throughput screening MD simulations, including those performed in the current study and in Summers *et al.*,[12] we plot every data point simulated via MD in Figure 3.1 and Figure 3.2.

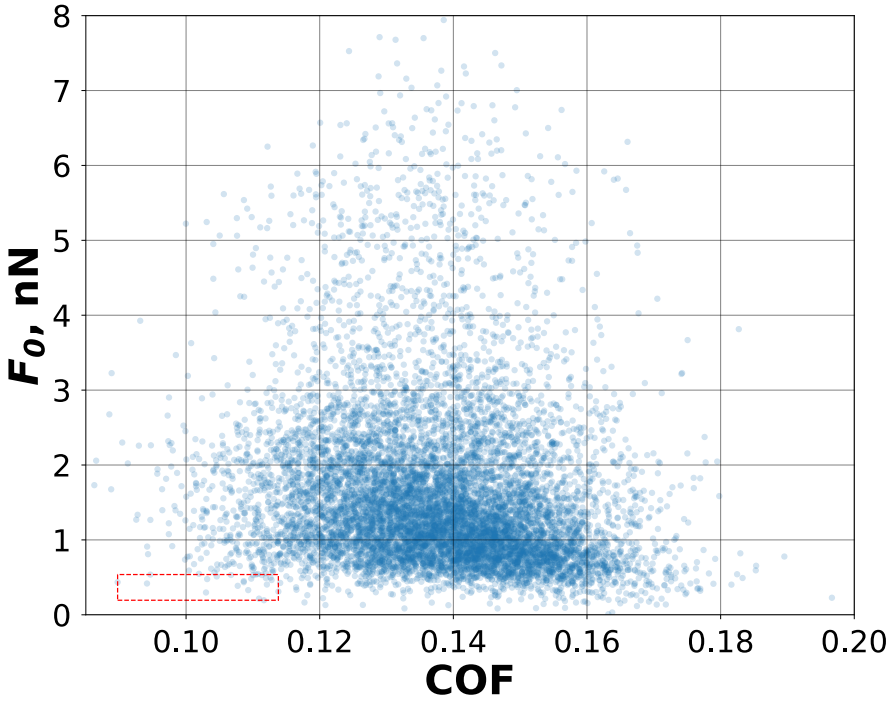

Figure 3.1: Distribution of simulated systems based on their COF and F_0 values. The 22 most-favorable systems, listed in Table 3.1, corresponds to data points confined within the red dashed box in the lower left quadrant of the figure.

We first note that the range of COF is narrower than that of F_0 and also possess different distributions. For COF, we observe a normal distribution, with values ranging from 0.074 to 0.1967, with the mean of 0.1377 and the standard distribution of 0.0143. While, for the F_0, the distribution appeared positively skewed, *i.e.*, more data points are of lower values than higher values, with reported values ranging from 0.007 *nN* to 7.942 *nN*. We note the skewed data explains the vast difference between the mean and the median value of F_0, being 1.644 *nN* and 1.328 *nN* respectively, with the standard distribution of 1.128 *nN*. The skewed data of the F_0 suggests that this value could be affected by property, which is related to a certain type of chemistry. Indeed,

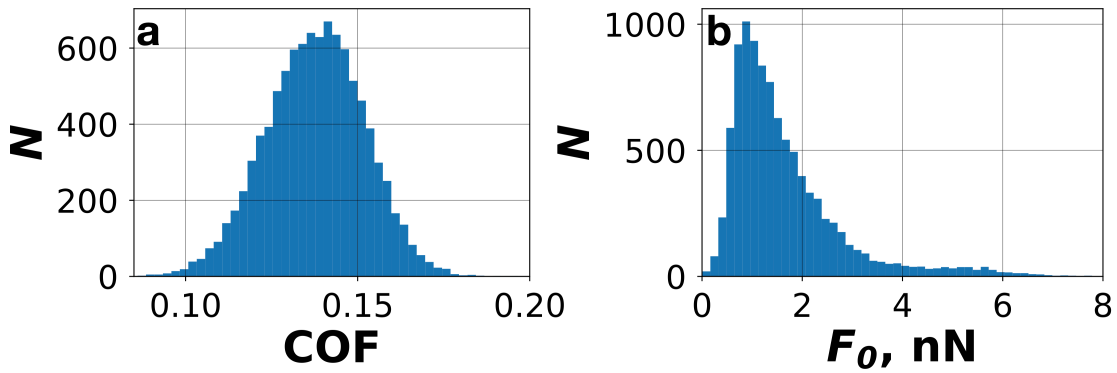

Figure 3.2: Distribution of (a) COF and (b) F_0 for systems considered in this study, obtained from MD simulations

this phenomenon could be explained by a conclusion drawn from the study by Summers *et al.*,[12] determining that the F_0 could be significantly elevated by the inter-monolayer hydrogen bonding capability of the system, which is only observed in certain terminal groups combinations.

From the data set, we also identify 22 monolayer designs that provide favorable frictional properties, *e.g.*, those that have low simulated COF and F_0 values (see Table 3.1 and Figure 3.1). The table was generated by the intersection of the 500 systems with lowest COF (values ranging from 0.074 to 0.114) with the 500 with lowest F_0 (values ranging from 0.007 *nN* to 0.541 *nN*). We first note that, in general, these results agree with observations made by Summers *et al.*, where it was noted that the COF of monolayers is mainly affected by the shape and size of the terminal group, with chemistries of small sizes and simple shapes (*e.g.*, *sp* hybridization) yielding the lowest COF.[12] While, the F_0 is most strongly affected by charge distribution, with polarity and hydrogen bond forming ability both elevating F_0.[12]

In agreement with these findings, we observe that a majority of the systems identified (19 out of 22) consists of a cyano homogeneous monolayer. The cyano group is small in size, has *sp* hybridization and does not readily form hydrogen bonds. These are characteristics that agree with previous work to identify chemistries that can lower the COF and F_0 of monolayers. We also note most systems in Table 3.1 are made up of 3 different components and only one system that consists of two homogeneous monolayers (first system in Table 3.1), which was simulated in the Summers *et al.* work.[12] This result suggests a slight advantage to having mixed monolayer designs. However, we also recognize that the data set is dominated with mixed monolayers compared to homogeneous monolayers, therefore the best performing systems are likely the result of the much larger representation of mixed monolayer systems compared to the homogeneous systems. Nonetheless, mixed monolayer systems could provide extra flexibility during the design process and allow for the optimization of other properties, such as thermal stability or environmental interactions, depending on

the specific application, giving these designs advantages over homogeneous monolayers.

Table 3.1: 22 most-favorable systems as determined by the intersection of the top 500 systems ranked by coefficient of friction (COF) and the top 500 systems ranked by adhesive force (F_0). The COF and F_0 mean values and standard deviation (std.) are calculated from the three replicate simulations.

	Terminal Group A	**Terminal Group B**	**Terminal Group C**	**A Fraction**	**B Fraction**	**COF - mean**	**COF - std**	**F_0 (nN) – mean**	**F_0 (nN) std.**
1	cyano	cyano	isopropyl	0.5	0.5	0.1032	0.0063	0.4768	0.1075
2	cyclopropyl	ethylene	cyano	0.5	0.5	0.1086	0.0142	0.4498	0.282
3	difluoromethyl	isopropyl	cyano	0.5	0.5	0.11	0.0098	0.5292	0.4261
4	difluoromethyl	methyl	cyano	0.5	0.5	0.1128	0.0036	0.464	0.1585
5	ethylene	isopropyl	cyano	0.5	0.5	0.1109	0.0108	0.2161	0.3768
6	ethylene	perfluoromethyl	cyano	0.5	0.5	0.1093	0.0119	0.3297	0.3124
7	isopropyl	methoxy	cyano	0.5	0.5	0.1113	0.0144	0.5313	0.5019
8	isopropyl	methyl	cyano	0.5	0.5	0.1121	0.0182	0.4041	0.3252
9	isopropyl	perfluoromethyl	cyano	0.5	0.5	0.103	0.0181	0.2978	0.196
10	difluoromethyl	carboxyl	isopropyl	0.25	0.75	0.1117	0.0021	0.1944	0.8319
11	difluoromethyl	isopropyl	carboxyl	0.25	0.75	0.1105	0.0078	0.5375	0.7366
12	difluoromethyl	methyl	cyano	0.25	0.75	0.1126	0.0041	0.4967	0.0876
13	ethylene	isopropyl	cyano	0.25	0.75	0.1046	0.0184	0.5122	0.1374
14	isopropyl	cyclopropyl	cyano	0.25	0.75	0.0898	0.0119	0.4354	0.1522
15	isopropyl	perfluoromethyl	cyano	0.25	0.75	0.1077	0.0073	0.3868	0.158
16	methoxy	cyano	ethylene	0.25	0.75	0.1086	0.0053	0.4346	0.434
17	methyl	isopropyl	cyano	0.25	0.75	0.0942	0.0148	0.4181	0.2017
18	perfluoromethyl	cyclopropyl	cyano	0.25	0.75	0.1138	0.007	0.3104	0.1799
19	perfluoromethyl	ethylene	cyano	0.25	0.75	0.1067	0.0002	0.5315	0.1965
20	perfluoromethyl	isopropyl	cyano	0.25	0.75	0.0947	0.0114	0.5366	0.2037
21	phenyl	isopropyl	cyano	0.25	0.75	0.1098	0.0156	0.4825	0.3196
22	toluene	ethylene	cyano	0.25	0.75	0.1119	0.0134	0.4907	0.6589

3.2 Notable Homogeneous Monolayer Terminal Group Chemistries

Based on the results above, we can take a more detailed look into the top performing systems; specifically, systems involving the cyano and isopropyl terminal groups that consistently appear with the highest frequency in Table 3.1. Subsequently, we will look at systems whose bottom monolayer, *i.e.,* the uniformly homogeneous monolayer, is terminated by either cyano or isopropyl groups, which is represented in the schematic as Terminal Group C in Figure 2.2 a. The tribological properties of these systems are then displayed in a heatmap (see Figure 3.3 and Figure 3.4), where the y and x axis represent the terminal groups involved in the top monolayer, *i.e.,* the mixed chemistry monolayer, matching with group A and B in Figure 2.2 a. The relative composition of groups A and B is also noted as either 0.5:0.5 or 0.25:0.75 in each corresponding figure. The magnitude of their properties, COF and F_0, is reflected in the color saturation of each cell. The remaining heatmaps of other terminal groups can be found in Appendix C. All heatmaps adopt the same color saturation range, determined by the COF/F_0 range of the entire data set. We note a few data points were not available due to failure to meet required conditions and appeared as white. This means it could be due to failure at the simulation step and appeared as white/empty cells or that a particular combination selected during the high-throuput screening set up failed to meet the criteria that chemistry A must be different than chemistry B (thus creating the white diagonal line in Figure 3.3 and Figure 3.4). These systems accounted for less than 1% of the entire systems considered, and hence does not affect our final conclusion.

3.2.1 Cyano Terminated Monolayer

Figure 3.3 represents the heatmap of systems containing a cyano terminated homogeneous monolayer (bottom monolayer) in addition to a heterogeneous top monolayer with two different chemistry combinations in either a 0.25:0.75 or 0.5:0.5 ratio. These systems have COF values ranging from 0.074 to 0.150, with the mean of 0.115 and standard deviation of 0.011. The mean COF value is 16.67% lower than the average COF of all systems surveyed, 0.1377 as reported in section 3.1, indicating that the cyano homogeneous monolayer proved to minimize the COF of the whole system. F_0, on the other hand, has a broader range, with the minimum and maximum of 0.0848 *nN* and 5.890 *nN* respectively, and the mean of 1.908 *nN* and a standard deviation of 1.180 *nN*. This average F_0 value is slightly greater than the average value of the entire data set shown in section 3.1, which is 1.644 *nN*. In other words, the result indicates that using the cyano terminating groups in an alkylsilane monolayer can lower COF during shear, regardless of the opposing surface. However, since this group has the tendency to form strong polar interactions and has the capability to form hydrogen bonds with certain neighbors, it can elevate the F_0 between surfaces. Therefore, one should be cautious when designing these thin film combinations in order achieve desirable properties for their applications.

3.2.2 Isopropyl Terminated Monolayer

Another terminal group that appears at high frequency from Table 3.1 is isopropyl terminal group chemistries. Thus, we take a detailed look into this group shown in Figure 3.4. Out of all the systems with isopropyl terminated groups, the monolayers have COF values ranging from 0.102 to 0.175 with the mean of 0.137 and standard deviation of 0.011; the F_0 ranges from 0.0887 nN to 1.537 nN, with the mean and standard deviation of 0.797 nN and 0.635 nN, respectively. At a glance, these systems have comparable COF values with the mean COF values for the entire data set shown in section 3.1, while providing significantly lower F_0, on average. In summary, contrasting with the cyano group, the isopropyl group does not offer any additional benefit for decreasing the COF. However, since this terminal group is made up of only carbons and hydrogens, thus lacking the capability to form neither hydrogen bonding networks nor strong polar-polar interactions, isopropyl terminated monolayer can help lower the F_0 of surfaces during shear. These results suggest that together the combination of cynao and isopropyl groups can create an optimal thin film coating, achieving both low COF and low F_0.

3.3 Effect of Different Mixing Ratio

From the available data, we also can compare the effect of changing the mixing ratio. Examining the heatmaps in Figure 3.3 and Figure 3.4 alone, we see that increases or decreases in the composition of certain terminal groups have an effect on the lubricating properties of the thin film system. These trends can be visualized by the increase or decrease in color saturation in the rows or columns when comparing the heatmaps of systems that have 0.5:0.5 mixing ratio to those that have 0.25:0.75 mixing ratio in the top monolayer. The increase in saturation of a column means that increasing the relative composition of that specific chemistry will elevate the property, while increase in saturation of row means that decreasing composition of the chemistry will lower the property. Since this logic is inversely related, we will only compare the change in saturation pattern of columns but not rows.

Comparing Figure 3.3 a and c shows that for the cyano groups, increasing the hydroxyl, perfluoromethyl, tolune and methoxy groups' relative composition from Figure 3.3 a to c in the top monolayer, cause the COF to become elevated; while increasing the cyano terminal group B from 0.5 to 0.75 relative composition actually lowers their COF. Regarding the F_0, comparing between the ratios in Figure 3.3 b and d, increasing the carboxyl or phenol group terminal group B composition causes an increase in F_0 values. However, these effects are slightly different in the homogeneous isopropyl monolayer systems shown in Figure 3.4. Comparing between the COF values in Figure 3.4 a and c, we see that the mixing ratio effects caused a spike in the COF for slightly different sets of terminal groups, namely perfluoromethyl, difluoromethyl, and toluene. Similar to Figure 3.3, the cyano terminal group B in Figure 3.4 decreases the system's COF. However

as a contrast to the cyano homogeneous monolayer, there is no noticeable difference between the F_0 of the film composition (see Figure 3.4 b and d). This inertness can be explained by the low polarity as well as lack of strong hydrogen bonding capability of the isopropyl monolayer.

We note that, besides the terminal groups mentioned above that induce a change in patterns, there is no consistent trends created by changes in film composition. This suggests that there is not a silver bullet combination that lowers both the COF and the F_0 currently. It is a complex balance of interactions involving intra-monlayer and/or inter-monolayer terminal groups interacting with one another affecting their final lubricating ability. The trends observed so far still conform to standard chemical intuition.

3.4 Conclusion

From the MD high-throughput screening process, we have been able to determine several promising terminal group chemistries that can be further studied, such as cyano and isopropyl groups. The MD results also unveil important data regarding the strategy to design thin films to achieve optimal lubricating properties producing low COF and F_0 values. The impact of altering the mixing ratio, however, does not provide clear trends, and one must still rely on chemical intuition to estimate the lubricating ability of different combinations. However, introducing varying mixing ratios significantly expanded the number of potential thin film designs. The data from the screening workflow can also serve as *in silico* data to new ML models, which will be further discussed in chapter 4, and provide predictions for systems with chemistries and compositions beyond those surveyed here.

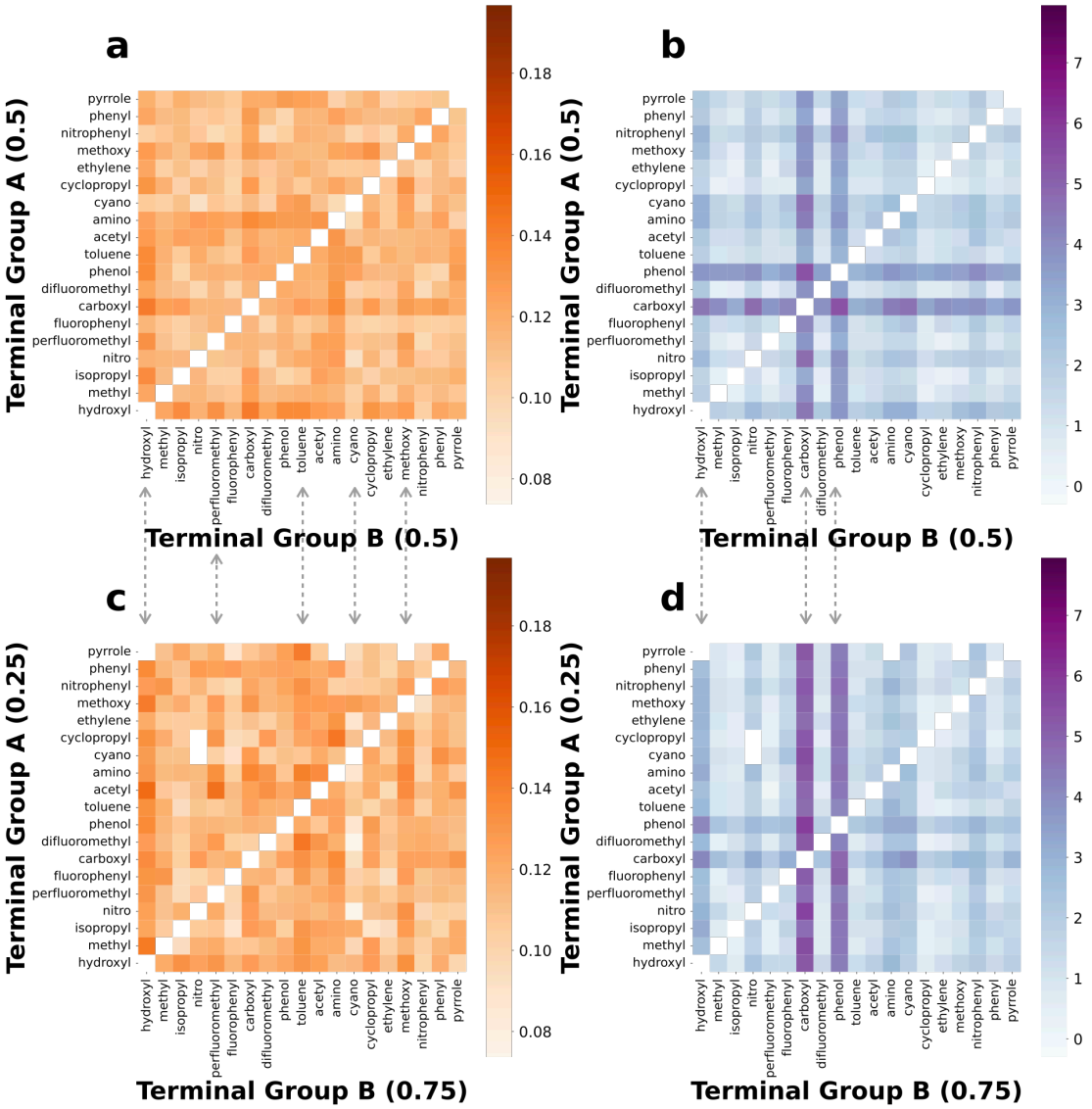

Figure 3.3: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only cyano terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal group (group A and B in Figure 2.2 a). Their relative composition is annotated in each individual figure. The dotted lines between figures (a)-(c) and (b)-(d) highlight groups whose increase in relative composition have a visible effect on the tribological properties of the system.

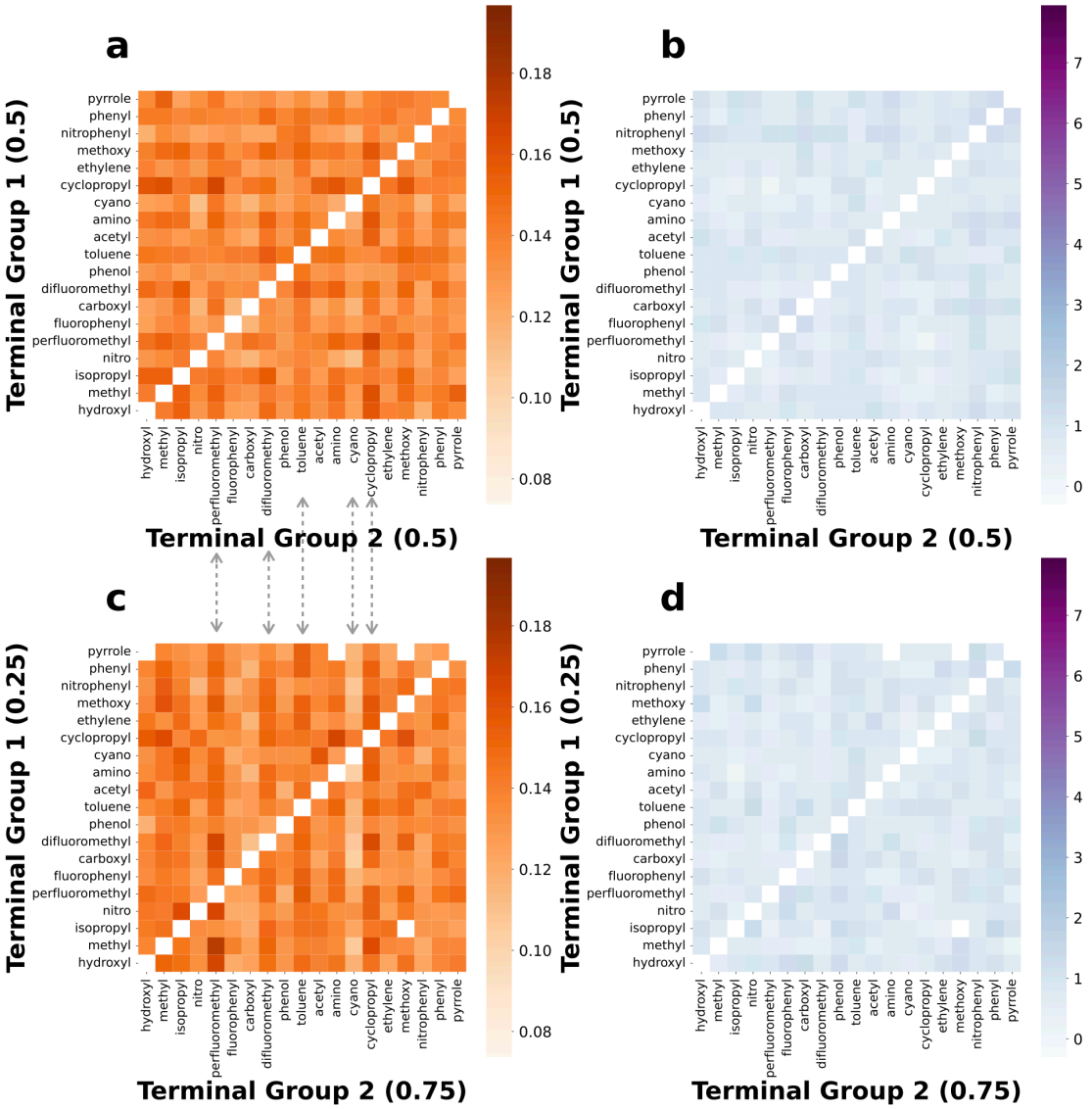

Figure 3.4: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only isopropyl terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal group (group A and B in Figure 2.2 a). Their relative composition is annotated in each individual figure. The dotted lines between figures (a)-(c) and (b)-(d) highlight groups whose increase in relative composition have a visible effect the tribological properties of the systems.

CHAPTER 4

Results and Discussion: Machine Learning

4.1 Comparing The Accuracy of ML Predictive Models

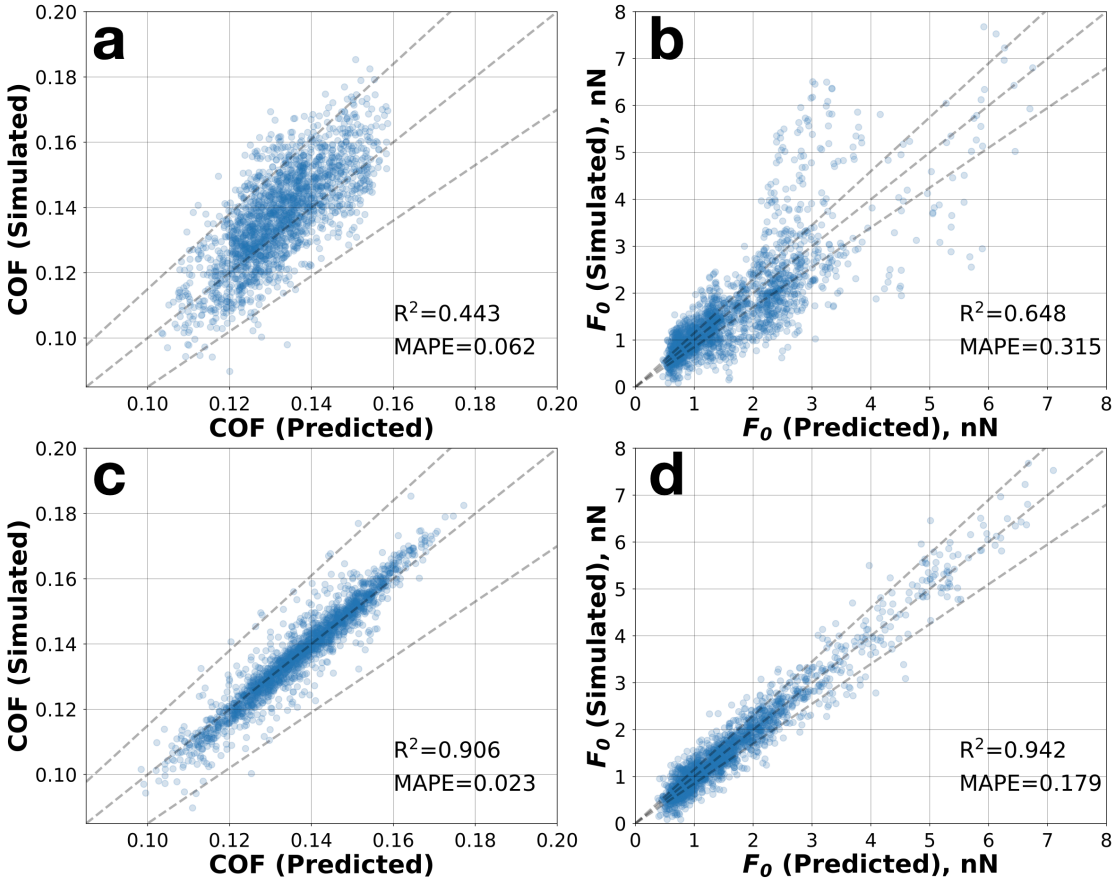

Figure 4.1: Predicted-versus-simulated plots for COF and F_0 for models trained with 100 simulation data points for uniform monolayers from Summers *et al.*[12] data set (a and b) and trained with 7816 data points as described in subsection 2.4.3 (c and d). The dotted line in the middle represents perfect prediction ($y = x$). The outer two lines represents the 15% variation from a perfect prediction ($y = 1.15x$ and $y = 0.85x$). The coefficient of determination (R^2) and mean absolute percentage error (MAPE) are included.

To begin, we train the first set of predictive models with the data generated with MD through the high-throughput screening process reported in chapter 3. The models are then applied to predicted tribological properties of systems in a test set, as described in chapter 2. The predicted results are then compared to the COF and F_0 results obtained directly from the MD simulations to determine the accuracy of the ML models (see Figure 4.1). Results are also included for the ML models trained in Summers *et al.*[12] with the same test set to show a comparison between the two ML models. When applied to the same testing set of nearly 2000

data points, the Summers *et al.* models provide R^2 values of 0.472 and 0.657 for COF and F_0, respectively, compared to 0.906 and 0.942 by the models trained with data generated in this study. While the R^2 values are considerably lower for the Summers *et al.* ML models, it is worth noting that the training data set used in their ML model did not include any information regarding mixed monolayer compositions; as such, the Summers *et al.* ML models still demonstrate adequate efficacy. This point is further demonstrated by their MAPE, where the Summers *et al.* models could predict the COF of a system with 0.056 (5.6%) error and predict F_0 with 0.266 (26.6%) error. These MAPE values are higher in comparison to those produced by the new set of models, but are justified by the significant size difference of the training data for the ML models (nearly 80 times different). We note the prediction of F_0 is less accurate in the higher adhesion regime for the Summers *et al.* ML models, but the prediction accuracy has been significantly improved in the newer ML models. This is likely related to the skewed distribution of F_0 values (see Figure 3.2 b), which make it more challenging to study systems in the higher F_0 regime for training, especially for smaller scale studies like that in Summers *et al.*.[12]

Nonetheless, this result suggests that ML models trained with limited data could still provide meaningful estimations, and that the use of the random forest regressor may lead to models that are predictive for chemistries and compositions outside of the training set. We note that, this relationship has not been tested with other forms of study, *e.g.*, different chemical systems and simulations, and would thus require further evaluation to draw a stronger and more generalized conclusion about ML models. We also note that for lower values of COF or F_0, both models deviate slightly in the positive direction, meaning they predict a slightly higher value compared to simulation; as the value of either COF or F_0 increases, a negative deviation is observed with the ML models predicting slightly better performance than is observed in the MD simulations. This trend is more apparent when looking at the prediction results from the Summers *et al.* models (see Figure 4.1 a and b). This skew in the predictions suggests that for favorable tribological conditions (*i.e.*, low COF and low F_0), the model will tend to overestimate the values, thus reducing the likelihood of incorrectly identifying poor performing films as viable options. Given that this behavior of the model minimizes the chances of exaggerating the performance of high performing systems (*i.e.,* those with low COF and F_0), this suggests the predictive ML models can be confidently used to screen over potential film candidates for possible applications. We also note that while the R^2 values for COF models are lower than those of F_0 models, giving an impression that the latter models outperform their COF counterparts. However, their MAPE values indicate the opposite; the F_0 models exhibit significantly greater percentage errors than the COF. This disparity is attributed to the difference in the range of these two properties; while COF values span a small range of values from roughly 0.08 to 0.2, F_0 can take values from $0\ nN$ to $8\ nN$ (see Figure 3.2 and Figure 3.1), which may affect how these metrics are calculated. Hence, it is important to recognize that using either R^2

or MAPE values to directly compare the predictive ability of the COF and F_0 models may be challenging and potentially misleading, though these metrics are useful to compare the performance of ML models of a similar type.

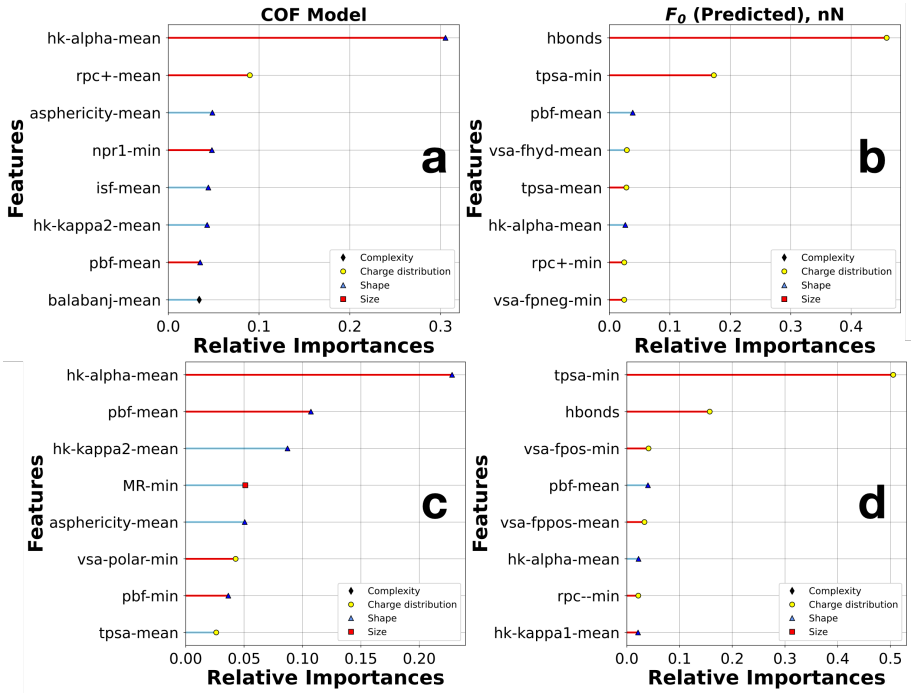

Figure 4.2: Feature importance of Summers *et al.,* model for (a) COF and (b) F_0 in comparison with those of ML models trained with data generated from this study for (c) COF and (d) F_0

Feature importance score in the random forest algorithm is calculated by measuring the impurity of the nodes of each decision tree that use a specific feature. The more impurity decreases, the more important the feature is considered.[51] The feature importance scores are normalized to add up to 1, and can be used to identify the most important features that have been utilized by the random forest predictive model. For our study, this means ranking the most important chemical/physical properties, described as cheminformatics, that have the most significant effects on the lubricating properties of the thin film design. Here, we can compare the feature importance of two models, one that trained with 100 data points from Summers *et al.,* and the other with nearly 8000 data points generated during the course of this study. This analysis can tell us if there is any major shift in trends, that is, only captured when sufficiently large enough data sets are provided. Comparing between COF models (Figure 4.2 a and c), we can see that the highest ranking feature, the hk-alpha, remains consistent, leaving the other feature importances order shuffled around. The adjustments in the feature importance ranking reflect the changes that occur as the forest decisions find descriptors that better classify the data, resulting in more accurate models. Between F_0 models (see Figure 4.2 c and d), we note the position of the two highest ranking features, hbonds and tpsa-min, remain unchanged. However, there is

some rearrangement among the lower ranking features. Despite the conclusions drawn from these two sets of models, the work from Summers *et al.* and the newly trained models, remain consistent for both COF and F_0. Specifically, the COF is mostly impacted by the shape and size of the terminal group chemistries. Signifying those with smaller size and linear organization, *i.e.*, having *sp* hybridization, yields the lowest values. While, F_0 is strongly impacted by the terminal group charge distribution and ability to form hydrogen bonds.

4.2 Screening of A Small Molecule Library

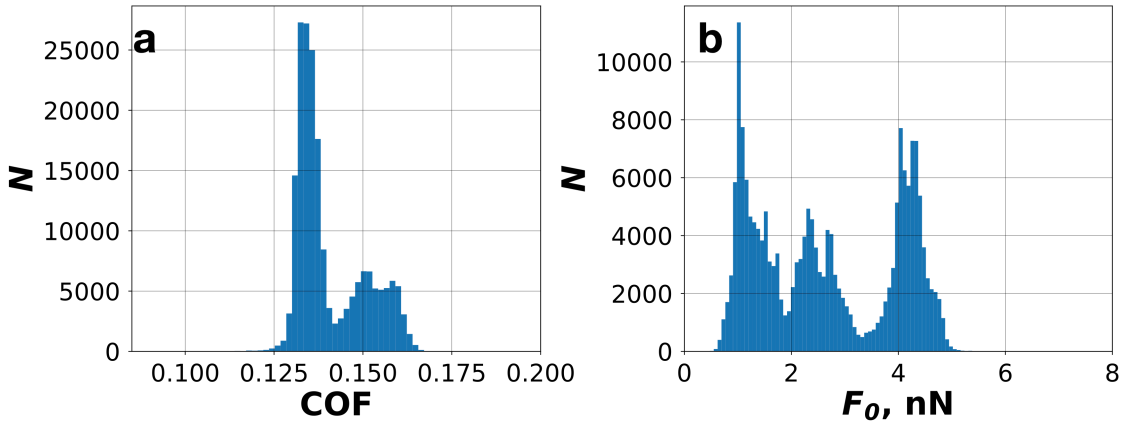

Figure 4.3: Distribution of (a) COF and (b) F_0 predicted by the ML models for 193,131 unique systems created with molecules from ChEMBL small molecules library.

As a proof-of-concept of using ML to pre-screen the design space, we perform a screening study using the above ML models. The chemical space for this screening was constructed using small molecule, whose molecular weight is between 4 and 99 amu, from the ChEMBL library.[34,35] This list of 981 small molecules undergo further filtering to remove irrelevant group, such as those containing metallic elements and those that cannot be processed by the RDKit library, *e.g.*, chiral or charged molecules, resulting in 621 unique chemistries. With these 621 groups, we created 193,131 unique systems in which each made up of two homogeneous monolayer (*i.e.*, containing only one species); mixed monolayer chemistries were not considered here due to the shear volume of data that would be generated. This simpler approach of using dual homogeneous monolayers was chosen to allow for consideration of more unique chemistries. Molecular descriptors of the 621 terminal groups were identified using their SMILES strings and the RDKit library[53]; these descriptors were then used to construct the "fingerprint" for each of the 193,131 unique dual-monolayer system (as described in subsection 2.4.2). The "fingerprint" was used as input to the ML models, which in turn, provided estimation of tribological properties of corresponding system. This screening process evaluated 385,641 systems total since duplicate systems (*i.e.*, systems whose top and bottom monolayers inverse of each other) were not removed. This screening took approximately 24 hours to predict on a standard laptop, or roughly

0.22 seconds per system, which is about 5 orders of magnitudes faster than the time required to perform a single MD simulation and without the need for expansive computational resources like a computing cluster.

Table 4.1: 20 best performing systems as determined by the intersection of the top 2000 systems ranked by coefficient of friction (COF) and the top 2000 systems ranked by adhesive force (F_0). The properties were predicted using the ML models trained with 7816 data points, as described in section 4.1.

	Terminal Group A	**Terminal Group B**	**COF**	F_0 (nN)
1	cyano	propyl	0.1144	0.7257
2	cyano	cyclopropyl	0.1151	0.4631
3	methyl	cyano	0.1153	0.5532
4	acetylene	1,1-difluoroethyl	0.118	0.7699
5	cyano	ethyl	0.1206	0.649
6	fulminic acid	cyclopropyl	0.1236	0.7117
7	ethylene	1,1-difluoroethyl	0.1244	0.7341
8	bromoethyl	1,2-diformylhydrazine	0.125	0.7695
9	methyl	fulminic acid	0.126	0.7704
10	cyano	difluoroethyl	0.1265	0.7279
11	bromoethyl	malononitrile	0.1269	0.7098
12	acetylene	ethyl	0.127	0.7254
13	1,1-difluoroethane	propene	0.1271	0.729
14	propyl	2,2-difluoroacetamide	0.128	0.7423
15	acetylene	propyl	0.1281	0.7777
16	methyl	acetylene	0.1281	0.7405
17	bromoethyl	1,2-dicyanoethyl	0.1282	0.7737
18	fulminic acid	ethyl	0.1283	0.7725
19	cyclopropyl	acrylonitrile	0.1283	0.7152
20	allyl	but-2-yne	0.1283	0.7546

The distribution of COF and F_0 of the systems estimated by the ML model are shown in Figure 4.3. We note the distributions differ from that of the data set screened using MD simulations (see Figure 3.2), which may be explained by expanded chemical design space. Using the first quartile of the COF distribution (0.1280) and F_0 distribution (0.8966 nN) obtained from MD high-throughput screening as a reference, the new data set has 5121 systems that can be considered to have good COF and 10,598 systems with good F_0. Two shortened lists of 2000 systems with lowest COF and 2000 systems with lowest F_0 are compiled. We were then able to reduce to the 20 most interesting systems by intersecting these two lists, and reported them in Table 4.1. We note that many of the chemistries that were surveyed in our MD screening studies (see Table 3.1) are found here; specifically, systems 2, 3, and 16 overlap with Summer *et al.*.[12] The result also suggests several other chemistries to examine in future studies, such as various alkenes (allyl, propene), alkynes (acetylene, but-2-yne), halocarbons (1,1-difluoroethyl, bromoethyl, vinyl chloride), and nitriles (cyano, malononitrile, acrylonitrile). We note that none of the systems reported here (in Table 4.1) provide better lubricating properties comparing to those identified in Table 3.1 through the MD high-throughput screening. This is potentially because the systems in Table 4.1 consist of 2 homogeneous monolayers, and hence, do not reap the benefits offered by the mixed monolayers, as discussed earlier. Nonetheless, this highlights the feasibility of combining ML with MD database screening to reduce computational cost and identify favorable candidates for further study via reducing the vast design space of mixed monolayer systems.

4.3 Integration of ML to Accelerate High-throughput Screening

As previously discussed, the capability of current high-throughput screening with MD can be highly restricted by available computing resources and can become impractical as the parameter space is expanded. Given the significant speed up that the machine learning models provide, we believe the further integration of ML techniques can help high-throughput screening processes be conducted more efficiently. Hence, with the available data, we experiment with a hypothetical scenario where the ML technique is integrated during an earlier stage of the high-throughput screening process. This encourages only those systems that have high potential of producing favorable lubricating properties to be examined. Specifically, we start by training a model with only 100 data points, randomly sampled from the available data set, and utilized the said predictive models to estimate tribological properties of remaining systems in the pre-designed parameter space. From the estimation, the top 100 systems with the most favorable tribological properties, *i.e.,* those with lowest COF and lowest F_0, will be appended to the previous data set, which in turn, is used to train subsequent ML models. This process can be done in an iterative fashion until a computing resource cutoff is reached.

Here, we perform a hypothetical 25 iterations as our cutoff, with an increment of 100 data points, creating 2,500 systems to be "simulated". The outcome is plotted in Figure 4.4, which quantifies the top 500 systems

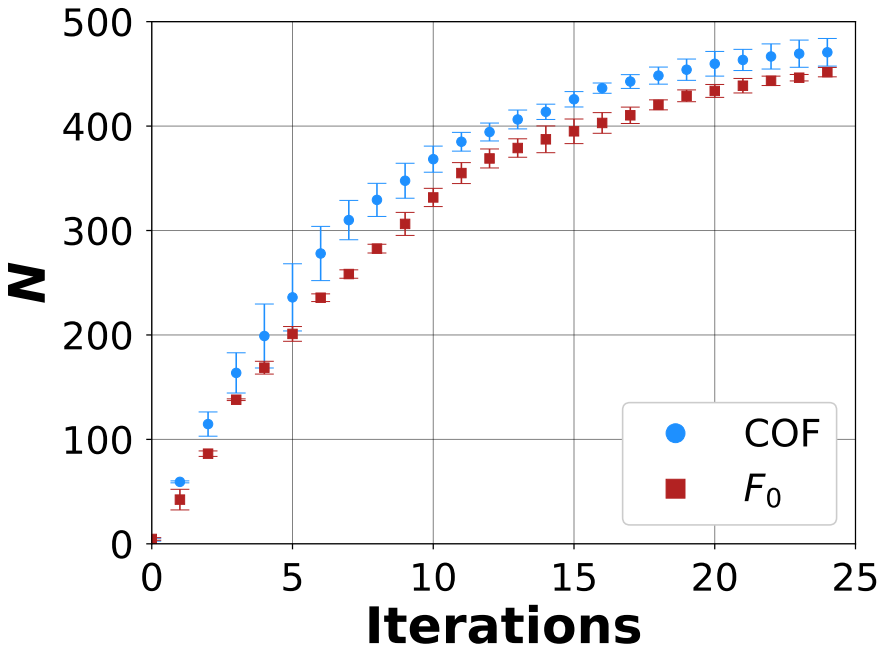

Figure 4.4: The amount of systems in the top 500 systems, ranked by either their simulated COF properties or F_0 properties, that is captured by a hypothetical iterative integration of ML model to the high-throughput screening process.

of the pre-designed parameter space that would have been "simulated" during this hypothetical process. The presented data is averaged from 3 trials, each differing by the initial 100 randomly seeded training data. So for example, Figure 4.4 shows that after 5 iterations, we were able to "simulate" 200 out of the 500 best systems in the entire workspace. Therefore, as the number of iterations increases, we approach and accumulate the top 500 systems. The COF is 406 ± 9 systems from the top 500 systems, or $81 \pm 2\%$ at the 13^{th} iteration and 471 ± 13, or $94 \pm 3\%$ at the 25^{th} iteration. For F_0, we observe a nearly identical trend, though slightly less impressive, with this approach being able to include 403 ± 10 systems, or $89 \pm 2\%$ at the 16^{th} iteration and 452 ± 5 systems, or $90 \pm 1\%$ at the final iteration.

In other words, this approach allows us to determine more than 80% of the most interesting systems by only simulating 1300-1600 systems, or 13-16% of the intended parameter landscape. We can achieve more than 90% of the best systems if 2500 simulations, or roughly a quarter of the entire landscape, is simulated. However, we note that these models will significantly under-perform on regimes of higher COF or F_0, due to the skewed training data. Nonetheless, this approach shows the benefit of scanning for properties on both extremes for screening processes.

4.4 Conclusion

We have demonstrated several key benefits of utilizing ML models for high-throughput screening processes. Applying iterative ML techniques to a traditional MD screening workflow can significantly improve the efficiency of the screening process by targeting regions of vast parameter spaces with the most favorable properties. The predictive ML models also allow for further extrapolation outside of the screened parameter space, extending the impact of the screening process.

Bibliography

1. Bhushan, B. & Sundararajan, S. Micro/Nanoscale Friction and Wear Mechanisms of Thin Films Using Atomic Force and Friction Force Microscopy. *Acta Mater.* **46,** 3793 (1998).

2. Tambe, N. S. & Bhushan, B. Nanotribological Characterization of Self-Assembled Monolayers Deposited on Silicon and Aluminium Substrates. *Nanotechnology* **16,** 1549 (2005).

3. Cummings, P. T., Docherty, H., Iacovella, C. R. & Singh, J. K. Phase transitions in nanoconfined fluids: The evidence from simulation and theory. *AIChE Journal,* NA–NA. https://doi.org/10.1002/aic.12226 (2010).

4. Erdemir, A. & Donnet, C. Tribology of diamond-like carbon films: recent progress and future prospects. *Journal of Physics D: Applied Physics* **39,** R311–R327. https://doi.org/10.1088/0022-3727/39/18/r01 (Sept. 2006).

5. Bouchet, M. I. D. B., Martin, J. M., Avila, J., Kano, M., Yoshida, K., Tsuruda, T., Bai, S., Higuchi, Y., Ozawa, N., Kubo, M. & Asensio, M. C. Diamond-like carbon coating under oleic acid lubrication: Evidence for graphene oxide formation in superlow friction. *Scientific Reports* **7.** https://doi.org/10.1038/srep46394 (Apr. 2017).

6. Zhao, J., Huang, Y., He, Y. & Shi, Y. Nanolubricant additives: A review. *Friction* **9,** 891–917. https://doi.org/10.1007/s40544-020-0450-8 (Dec. 2020).

7. Hoffmann, T., Lehmann, D. & Schäffler, M. Additives for lubricants containing poly(tetrafluoroethylene). Part 2: Tribological characterization. *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology* **226,** 222–229. https://doi.org/10.1177/1350650111429916 (Jan. 2012).

8. Asay, D. B., Dugger, M. T. & Kim, S. H. In-situ Vapor-Phase Lubrication of MEMS. *Tribology Letters* **29,** 67–74. https://doi.org/10.1007/s11249-007-9283-0 (Nov. 2007).

9. Berman, D., Erdemir, A. & Sumant, A. V. Graphene: a new emerging lubricant. *Materials Today* **17,** 31–42. ISSN: 1369-7021. https://www.sciencedirect.com/science/article/pii/S1369702113004574 (2014).

10. Vilt, S. G., Leng, Z., Booth, B. D., McCabe, C. & Jennings, G. K. Surface and Frictional Properties of Two-Component Alkylsilane Monolayers and Hydroxyl-Terminated Monolayers on Silicon. *The Journal of Physical Chemistry C* **113,** 14972–14977. ISSN: 1932-7447. eprint: https://doi.org/10.1021/jp904809h (Aug. 2009).

11. Quach, C. D., Gilmer, J. B., Pert, D. O., Mason-Hogans, A., Iacovella, C. R., Cummings, P. T. & McCabe, C. High-Throughput Screening of Tribological Properties of Monolayer Films Using Molecular Dynamics and Machine Learning. *The Journal of Chemical Physics,* 5.0080838. ISSN: 0021-9606, 1089-7690 (Feb. 2022).

12. Summers, A. Z., Gilmer, J. B., Iacovella, C. R., Cummings, P. T. & McCabe, C. MoSDeF, a Python Framework Enabling Large-Scale Computational Screening of Soft Matter: Application to Chemistry-Property Relationships in Lubricating Monolayer Films. *Journal of Chemical Theory and Computation* **0.** PMID: 32004433, null. ISSN: 1549-9618. eprint: https://doi.org/10.1021/acs.jctc.9b01183 (Mar. 0).

13. Yu, B., Qian, L., Yu, J. & Zhou, Z. Effects of Tail Group and Chain Length on the Tribological Behaviors of Self-Assembled Dual-Layer Films in Atmosphere and in Vacuum. *Tribology Letters* **34,** 1–10. ISSN: 1023-8883, 1573-2711 (Apr. 2009).

14. Brewer, N. J., Beake, B. D. & Leggett, G. J. Friction Force Microscopy of Self-Assembled Monolayers: Influence of Adsorbate Alkyl Chain Length, Terminal Group Chemistry, and Scan Velocity. *Langmuir* **17,** 1970–1974. ISSN: 0743-7463 (Mar. 2001).

15. Lewis, J. B., Vilt, S. G., Rivera, J. L., Jennings, G. K. & McCabe, C. Frictional Properties of Mixed Fluorocarbon/Hydrocarbon Silane Monolayers: A Simulation Study. *Langmuir* **28,** 14218–14226. ISSN: 0743-7463, 1520-5827 (Oct. 2012).

16. Le, T., Epa, V. C., Burden, F. R. & Winkler, D. A. Quantitative Structure-Property Relationship Modeling of Diverse Materials Properties. *Chemical Reviews* **112,** 2889–2919. ISSN: 0009-2665 (May 2012).

17. Rivera, J. L., Jennings, G. K. & McCabe, C. Examining the Frictional Forces between Mixed Hydrophobic Hydrophilic Alkylsilane Monolayers. *The Journal of Chemical Physics* **136,** 244701. ISSN: 0021-9606, 1089-7690 (June 2012).

18. Mazyar, O. A., Jennings, G. K. & McCabe, C. Frictional Dynamics of Alkylsilane Monolayers on SiO_2 : Effect of 1- *n* - Butyl-3-methylimidazolium Nitrate as a Lubricant. *Langmuir* **25,** 5103–5110. ISSN: 0743-7463, 1520-5827 (May 2009).

19. Contributors, M. *MoSDeF Webpage* https://mosdef.org. Accessed: 2022-03-13.

20. Dice, B. D., Butler, B. L., Ramasubramani, V., Travitz, A., Henry, M. M., Ojha, H., Wang, K. L., Adorf, C. S., Jankowski, E. & Glotzer, S. C. *Signac: Data Management and Workflows for Computational Researchers* in *Proceedings of the 20th Python in Science Conference* (2021), 23–32.

21. Thompson, M. W., Matsumoto, R., Sacci, R. L., Sanders, N. C. & Cummings, P. T. Scalable Screening of Soft Matter: A Case Study of Mixtures of Ionic Liquids and Organic Solvents. *The Journal of Physical Chemistry B* **123,** 1340–1347. ISSN: 1520-6106 (Feb. 2019).

22. Shamaprasad, P., Frame, C. O., Moore, T. C., Yang, A., Iacovella, C. R., Bouwstra, J. A., Bunge, A. L. & McCabe, C. Using molecular simulation to understand the skin barrier. *Progress in Lipid Research* **88,** 101184. https://doi.org/10.1016/j.plipres.2022.101184 (Nov. 2022).

23. Thompson, M. W., Gilmer, J. B., Matsumoto, R. A., Quach, C. D., Shamaprasad, P., Yang, A. H., Iacovella, C. R., McCabe, C. & Cummings, P. T. Towards Molecular Simulations That Are Transparent, Reproducible, Usable by Others, and Extensible (TRUE). *Molecular Physics* **118,** e1742938. ISSN: 0026-8976, 1362-3028 (June 2020).

24. Cummings, P. T., McCabe, C., Iacovella, C. R., Ledeczi, A., Jankowski, E., Jayaraman, A., Palmer, J. C., Maginn, E. J., Glotzer, S. C., Anderson, J. A., Ilja Siepmann, J., Potoff, J., Matsumoto, R. A., Gilmer, J. B., DeFever, R. S., Singh, R. & Crawford, B. Open-Source Molecular Modeling Software in Chemical Engineering Focusing on the Molecular Simulation Design Framework. *AIChE Journal* **67,** e17206. eprint: https://aiche.onlinelibrary.wiley.com/doi/pdf/10.1002/aic.17206 (2021).

25. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **596,** 583–589. ISSN: 1476-4687 (Aug. 2021).

26. Doerr, S., Majewski, M., Pérez, A., Kr ämer, A., Clementi, C., Noe, F., Giorgino, T. & De Fabritiis, G. TorchMD: A Deep Learning Framework for Molecular Simulations. *Journal of Chemical Theory and Computation* **17,** 2355–2363. ISSN: 1549-9618, 1549-9626 (Apr. 2021).

27. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less Is More: Sampling Chemical Space with Active Learning. *The Journal of Chemical Physics* **148,** 241733. ISSN: 0021-9606, 1089-7690 (June 2018).

28. Shmilovich, K., Mansbach, R. A., Sidky, H., Dunne, O. E., Panda, S. S., Tovar, J. D. & Ferguson, A. L. Discovery of Self-Assembling π-Conjugated Peptides by Active Learning-Directed Coarse-Grained Molecular Simulation. *The Journal of Physical Chemistry B* **124,** 3873–3891. ISSN: 1520-6106, 1520-5207 (May 2020).

29. Afzal, M. A. F., Haghighatlari, M., Ganesh, S. P., Cheng, C. & Hachmann, J. Accelerated Discovery of High-Refractive-Index Polyimides via First-Principles Molecular Modeling, Virtual High-Throughput Screening , and Data Mining. *The Journal of Physical Chemistry C* **123,** 14610–14618. ISSN: 1932-7447 (June 2019).

30. Kuenneth, C., Schertzer, W. & Ramprasad, R. Copolymer Informatics with Multitask Deep Neural Networks. *Macromolecules* **54,** 5957–5961. ISSN: 0024-9297, 1520-5835 (July 2021).

31. Faber, F. A., Hutchison, L., Huang, B., Gilmer, J., Schoenholz, S. S., Dahl, G. E., Vinyals, O., Kearnes, S., Riley, P. F. & von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *Journal of Chemical Theory and Computation* **13,** 5255–5264. ISSN: 1549-9618, 1549-9626 (Nov. 2017).

32. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chemical Science* **8,** 3192–3203. ISSN: 2041-6520, 2041-6539 (2017).

33. Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P. & Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* **43,** 1947–1958. ISSN: 0095-2338 (Nov. 2003).

34. Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L. & Overington, J. P. ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities. *Nucleic Acids Research* **43,** W612–620. ISSN: 1362-4962 (July 2015).

35. Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I. & Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Research* **45,** D945–D954. ISSN: 1362-4962 (Jan. 2017).

36. Summers, A. Z., Iacovella, C. R., Cummings, P. T. & McCabe, C. Investigating Alkylsilane Monolayer Tribology at a Single-Asperity Contact with Molecular Dynamics Simulation. *Langmuir* **33,** 11270–11280. ISSN: 0743-7463 (Oct. 2017).

37. Black, J. E., Iacovella, C. R., Cummings, P. T. & McCabe, C. Molecular Dynamics Study of Alkylsilane Monolayers on Realistic Amorphous Silica Surfaces. *Langmuir* **31,** 3086–3093. ISSN: 0743-7463, 1520-5827 (Mar. 2015).

38. Klein, C., Sallai, J., Jones, T. J., Iacovella, C. R., McCabe, C. & Cummings, P. T. in *Foundations of Molecular Modeling and Simulation: Select Papers from FOMMS 2015* (eds Snurr, R. Q., Adjiman, C. S. & Kofke, D. A.) 79–92 (Springer Singapore, Singapore, 2016). ISBN: 978-981-10-1128-3.

39. Klein, C., Summers, A. Z., Thompson, M. W., Gilmer, J. B., McCabe, C., Cummings, P. T., Sallai, J. & Iacovella, C. R. Formalizing Atom-Typing and the Dissemination of Force Fields with Foyer. *Computational Materials Science* **167,** 215–227. ISSN: 0927-0256. http://www.sciencedirect.com/science/article/pii/S0927025619303040 (Sept. 2019).

40. *mBuild Github repository* https://github.com/mosdef-hub/mbuild (2018).

41. *Foyer Github repository* https://github.com/mosdef-hub/foyer (2018).

42. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **118,** 11225–11236. ISSN: 0002-7863 (Nov. 1996).

43. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91,** 43–56. ISSN: 0010-4655 (Sept. 1995).

44. Abraham, M. J., Murtola, T., Schulz, R., Pá ll, S., Smith, J. C., Hess, B. & Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **1-2,** 19–25. ISSN: 2352-7110 (Sept. 2015).

45. Lorenz, C., Webb, E., Stevens, M., Chandross, M. & Grest, G. Frictional Dynamics of Perfluorinated Self-Assembled Monolayers on Amorphous SiO2. *Tribology Letters* **19,** 93–98. ISSN: 1023-8883, 1573-2711 (June 2005).

46. Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Physical Review A: Atomic, Molecular, and Optical Physics* **31,** 1695–1697 (3 Mar. 1985).

47. Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics* **117,** 1–19. ISSN: 0021-9991. http://www.sciencedirect.com/science/article/pii/S002199918571039X (Mar. 1995).

48. Thompson, A. P., Aktulga, H. M., Berger, R., Bolintineanu, D. S., Brown, W. M., Crozier, P. S., in 't Veld, P. J., Kohlmeyer, A., Moore, S. G., Nguyen, T. D., Shan, R., Stevens, M. J., Tranchida, J., Trott, C. & Plimpton, S. J. LAMMPS - a Flexible Simulation Tool for Particle-Based Materials Modeling at the Atomic, Meso, and Continuum Scales. **271,** 108171 (2022).

49. Contributors, S. *Signac Framework Webpage* https://signac.io. Accessed: 2022-03-13.

50. McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L.-P., Lane, T. J. & Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **109,** 1528–1532. ISSN: 0006-3495 (Oct. 2015).

51. Pavlov, Y. L. *Random Forests* ISBN: 978-3-11-094197-5 (2019).

52. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. d. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830. ISSN: 1532-4435. arXiv: 1201.0490 (2012).

53. Landrum, G. *et al.* RDKit: Open-source cheminformatics (2006).

54. Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Modeling* **28,** 31–36. ISSN: 1549-9596 (Feb. 1988).

55. Patel, R., Borca, C. & Webb, M. *Featurization Strategies for Polymer Sequence or Composition Design by Machine Learning* Preprint (Chemistry, Nov. 2021).

56. Vishwakarma, G., Sonpal, A. & Hachmann, J. Metrics for Benchmarking and Uncertainty Quantification: Quality, Applicability, and Best Practices for Machine Learning in Chemistry. *Trends in Chemistry* **3,** 146–156. ISSN: 2589-5974 (Feb. 2021).

57. Labute, P. A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling* **18,** 464–477. ISSN: 1093-3263. https://www.sciencedirect.com/science/article/pii/S1093326300000681 (2000).

58. Baumgärtner, A. Shapes of Flexible Vesicles at Constant Volume. *The Journal of Chemical Physics* **98,** 7496–7501. ISSN: 0021-9606, 1089-7690 (May 1993).

59. Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chemical Physics Letters* **89,** 399–404. ISSN: 0009-2614 (1982).

60. Bertz, S. H. Convergence, Molecular Complexity, and Synthetic Analysis. *Journal of the American Chemical Society* **104,** 5801–5803. ISSN: 0002-7863 (Oct. 1982).

61. Hall, L. H. & Kier, L. B. in *Reviews in Computational Chemistry* 367–422 (John Wiley & Sons, Ltd, 1991). ISBN: 978-0-470-12579-3. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470125793.ch9.

62. Arteca, G. in *Reviews in Computational Chemistry, Vol 20* 191–253 (Jan. 2007). ISBN: 978-0-470-12586-1.

63. Todeschini, R. & Consonni, V. *Handbook of Molecular Descriptors* ISBN: 978-3-527-61310-6 978-3-527-61311-3 (2011).

64. Bonchev, D. & Trinajstić, N. Information Theory, Distance Matrix, and Molecular Branching. *The Journal of Chemical Physics* **67,** 4517–4533. eprint: https://doi.org/10.1063/1.434593 (1977).

65. Wildman, S. A. & Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences* **39,** 868–873. ISSN: 0095-2338 (Sept. 1999).

66. Sauer, W. H. B. & Schwarz, M. K. Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity. *Journal of Chemical Information and Computer Sciences* **43,** 987–1003. ISSN: 0095-2338 (May 2003).

67. Firth, N. C., Brown, N. & Blagg, J. Plane of Best Fit: A Novel Method to Characterize the Three-Dimensionality of Molecules. *Journal of Chemical Information and Modeling* **52,** 2516–2525. ISSN: 1549-9596 (Oct. 2012).

68. Robinson, D. D., Barlow, T. W. & Richards, W. G. The Utilization of Reduced Dimensional Representations of Molecular Structure for Rapid Molecular Similarity Calculations. *Journal of Chemical Information and Computer Sciences* **37,** 943–950. ISSN: 0095-2338 (Sept. 1997).

69. Ertl, P., Rohde, B. & Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *Journal of Medicinal Chemistry* **43,** 3714–3717. ISSN: 0022-2623 (Oct. 2000).

Appendix A

Accessing the source code and data

A.1 Using the released Repository

To install `conda` to a local machine, run the following commands in your terminal based on your operating system (note that the initial "$" is meant to denote a line on the command line):

A.1.1 `MacOS`

```
$ cd ${HOME}
$ curl -O https://repo.anaconda.com/miniconda/Miniconda3-latest-MacOSX-x86_64.sh
$ /bin/bash Miniconda3-latest-MacOSX-x86_64.sh
```

A.1.2 `Linux`

```
$ cd ${HOME}
$ curl -O https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
$ /bin/bash Miniconda3-latest-Linux-x86_64.sh
```

A.2 Cloning the repository and creating the python environment

Once the `conda` package manager has been installed, the reader can proceed to clone the repository and create a working environment. We note, due to the sheer size of the simulation data, only the analysis routines, along with the summarized data, is hosted on GitHub for easy access; the full simulation workflow is archived and uploaded Zenodo.

`MacOS` and `Linux`:

```
$ git clone https://github.com/daico007/iMoDELS-supplements.git
$ cd iModels-supplements
$ conda install -c conda-forge mamba
$ mamba env create --file env.yml
$ conda activate screeni
```

A.3 Utilizing the repository

The `iModels-supplements` repository includes the summarized data generated from the MD simulations and routines used to train and evaluate the efficacy of the ML models. The raw data contains information regarding the systems as well as the calculated tribological properties, while the analysis routines can be found in a collection of Python scripts (used to perform the training and evaluation of the ML models) and `Jupyter` notebooks (used to plot and visualize data). We have set up a `Jupyter`

notebook, named `Data-Lookup.ipynb`, to specifically assist with 1) Looking up data generated from MD and 2) utilize the trained ML models to predict tribological properties of new systems. More details of the structure of the repository can be found on the GitHub page accompanied the paper at $https://github.com/daico007/iMoDELS-supplements/$.

Appendix B

Additional Force Field Details

This study utilized the Optimized Potential for Liquid System - All Atom, consistent with that used in Summers *et al.*, study[12,42,45]. Beyond parameters for chemistries studied in the previous work, interaction parameters for the three new chemistries (toluene, phenol, and difluoromethyl) are presented below. The complete list of parameters is stored in a `foyer`-compatible XML file, named `oplsaa.xml`, and included with the workflow repository at https://github.com/daico007/iMoDELS-supplements/.

B.1 Toluene

Table B.1: Toluene nonbonded parameters.

Nonbonded parameters					
Atom Type	Element	Charge	Sigma, Å	Epsilon, kcal mol^{-1}	Reference
opls_140	H	0.06	2.5	0.03	[42]
opls_148	C	-0.065	3.5	0.066	[42]
opls_145	C	-0.115	3.55	0.07	[42]
opls_146	H	0.115	2.42	0.03	[42]

Table B.2: Toluene bonded parameters.

Harmonic Bond parameters				
Bond	Elements	k, kcal mol^{-1} Å^{-2}	r_0, Å	Reference
opls_149-opls_140	C-H	340	1.09	[42]
opls_145-opls_148	C-C	317	1.51	[42]
opls_145-opls_145	C-C	469	1.4	[42]
opls_145-opls_149	C-C	317	1.51	[42]
opls_140-opls_148	H-C	340	1.09	[42]

Table B.3: Toluene angle parameters.

Harmonic Angle parameters				
Angle	Elements	k, kcal mol^{-1} deg^{-2}	θ_0, deg	Reference
opls_149-opls_145-opls_145	C-C-C	70	120	[42]
opls_140-opls_149-opls_145	H-C-C	35	109.5	[42]
opls_136-opls_149-opls_145	C-C-C	63	114	[42]
opls_148-opls_145-opls_145	C-C-C	70	120	[42]
opls_145-opls_148-opls_140	C-C-H	35	109.5	[42]
opls_145-opls_145-opls_145	C-C-C	63	120	[42]
opls_145-opls_145-opls_146	C-C-H	35	120	[42]
opls_140-opls_148-opls_140	H-C-H	33	107.8	[42]

Table B.4: Toluene dihedral parameters.

Dihedral parameters						
Dihedral	Elements	k_1	k_2	k_3	k_4	Reference
		kcal mol^{-1}				
opls_149-opls_145-opls_145-opls_145	C-C-C-C	0	7.25	0	0	[42]
opls_149-opls_145-opls_145-opls_146	C-C-C-H	0	7.25	0	0	[42]
opls_140-opls_149-opls_145-opls_145	H-C-C-C	0	0	0	0	[42]
opls_136-opls_149-opls_145-opls_145	C-C-C-C	0	0	0	0	[42]
opls_140-opls_136-opls_149-opls_145	H-C-C-C	-1.2×10^{-6}	0	0.462	0	[42]
opls_148-opls_145-opls_145-opls_145	C-C-C-C	0	7.25	0	0	[42]
opls_148-opls_145-opls_145-opls_146	C-C-C-H	0	7.25	0	0	[42]
opls_145-opls_145-opls_145-opls_145	C-C-C-C	0	7.25	0	0	[42]
opls_145-opls_145-opls_145-opls_146	C-C-C-H	0	7.25	0	0	[42]
opls_145-opls_145-opls_148-opls_140	C-C-C-H	0	0	0	0	[42]
opls_146-opls_145-opls_145-opls_146	H-C-C-H	0	7.25	0	0	[42]

Table B.5: Toluene improper parameters.

Improper parameters[1]					
Improper[1]	Elements	K_ϕ, kcal mol^{-1}	n	γ, deg	Reference
opls_148-opls_145-opls_145-opls_145	C-C-C-C	1.1	2	180	[42]
opls_145-opls_145-opls_145-opls_146	C-C-C-H	1.1	2	180	[42]
opls_149-opls_145-opls_145-opls_145	C-C-C-C	1.1	2	180	[42]

[1] Dihedral OPLS parameters are converted from Ryckaert-Bell parameters stored in the "oplsaa.xml".

B.2 Phenol

Table B.6: Phenol nonbonded parameters.

Nonbonded parameters					
Atom Type	Element	Charge	Sigma, Å	Epsilon, kcal mol^{-1}	Reference
opls_145	C	-0.115	3.55	0.07	[42]
opls_166	C	0.15	3.55	0.07	[42]
opls_167	O	-0.585	3.07	0.17	[42]
opls_146	H	0.115	2.42	0.03	[42]
opls_168	H	0.435	10	0.0	[42]

Table B.7: Phenol bonded parameters.

Harmonic Bond parameters				
Bond	Elements	k, kcal mol^{-1} Å^{-2}	r_0, Å	Reference
opls_145-opls_149	C-C	317	1.51	[42]
opls_145-opls_145	C-C	469	1.4	[42]
opls_145-opls_166	C-C	469	1.4	[42]
opls_167-opls_166	O-C	450	1.364	[42]
opls_146-opls_145	H-C	367	1.08	[42]
opls_168-opls_167	H-O	553	0.945	[42]

Table B.8: Phenol angle parameters.

Harmonic Angle parameters				
Angle	Elements	k, kcal mol^{-1} deg^{-2}	θ_0, deg	Reference
opls_149-opls_145-opls_145	C-C-C	70	120	[42]
opls_140-opls_149-opls_145	H-C-C	35	109.5	[42]
opls_136-opls_149-opls_145	C-C-C	63	114	[42]
opls_145-opls_145-opls_145	C-C-C	63	120	[42]
opls_145-opls_145-opls_146	C-C-H	35	120	[42]
opls_145-opls_145-opls_166	C-C-C	63	120	[42]
opls_145-opls_166-opls_145	C-C-C	63	120	[42]
opls_145-opls_166-opls_167	C-C-O	70	120	[42]
opls_166-opls_145-opls_146	C-C-H	35	120	[42]
opls_166-opls_145-opls_145	C-C-C	63	120	[42]
opls_166-opls_167-opls_168	C-O-H	35	113	[42]

Table B.9: Phenol dihedral parameters.

Dihedral parameters						
Dihedral	Elements	k_1	k_2	k_3	k_4	Reference
		kcal mol^{-1}				
opls_149-opls_145-opls_145-opls_145	C-C-C-C	0	7.25	0	0	[42]
opls_149-opls_145-opls_145-opls_146	C-C-C-H	0	7.25	0	0	[42]
opls_140-opls_149-opls_145-opls_145	H-C-C-C	0	0	0	0	[42]
opls_136-opls_149-opls_145-opls_145	C-C-C-C	0	0	0	0	[42]
opls_140-opls_136-opls_149-opls_145	H-C-C-C	-1.20E-06	0	0.46199928	0	[42]
opls_145-opls_145-opls_145-opls_166	C-C-C-C	0	7.25	0	0	[42]
opls_145-opls_145-opls_145-opls_146	C-C-C-H	0	7.25	0	0	[42]
opls_145-opls_145-opls_145-opls_145	C-C-C-C	0	7.25	0	0	[42]
opls_145-opls_145-opls_166-opls_145	C-C-C-C	0	7.25	0	0	[42]
opls_145-opls_145-opls_166-opls_167	C-C-C-O	0	7.25	0	0	[42]
opls_145-opls_166-opls_145-opls_145	C-C-C-C	0	7.25	0	0	[42]
opls_145-opls_166-opls_145-opls_146	C-C-C-H	0	7.25	0	0	[42]
opls_145-opls_166-opls_167-opls_168	C-C-O-H	0	1.68200048	0	0	[42]
opls_166-opls_145-opls_145-opls_146	C-C-C-H	0	7.25	0	0	[42]
opls_167-opls_166-opls_145-opls_146	O-C-C-H	0	7.25	0	0	[42]
opls_146-opls_145-opls_145-opls_146	H-C-C-H	0	7.25	0	0	[42]

Table B.10: Phenol improper parameters.

Improper parameters[1]					
Improper[1]	Elements	K_ϕ, kcal mol^{-1}	n	γ, deg	Reference
opls_149-opls_145-opls_145-opls_145	C-C-C-C	1.1	2	180	[42]
opls_145-opls_145-opls_145-opls_146	C-C-C-H	1.1	2	180	[42]
opls_145-opls_166-opls_145-opls_146	C-C-C-H	1.1	2	180	[42]
opls_145-opls_145-opls_166-opls_167	C-C-C-O	1.1	2	180	[42]
opls_166-opls_145-opls_145-opls_146	C-C-C-H	1.1	2	180	[42]

[1] Dihedral OPLS parameters are converted from Ryckaert-Bell parameters stored in the "oplsaa.xml".

B.3 Difluoromethyl

Table B.11: Difluoromethyl nonbonded parameters.

Nonbonded parameters					
Atom Type	Element	Charge	Sigma, Å	Epsilon, kcal mol^{-1}	Reference
opls_140	H	0.06	2.5	0.03	[42]
opls_962	C	0.24	3.5	0.066	[42]
opls_965	F	-0.12	2.95	0.053	[42]

Table B.12: Difluoromethyl bonded parameters.

Harmonic Bond parameters				
Bond	Elements	k, kcal mol^{-1} Å^{-2}	r_0, Å	Reference
opls_136-opls_140	C-H	340	1.09	[42]
opls_140-opls_136	H-C	340	1.09	[42]
opls_136-opls_140	C-H	340	1.09	[42]
opls_140-opls_136	H-C	340	1.09	[42]
opls_1004-opls_140	C-H	340	1.09	[42]
opls_140-opls_1004	H-C	340	1.09	[42]
opls_962-opls_136	C-C	268	1.529	[42]
opls_965-opls_962	F-C	367	1.332	[42]
opls_965-opls_962	F-C	367	1.332	[42]
opls_140-opls_962	H-C	340	1.09	[42]

Table B.13: Difluoromethyl angle parameters.

Harmonic Angle parameters				
Angle	Elements	k, kcal mol^{-1} deg^{-2}	θ_0, deg	Reference
opls_136-opls_962-opls_965	C-C-F	50	109.5	[42]
opls_136-opls_962-opls_140	C-C-H	37.5	110.7	[42]
opls_140-opls_136-opls_962	H-C-C	37.5	110.7	[42]
opls_136-opls_136-opls_962	C-C-C	58.3500239	112.7	[42]
opls_965-opls_962-opls_965	F-C-F	77	109.1	[42]
opls_965-opls_962-opls_140	F-C-H	40	107	[42]

Table B.14: Difluoromethyl dihedral parameters.

Dihedral parameters						
Dihedral	Elements	k_1	k_2	k_3 kcal mol^{-1}	k_4	Reference
opls_140-opls_136-opls_962-opls_965	H-C-C-F	0	0	0.4	0	[42]
opls_140-opls_136-opls_962-opls_140	H-C-C-H	0	0	0.3	0	[42]
opls_136-opls_136-opls_962-opls_965	C-C-C-F	0.3	0	0.4	0	[42]
opls_136-opls_136-opls_962-opls_140	C-C-C-H	0	0	0.3	0	[42]
opls_140-opls_136-opls_136-opls_962	H-C-C-C	0	0	0.3	0	[42]

Appendix C

Additional High-throughput Screening Result

The entirety of the MD high-throughput data can be presented as heatmap. In in the sections below, the terminal group of the bottom monolayer, *i.e.*, the uniform monolayer, will be held constant, while the corresponding terminal group combinations in the top monolayer, *i.e.*, the mixed monolayer, will be shown as the x and y axis of each heatmap. Also, their relative composition will be denoted on the individual axes.

C.1 Hydroxyl Terminated Monolayer

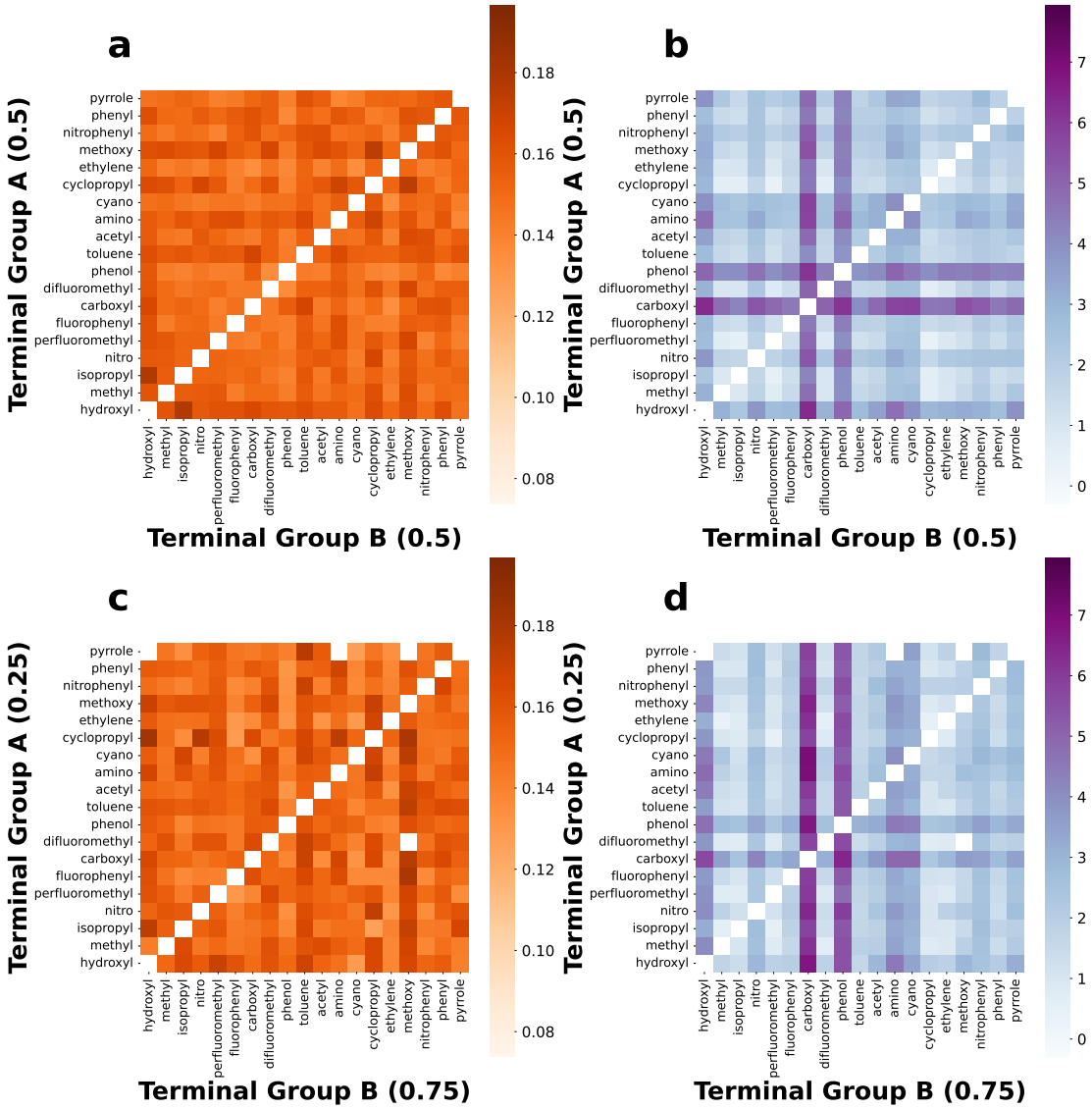

Figure C.1: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only hydroxyl terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.2 Methyl Terminated Monolayer

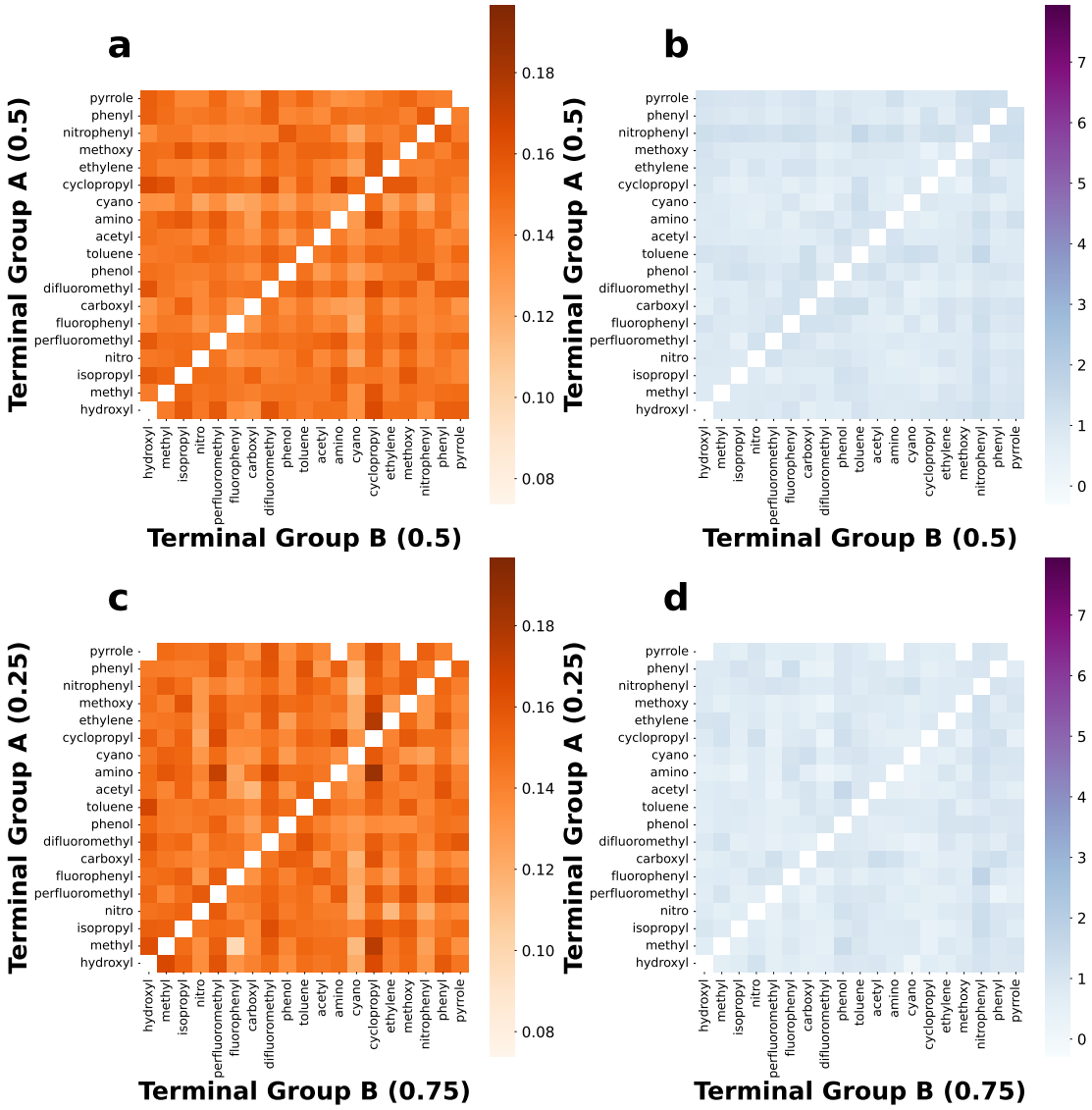

Figure C.2: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only methyl terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.3 Isopropyl Terminated Monolayer

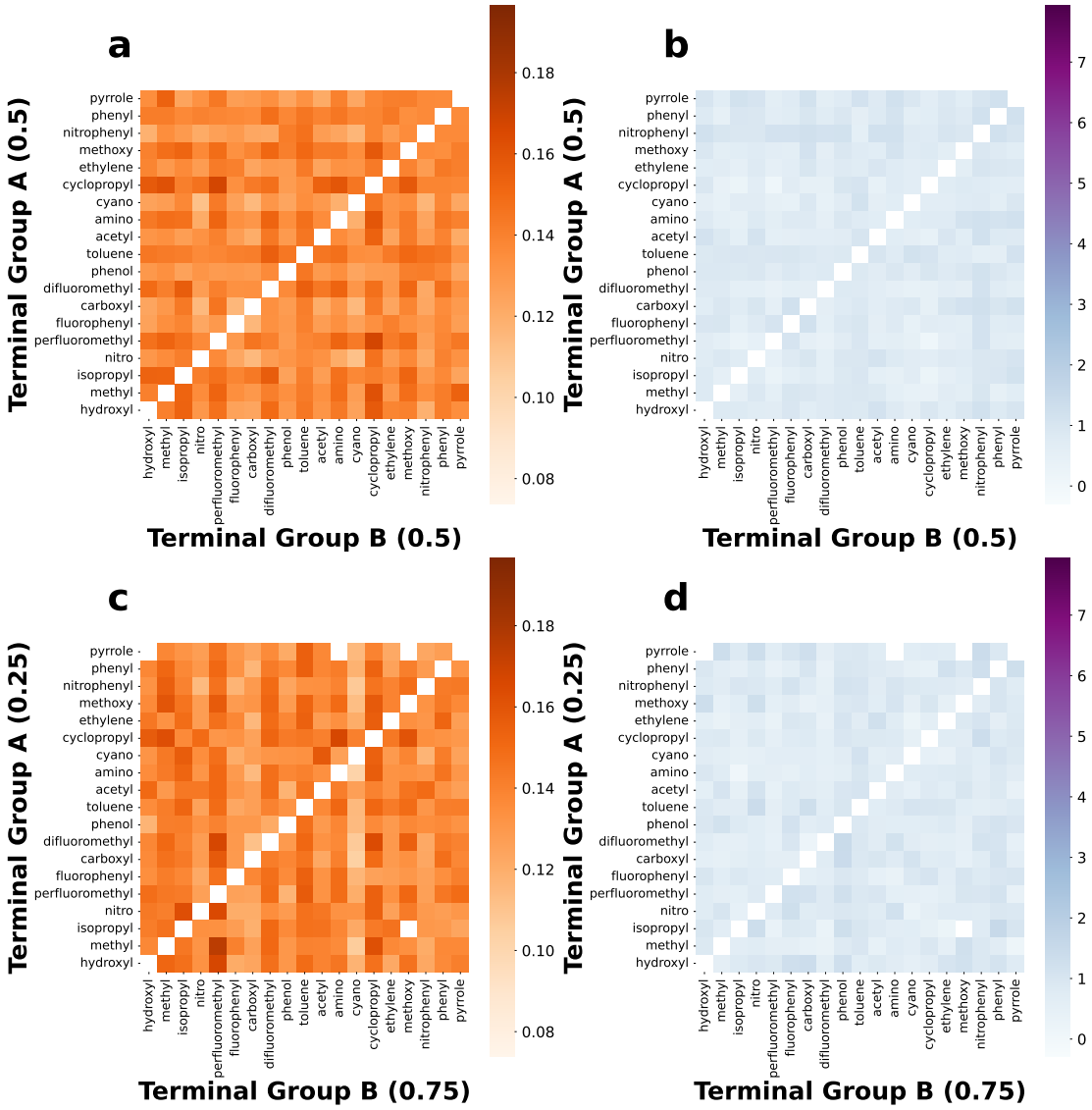

Figure C.3: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only isopropyl terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.4 Nitro Terminated Monolayer

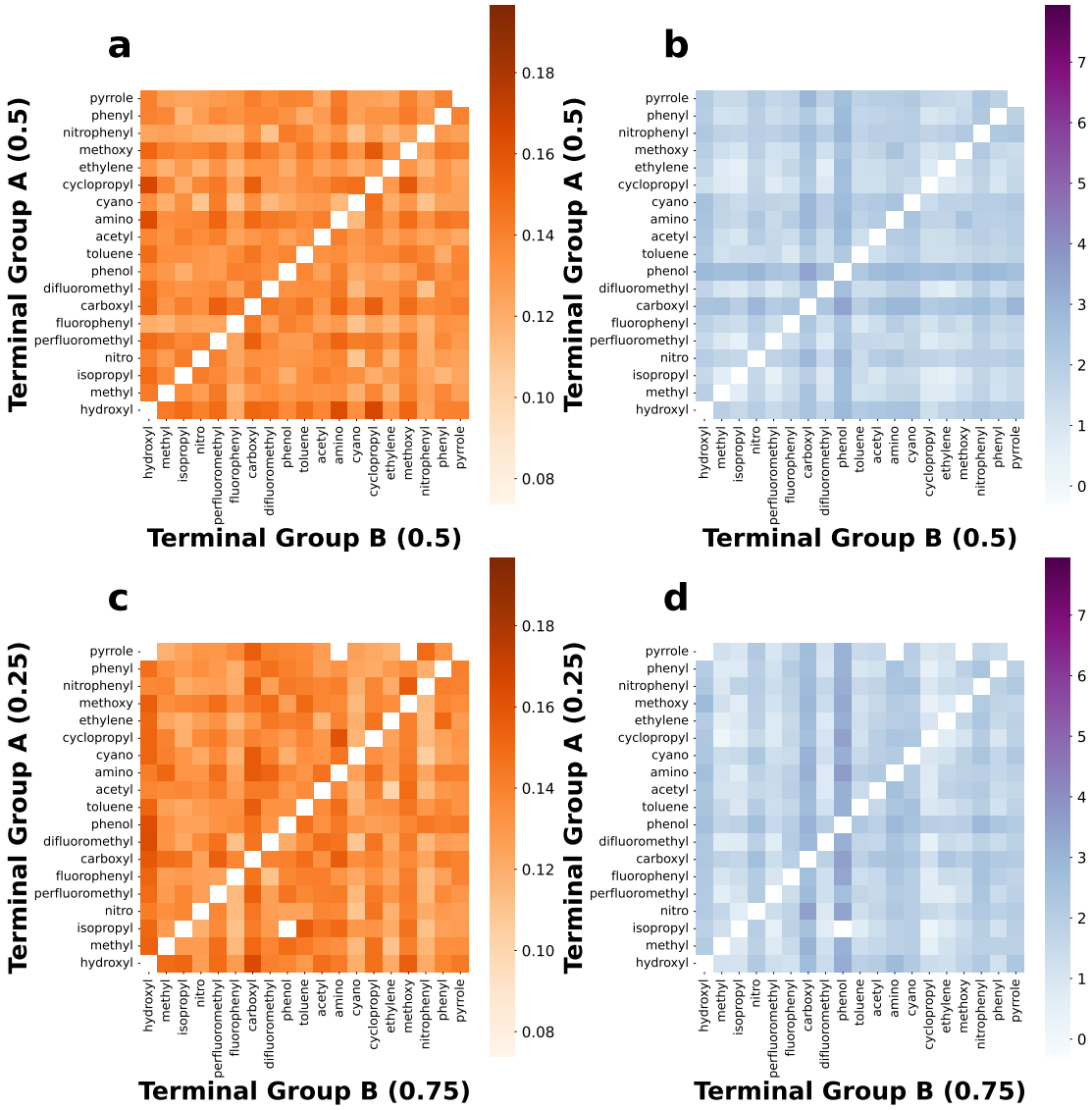

Figure C.4: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only nitro terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.5 Perfluoromethyl Terminated Monolayer

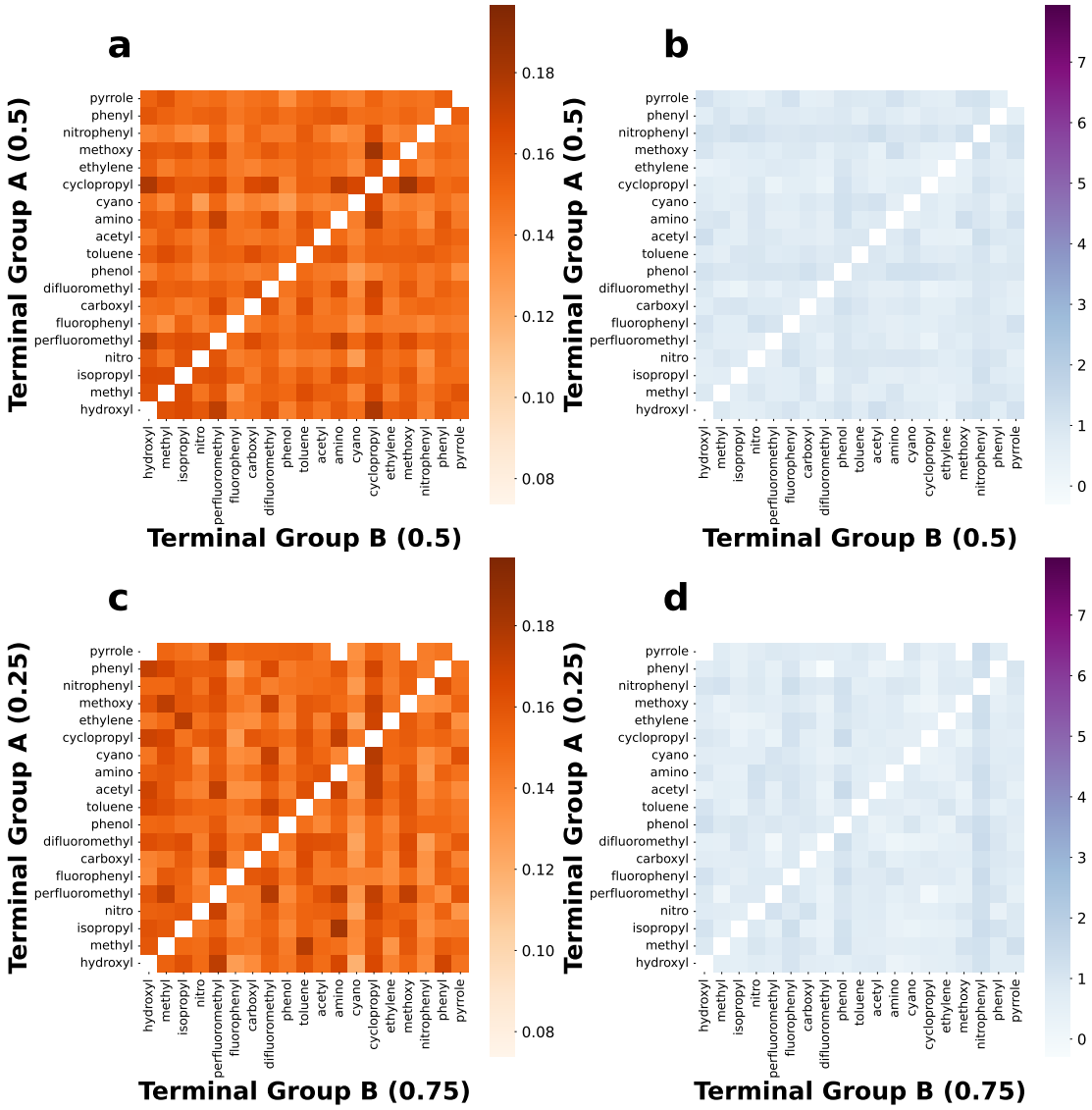

Figure C.5: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only perfluoromethyl terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.6 Fluorophenyl Terminated Monolayer

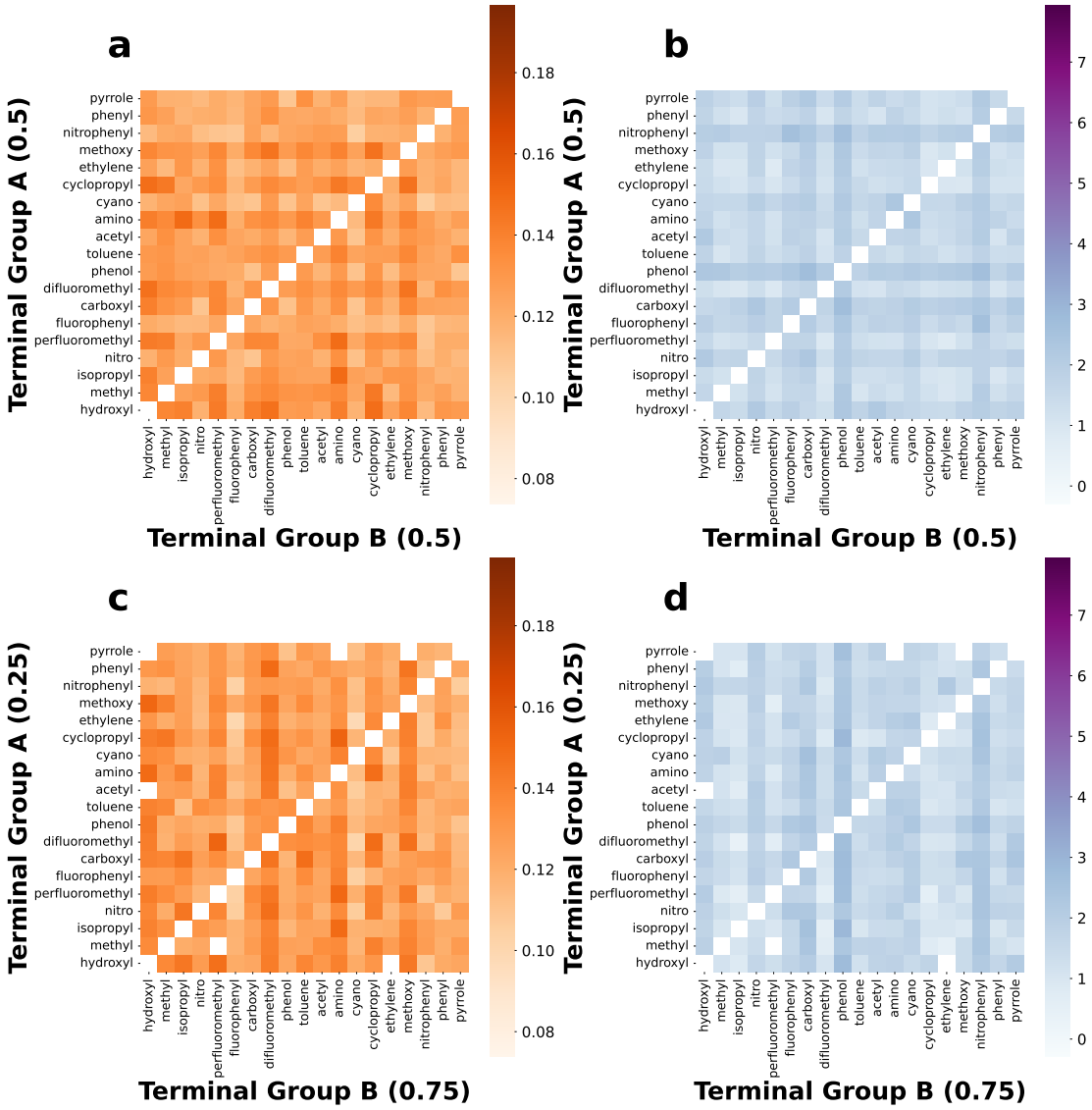

Figure C.6: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only fluorophenyl terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.7 Carboxyl Terminated Monolayer

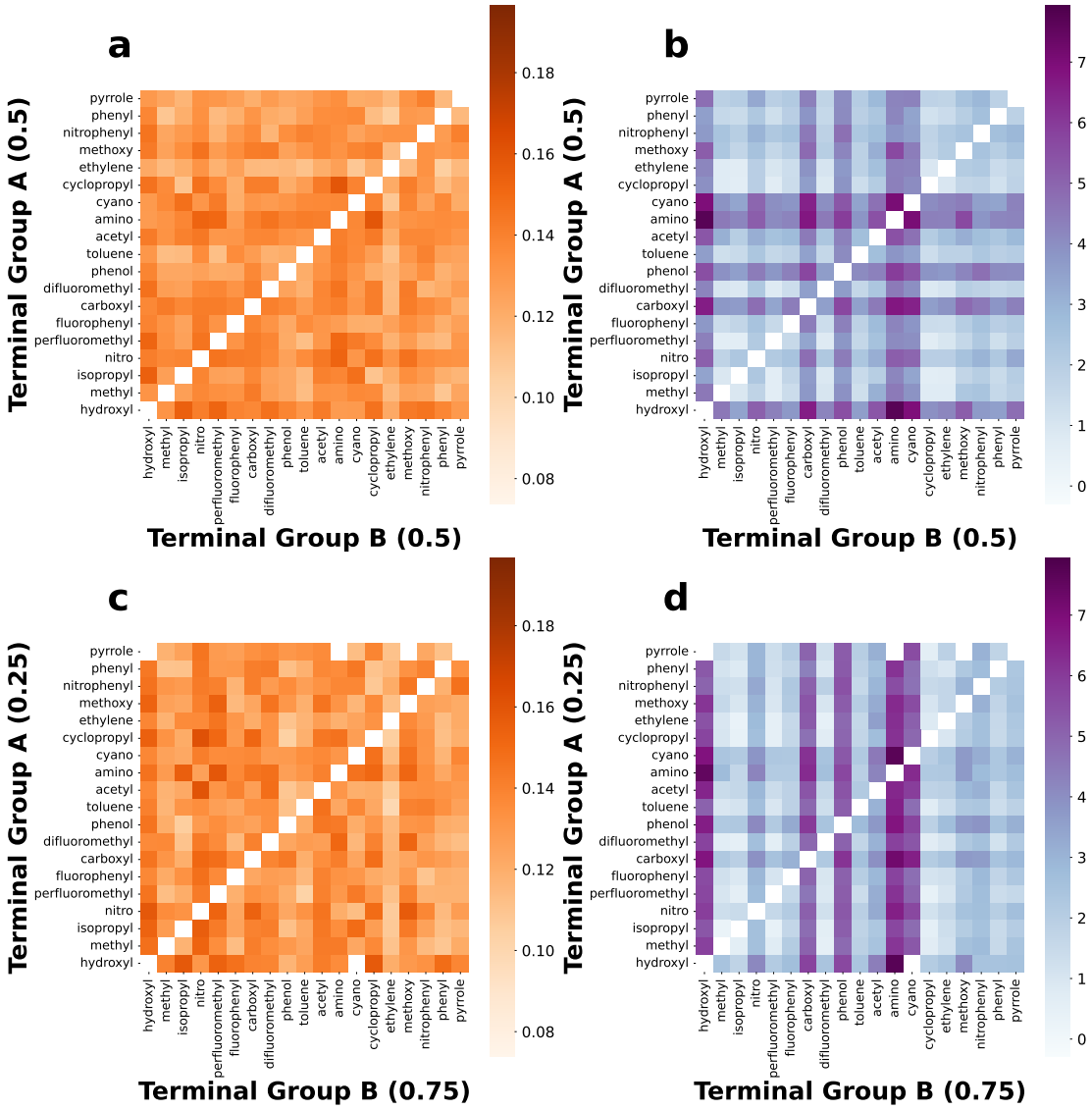

Figure C.7: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only carboxyl terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.8 Difluoromethyl Terminated Monolayer

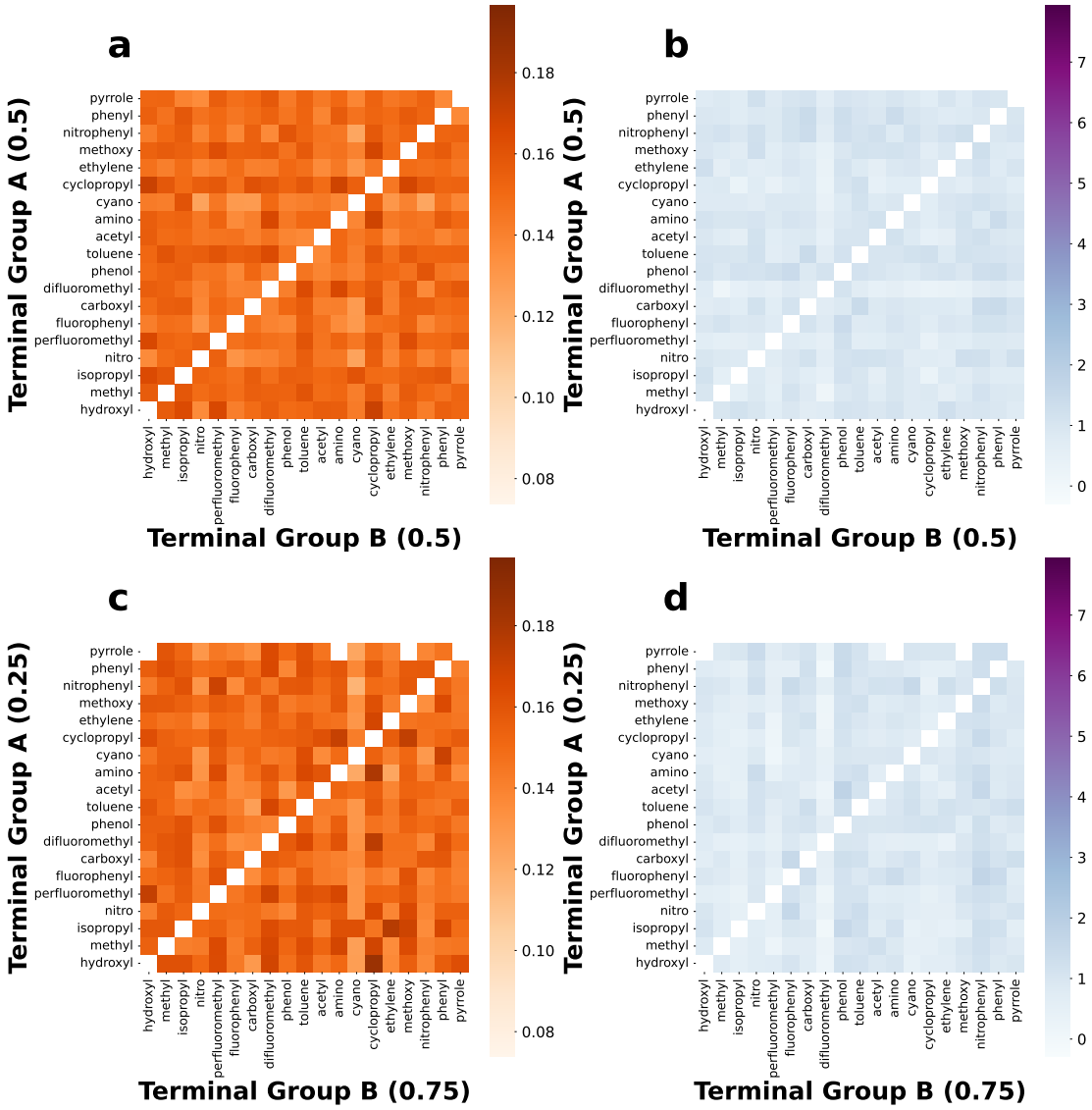

Figure C.8: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only difluoromethyl terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.9 Phenol Terminated Monolayer

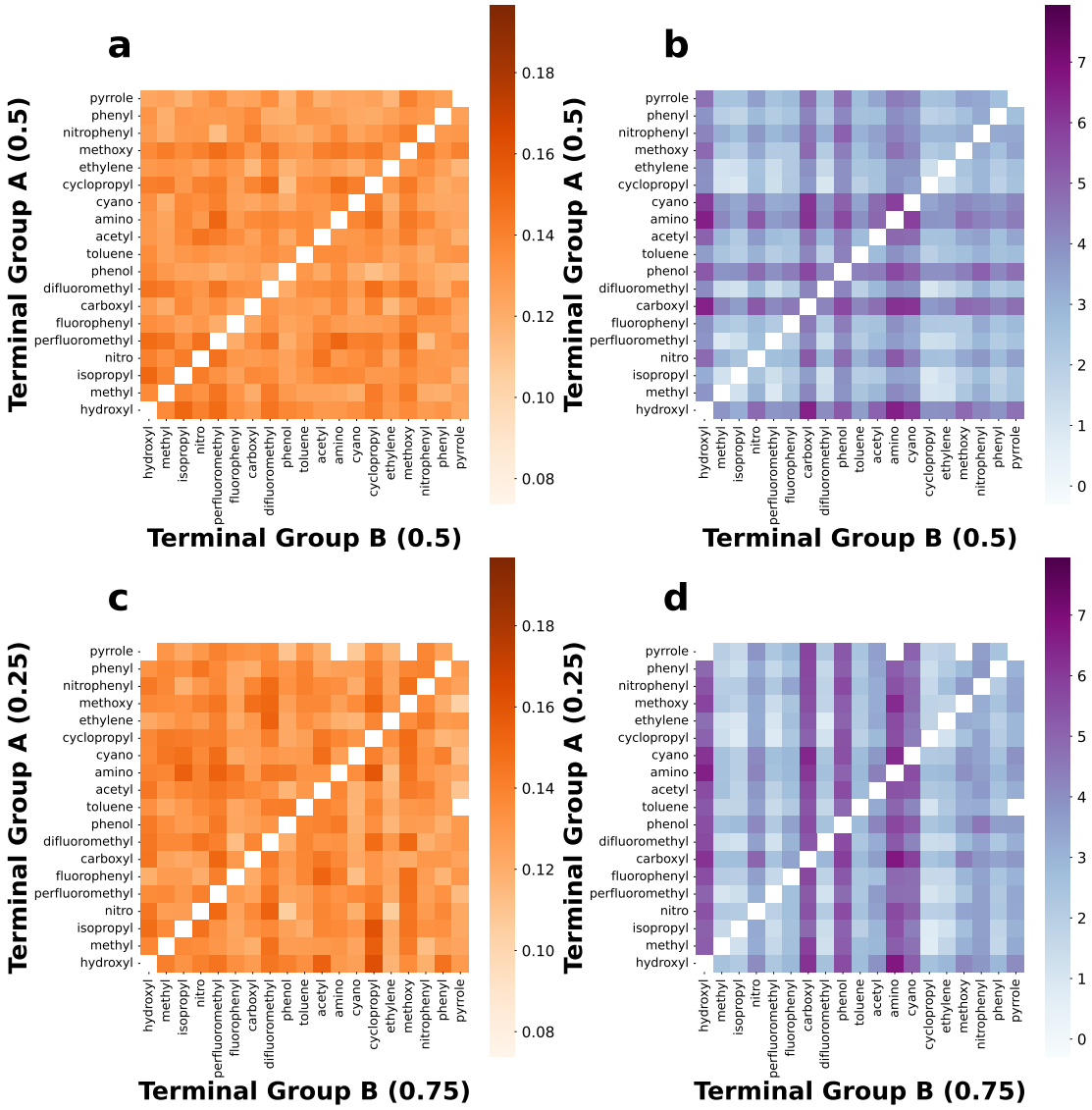

Figure C.9: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only phenol terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.10 Toluene Terminated Monolayer

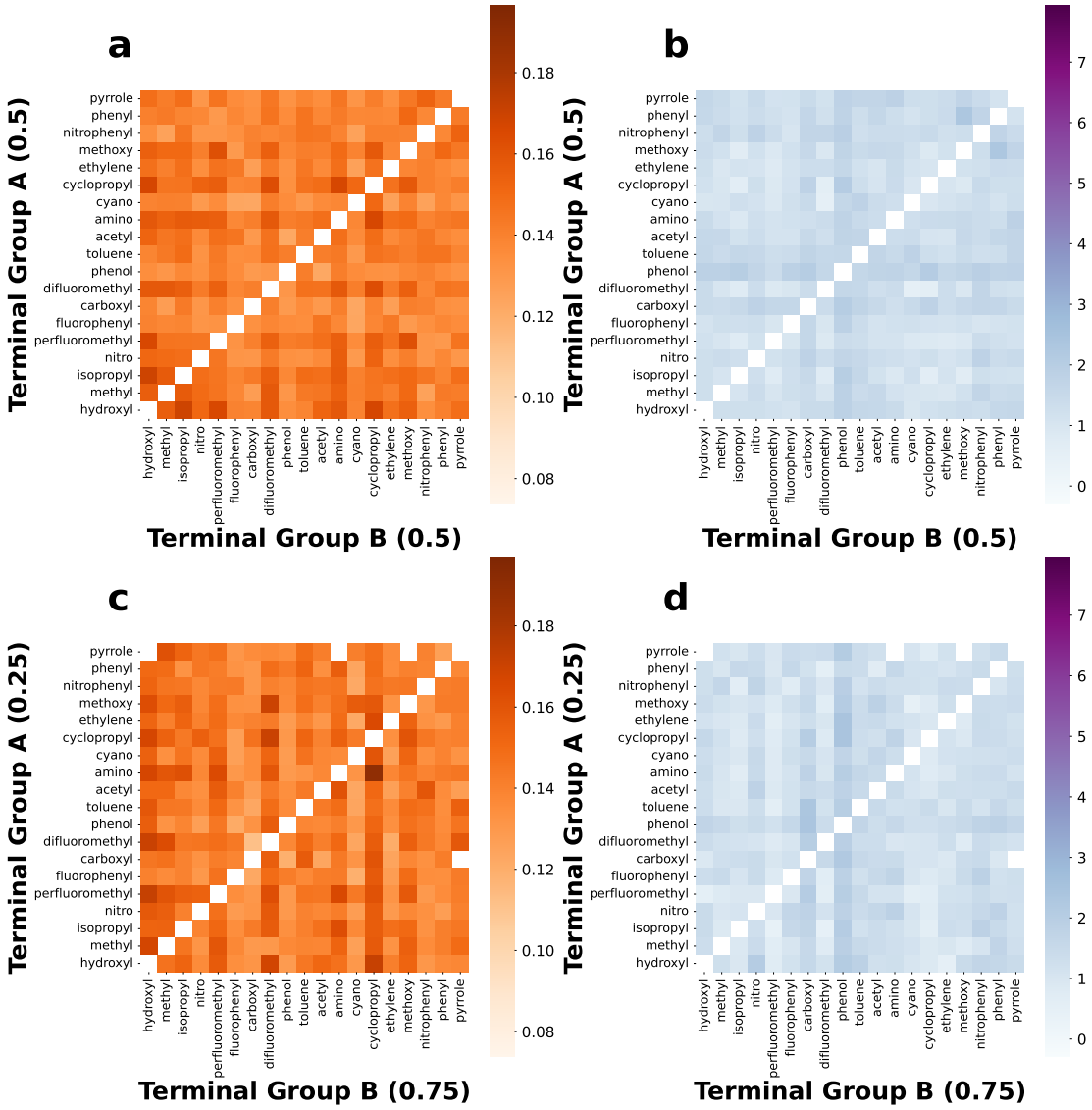

Figure C.10: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only toluene terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.11 Acetyl Terminated Monolayer

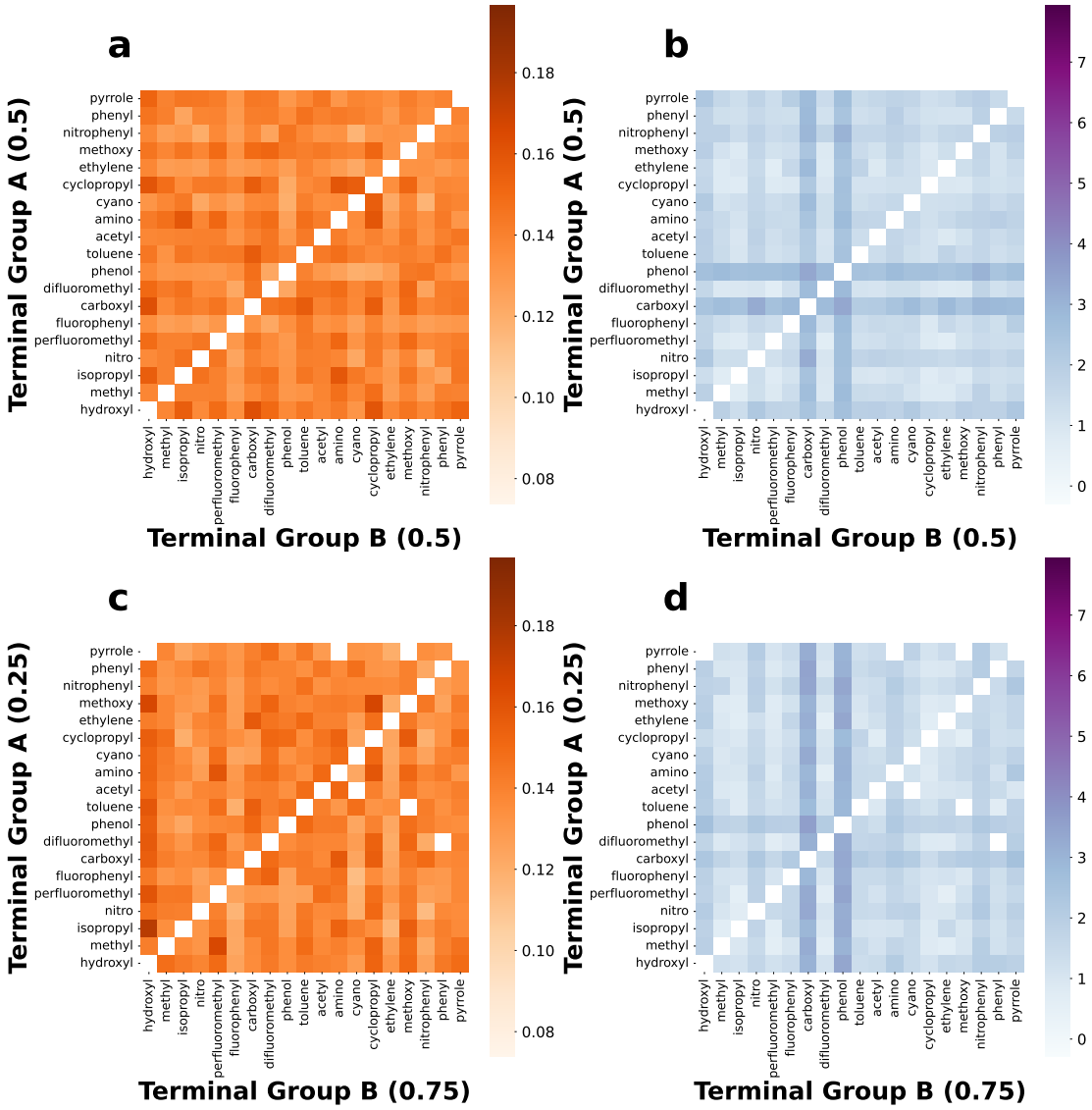

Figure C.11: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only acetyl terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.12 Amino Terminated Monolayer

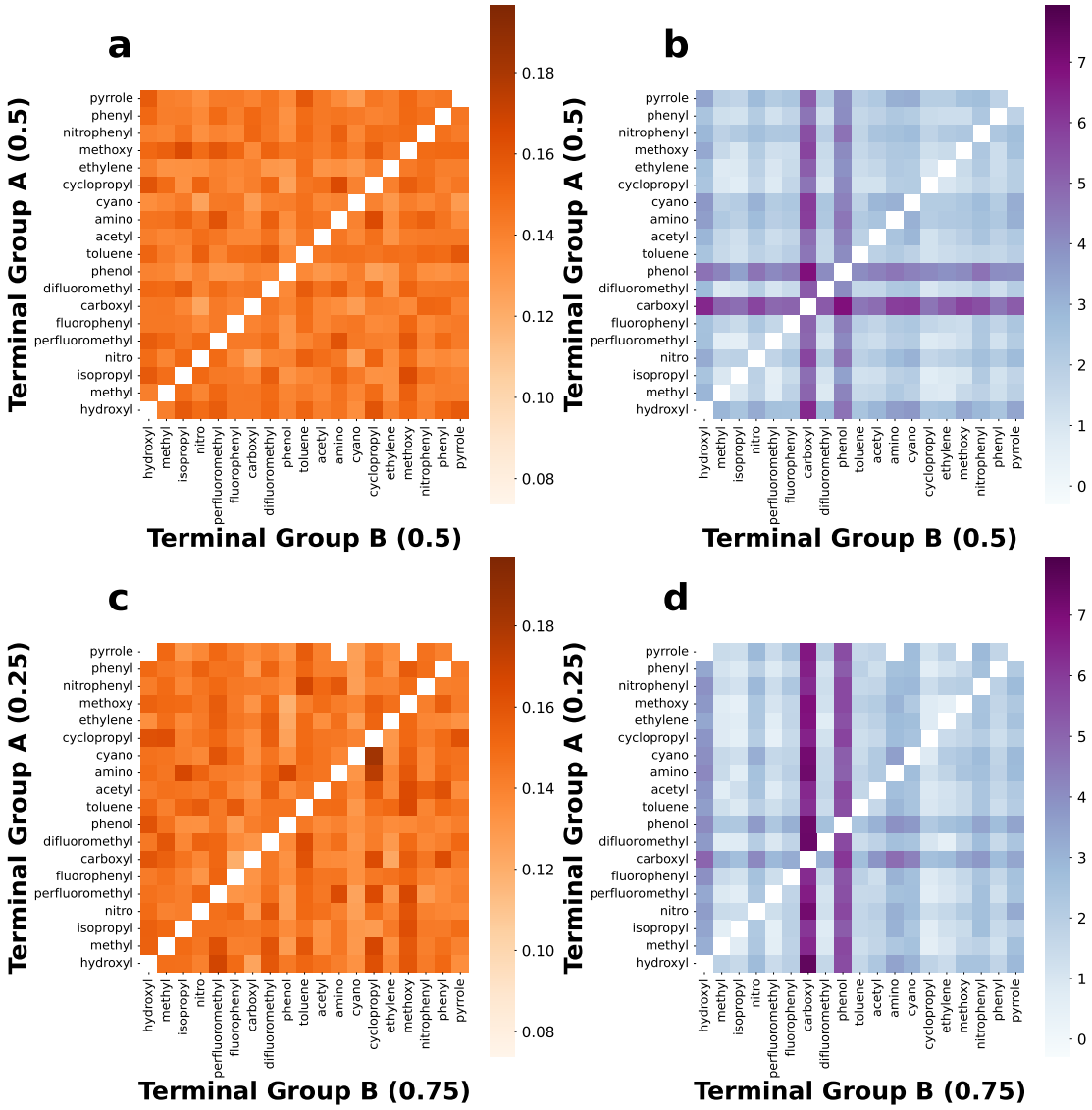

Figure C.12: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only amino terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.13 Cyano Terminated Monolayer

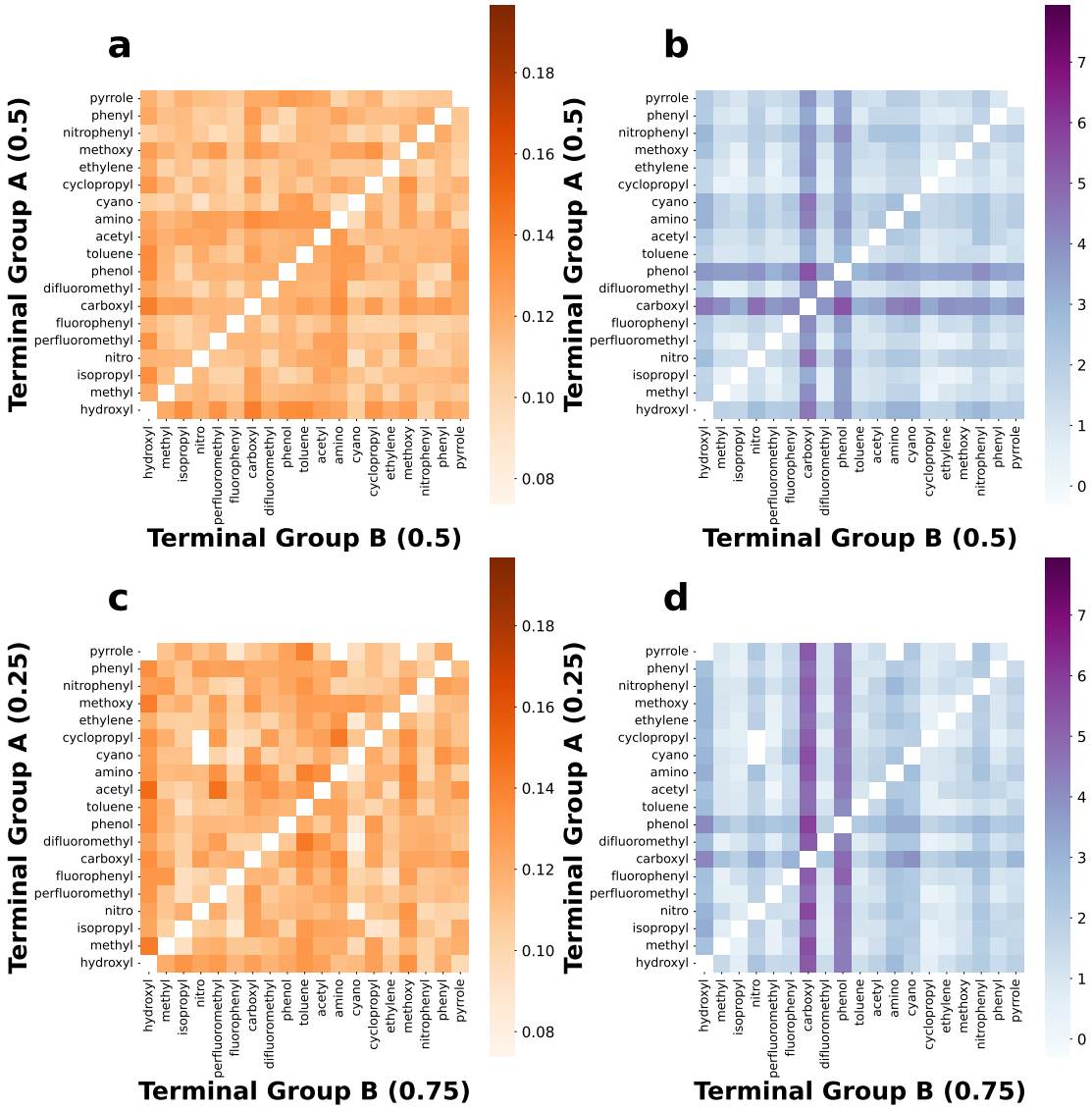

Figure C.13: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only cyano terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.14 Cyclopropyl Terminated Monolayer

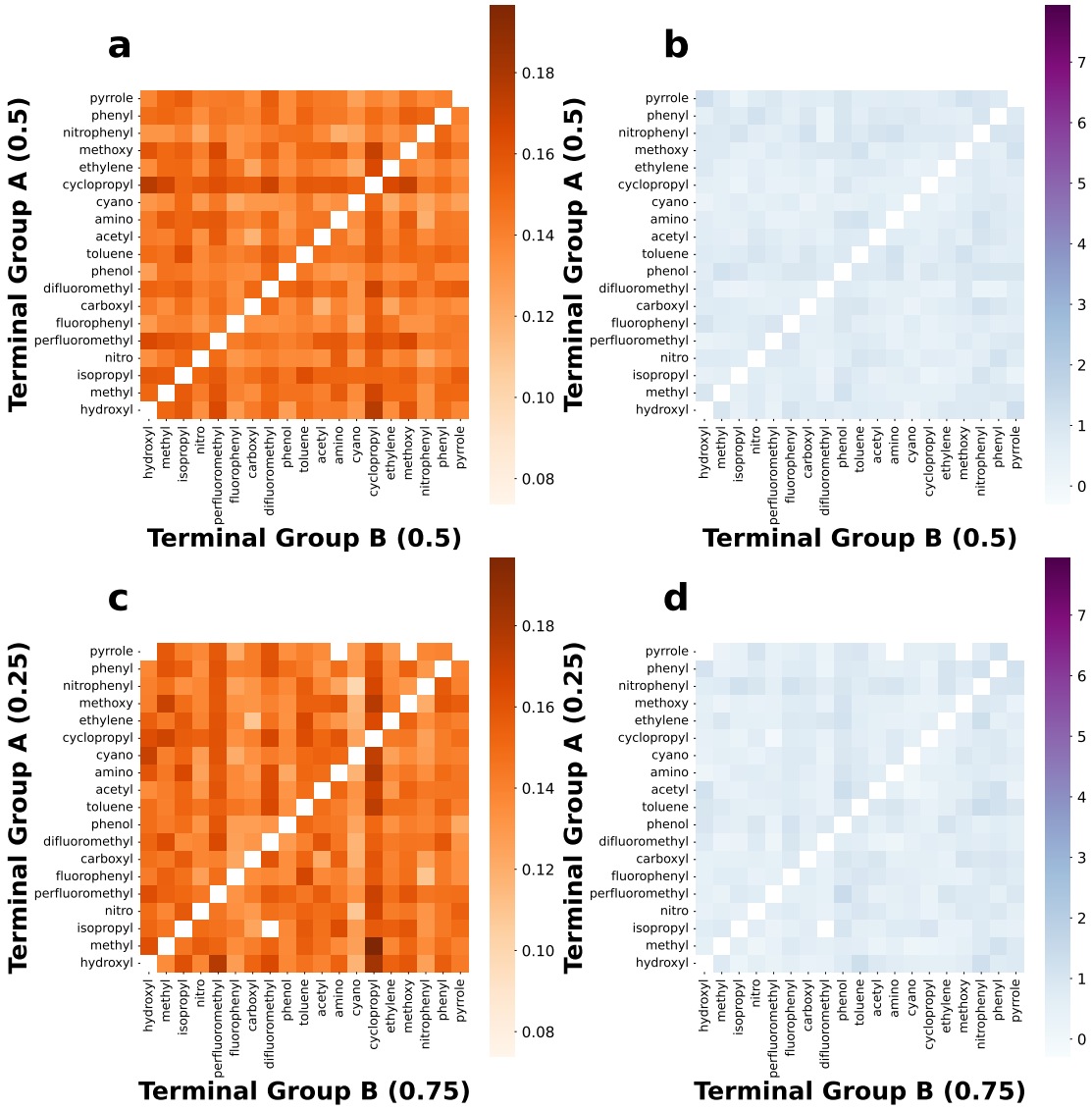

Figure C.14: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only cyclopropyl terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.15 Ethylene Terminated Monolayer

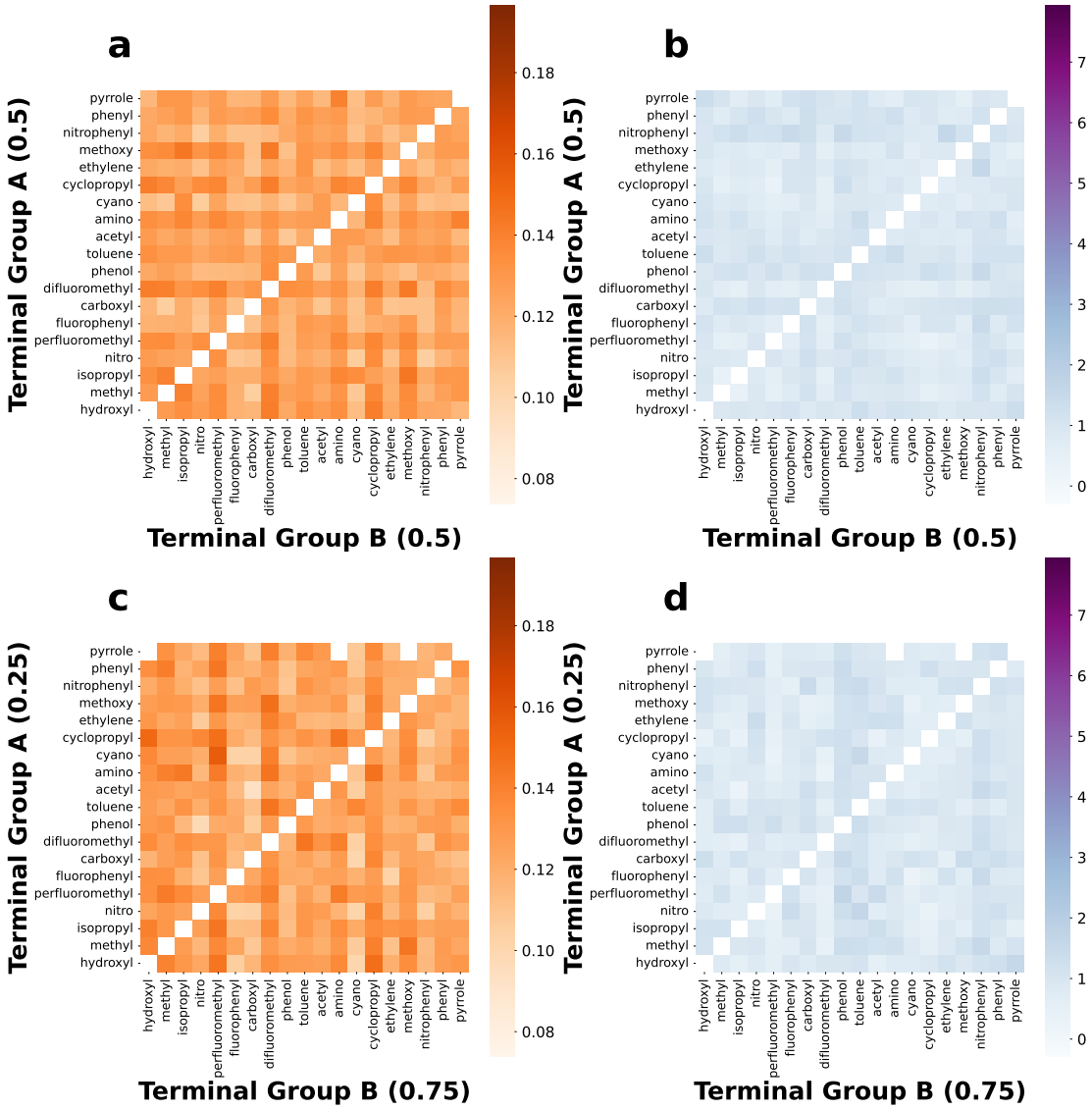

Figure C.15: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only ethylene terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.16 Methoxy Terminated Monolayer

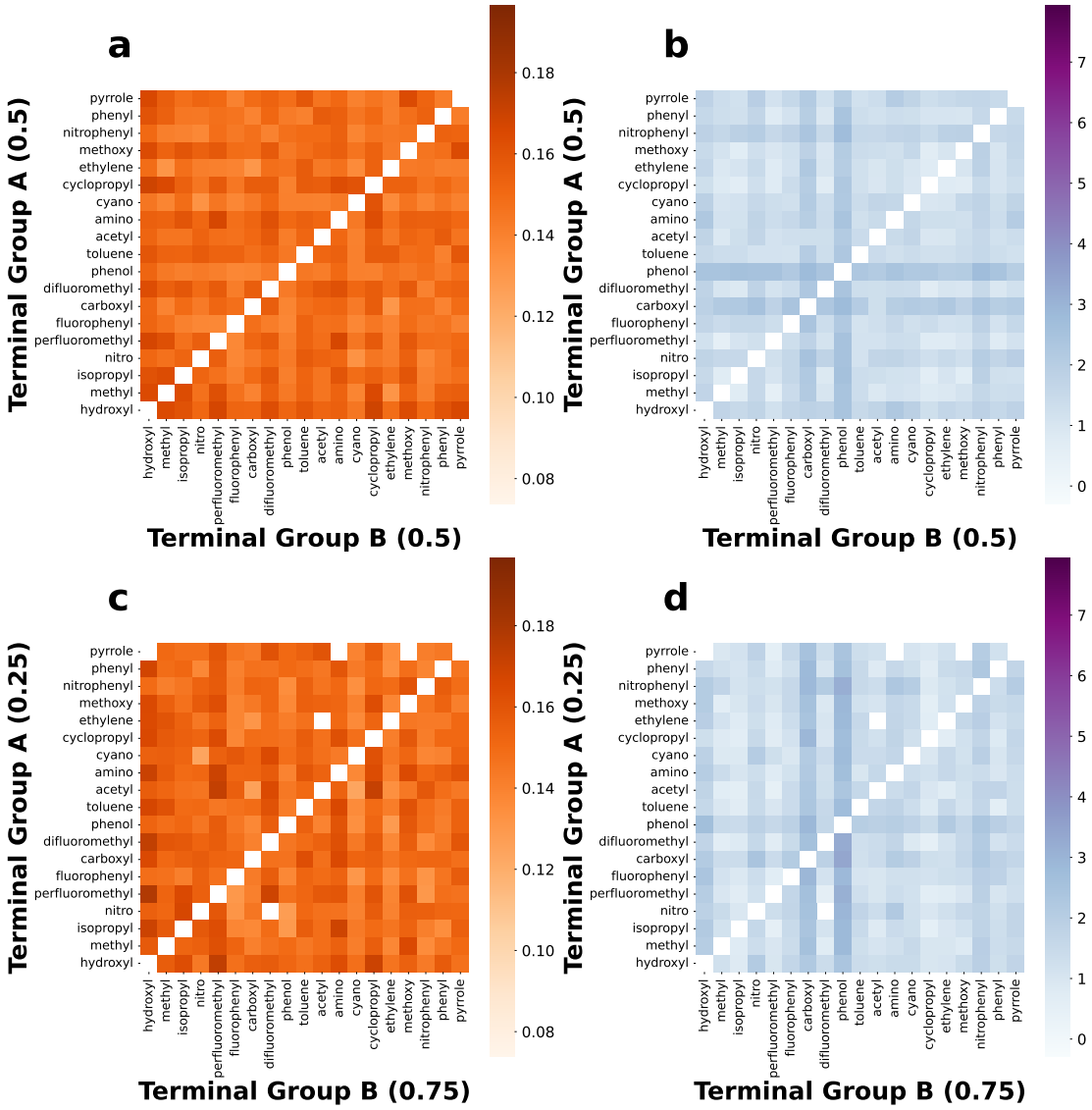

Figure C.16: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only methoxy terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.17 Nitrophenyl Terminated Monolayer

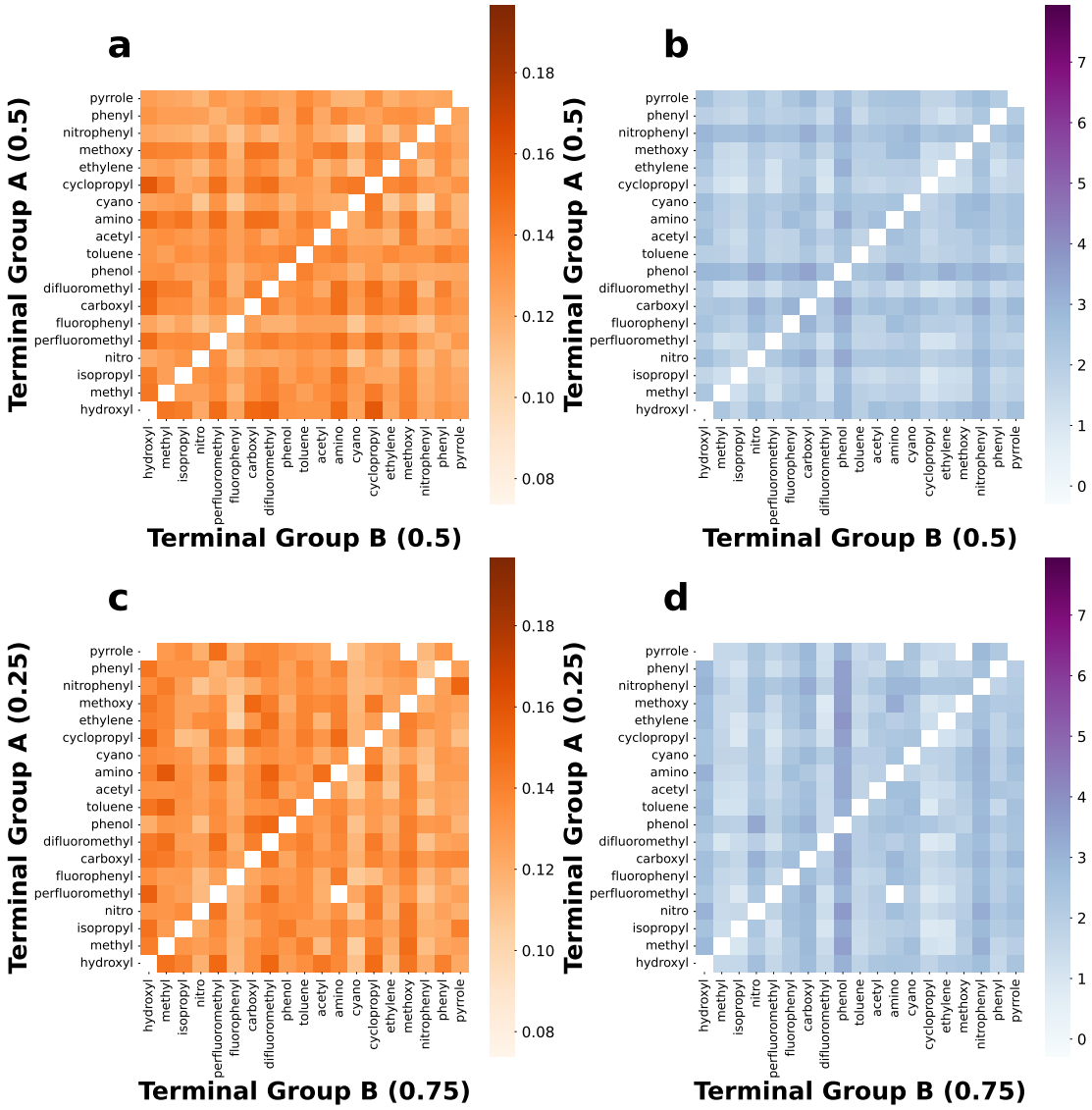

Figure C.17: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only nitrophenyl terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.18 Phenyl Terminated Monolayer

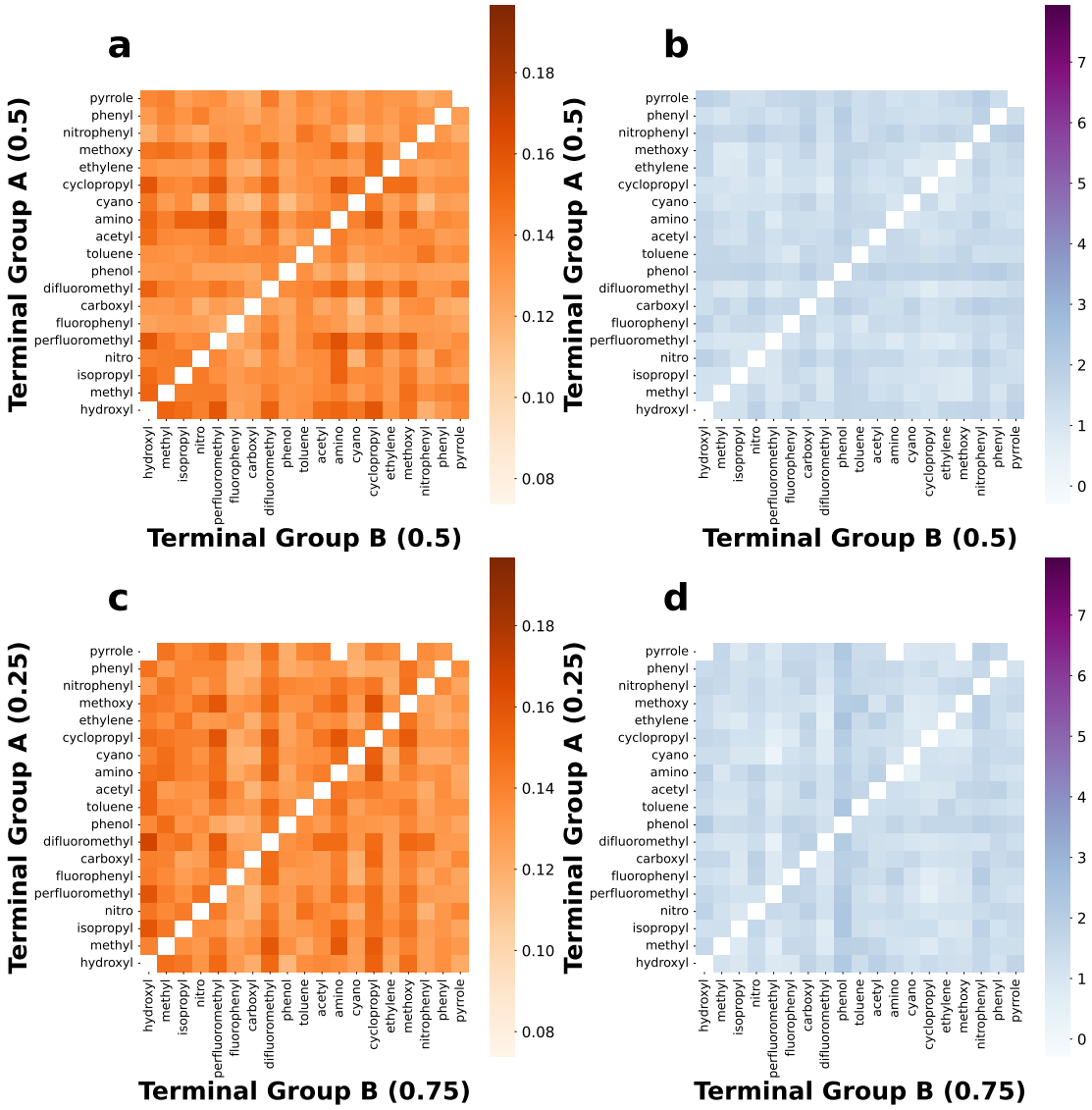

Figure C.18: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only phenyl terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

C.19 Pyrrole Terminated Monolayer

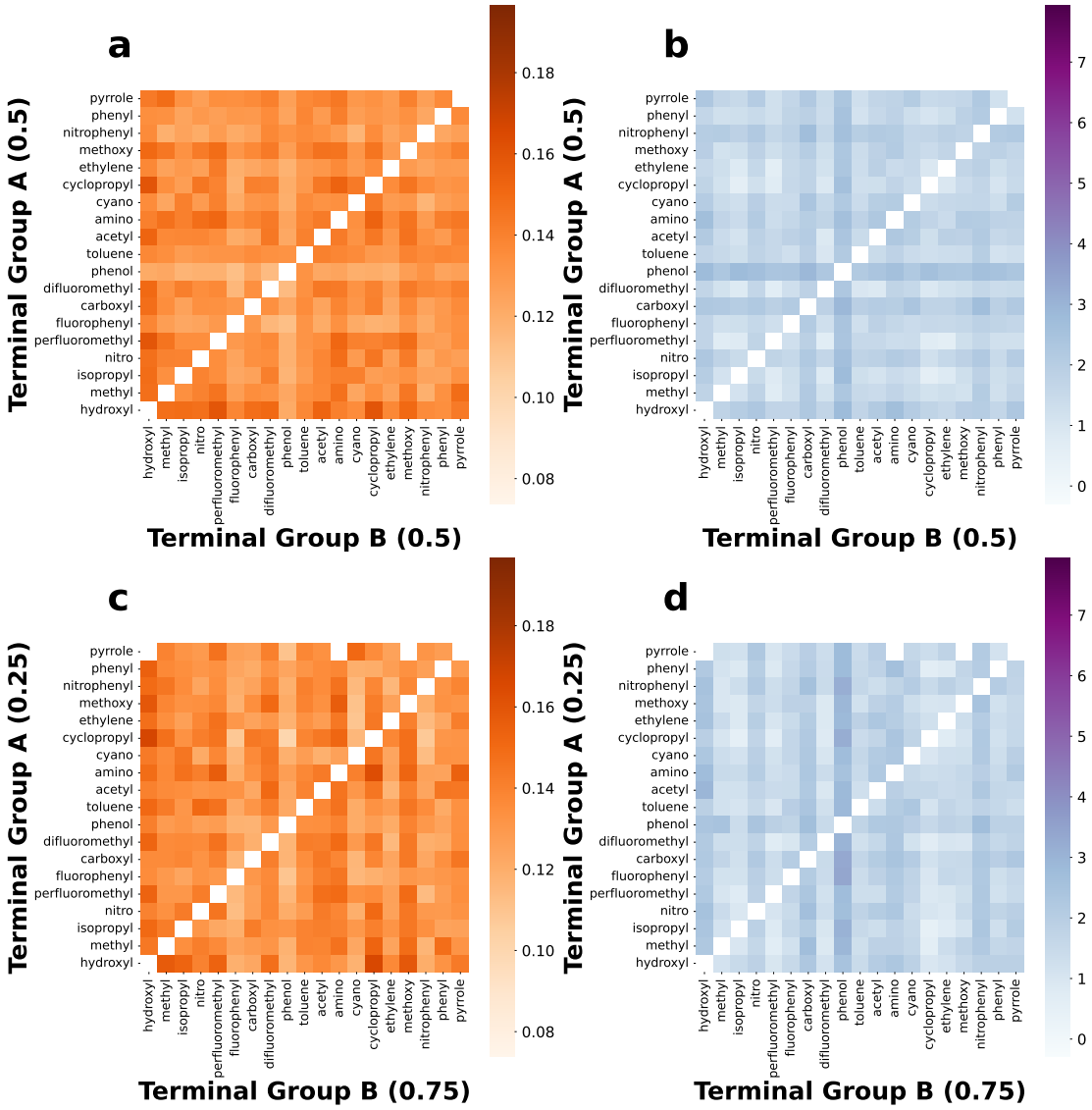

Figure C.19: Heatmaps showing the COF (a, c) and F_0 (b, d) of systems whose bottom monolayer consists of only pyrrole terminal group (chemistry C in Figure 2.2 a), while the top monolayer is a mixture of two different terminal groups (group A and B in Figure 2.2 a). Their relative composition annotated in each individual figure

Appendix D

Molecular Descriptors

Table D.1: Molecular descriptors from RDKit .

Molecular descriptor	Description	Category
Approximate Surface Area	Approximation of molecular surface area using the approach defined by Labute[57]	Size
Asphericity	Measure of molecular shape (from Baumgartner[58]); $A = 0$ for spherical shape, $A = 1$ for highly prolate shapes, and $A = 0 : 25$ for oblate shapes	Shape
Balaban J	Related to connectivity, degree of branching[59]	Complexity
Bertz CT	Measure of molecular complexity through connectivity[60]	Complexity
Chi0, Chi1	Connectivity indices[61]	Complexity
Chi0n - Chi4n	Connectivity indices over various molecular fragments (0=atoms, 1=one bond fragments, 2=two bond fragments, etc.)[61]	Complexity
Chi0v - Chi4v	Valence connectivity indices (0=atoms, 1=one bond fragments, 2=two bond fragments, etc.)[61]	Complexity
Eccentricity	Shape descriptor calculated from the inertiamatrix (0=spherical, 1=linear), from Arteca[62]	Shape
Hall-Kier alpha	Modifying term for kappa descriptors, related to shape/flexibility[63]	Shape
Hall-Kier kappa1	Alpha-modified topological shape descriptor; related to complexity/number of cycles (rings) in the bond graph[63]	Shape
Hall-Kier kappa2	Alpha-modified topological shape descriptor; related to degree of star-like bond graph vs.linearity[63]	Shape
Hall-Kier kappa3	Alpha-modified topological shape descriptor; related to ”centrality” of branching[63]	Shape
Hydrogen bond factor	Developed in Summers *et al.*[12] work; related to ability for formation of inter-monolayer hydrogen bonds	Charge distribution/ Misc.
IPC	Complexity/connectivity descriptor estimated from adjacency matrix of bond graph[64]	Complexity
Inertial shape factor	Characterization of molecular shape from principal moments of inertia (pm2/(pm1 *pm3), where pm1–3 are the three principal moments), from Todeschini and Consoni[63]	Shape
logP	Octanol - water partition coefficient estimated through the method of Wildman and Crippen;[65] measure of hydrophobicity	Charge distribution/ Misc.
Molar refractivity	Estimation of molecular polarizability; calculated through the method of Wildman and Crippen[65]	Size
Molecular weight	-	Size

Molecular weight (heavy atoms)	Molecular weight excluding hydrogens	Size
Normalized principal moments ratios (NPR1, NPR2)	Used to characterize molecular shape, from Sauer and Schwarz[66]	Shape
Number of heavy atoms	Number of non-hydrogen atoms	Size
Number of rotatable bonds	-	Size/Shape
Number of valence electrons	-	Size
Plane of best fit	Measure of molecular planarity (0=planar, in-creasing with less planarity)[67]	Shape
Principal moments of inertia (PMI1, PMI2, PMI3)	Three principal moments of inertia for the molecule (1=smallest, 3=largest)	Shape
Radius of gyration	(From Arteca[62]) Characterizes molecular shape, specifically, elongation	Shape/Size
Sphericity	Measure of molecular shape (0=spherical, 1=flat), from Robinson *et al.*[68]	Shape
Topological polar surface area	Estimation of surface area of only polar atoms, from Ertl *et al.*[69]	Charge distribution
Total hydrophobic VSA	Sum of SA contributions from atoms with $-0.20 \leq q < 0.20$	Charge distribution
Total negative van der Waals surface area (VSA)	Sum of SA contributions from atoms with $q < 0.0$	Charge distribution
Total negative polar VSA	Sum of SA contributions from atoms with $q < 0.20$	Charge distribution
Total polar VSA	Sum of SA contributions from atoms with $\|q\| > 0.20$	Charge distribution
Total positive VSA	Sum of SA contributions from atoms with $q > 0.0$	Charge distribution
Total positive polar VSA	Sum of SA contributions from atoms with $q \geq 0.20$	Charge distribution
Fractional hydrophobic VSA	Total hydrophobic VSA / Total VSA	Charge distribution
Fractional negative VSA	Total negative VSA / Total VSA	Charge distribution
Fractional negative polar VSA	Total negative polar VSA / Total VSA	Charge distribution
Fractional polar VSA	Total polar VSA / Total VSA	Charge distribution
Fractional positive VSA	Total positive VSA / Total VSA	Charge distribution
Fractional positive polar VSA	Total positive polar VSA / Total VSA	Charge distribution