High-throughput screening of tribological properties of monolayer films using molecular dynamics and machine learning

Cite as: J. Chem. Phys. 156, 154902 (2022); https://doi.org/10.1063/5.0080838
Submitted: 03 December 2021 • Accepted: 21 February 2022 • Accepted Manuscript Online: 21 February 2022 • Published Online: 19 April 2022

© 2022 Author(s).

ARTICLES YOU MAY BE INTERESTED IN

Machine learning investigation of viscosity and ionic conductivity of protic ionic liquids in water mixtures
The Journal of Chemical Physics 156, 154503 (2022); https://doi.org/10.1063/5.0085592

Molecular theory of the static dielectric constant of dipolar fluids
The Journal of Chemical Physics 156, 154111 (2022); https://doi.org/10.1063/5.0079511

Descriptor-free unsupervised learning method for local structure identification in particle packings
The Journal of Chemical Physics 156, 154504 (2022); https://doi.org/10.1063/5.0088056
High-throughput screening of tribological properties of monolayer films using molecular dynamics and machine learning

Cite as: J. Chem. Phys. 156, 154902 (2022); doi: 10.1063/5.0080838
Submitted: 3 December 2021 • Accepted: 21 February 2022 • Published Online: 19 April 2022

Co D. Quach,1,2 Justin B. Gilmer,2,3 Daniel Pert,1,2 Akanke Mason-Hogans,1,2 Christopher R. Iacovella,1,2 Peter T. Cummings,1,2 and Clare M. McCabe1,2,a)

AFFILIATIONS
1 Department of Chemical and Biomolecular Engineering, Vanderbilt University, Nashville, Tennessee 37235, USA
2 Interdisciplinary Materials Science, Vanderbilt University, Nashville, Tennessee 37235, USA
3 Multiscale Modeling and Simulation Center, Vanderbilt University, Nashville, Tennessee 37235, USA
4 Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235, USA
a) Author to whom correspondence should be addressed: c.mccabe@vanderbilt.edu

ABSTRACT

Monolayer films have shown promise as a lubricating layer to reduce friction and wear of mechanical devices with separations on the nanoscale. These films have a vast design space with many tunable properties that can affect their tribological effectiveness. For example, terminal group chemistry, film composition, and backbone chemistry can all lead to films with significantly different tribological properties. This design space, however, is very difficult to explore without a combinatorial approach and an automatable, reproducible, and extensible workflow to screen for promising candidate films. Using the Molecular Simulation Design Framework (MoSDeF), a combinatorial screening study was performed to explore 9747 unique monolayer films (116 964 total simulations) and a machine learning (ML) model using a random forest regressor, an ensemble learning technique, to explore the role of terminal group chemistry and its effect on tribological effectiveness. The most promising films were found to contain small terminal groups such as cyano and ethylene. The ML model was subsequently applied to screen terminal group candidates identified from the ChEMBL small molecule library. Approximately 193 131 unique film candidates were screened with approximately a five order of magnitude speed-up in analysis compared to simulation alone. The ML model was thus able to be used as a predictive tool to greatly speed up the initial screening of promising candidate films for future simulation studies, suggesting that computational screening in combination with ML can greatly increase the throughput in combinatorial approaches to generate in silico data and then train ML models in a controlled, self-consistent fashion.

Published under an exclusive license by AIP Publishing. https://doi.org/10.1063/5.0080838

INTRODUCTION

Monolayer films have shown promise as a means of reducing friction and wear of mechanical devices with nanoscale surface separations (e.g., nano- and micro-electromechanical systems, NEMS and MEMS, respectively). Such films are highly tunable through modification of their terminal group chemistries, backbone chain lengths, backbone chemistries, and film compositions, all of which have been demonstrated to impact their tribological effectiveness along with other properties, such as durability, solvent interactions, and thermal response. This tunability presents a rich chemical parameter space that can be explored for the optimization of film properties as well as gleaning useful information about the quantitative structure–property relationship (QSPR) of these systems to develop better predictive models for design considerations. Of these various modifications, the chemical and physical characteristics (descriptors) of the terminal groups play a dominant role in the tribological response. For example, Yu et al. showed that phenyl-terminated monolayer thin films yield higher frictional forces than methyl-terminated films, explained by the phenyl groups’ ability to twist and impede movement during shear (these can be thought of as structural molecular descriptors according to QSPR). Hydroxylated and carboxylated thin films have been shown to have high frictional and adhesive forces relative to methyl-terminated films, attributed
to their ability to form inter-monolayer hydrogen bonds during contact (physicochemical descriptors according to QSPR). Similar trends found in molecular dynamics (MD) simulations further support these relationships. Moreover, this interfacial region may feature not just a single terminal group chemistry but instead multiple chemistries. For example, 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 coefficient of friction (COF) compared to the same tip in contact with a methyl-terminated monolayer. Monolayers composed of two or more terminal group chemistries within the same layer at varied relative compositions may also provide a means to further tune and improve performance. For example, computational studies by Lewis et al. for monolayers composed of methyl-terminated alkanes mixed with perfluoroalkanes showed a regime where the COF was reduced compared to either pure component system. There are a multitude of complex relationships between these various chemical/molecular descriptors when translated to monolayer polymer systems as hinted by Le et al.

MD simulations are a useful tool to perform large-scale sweeps of the accessible parameter space to create an in silico self-consistent dataset. Computational examination avoids the need to develop experimental synthesis techniques, which may be non-trivial and time intensive. Simulations can also more effectively reveal the intrinsic properties associated with defect-free films on pristine contaminant-free surfaces. This approach has been utilized to study and optimize various parameters describing the monolayer, such as backbone chain lengths and chain densities. Our recent development of the Molecular Simulation and Design Framework (MoSDeF) combined with the signac framework by Glotzer et al., enables the large-scale screening of soft matter systems, allowing the reproducible initialization and parameterization of systems, and the management of large datasets. MoSDeF has been used to perform large-scale screening studies of soft matter systems in several recent papers as well as used to fully capture the provenance of simulation workflows for increased reproducibility in other studies.

Despite its utility, the vast parameter space to be explored makes brute force computational screening of monolayers impractical. Instead, a promising approach is to combine computational screening with machine learning (ML) techniques in order to accelerate and better direct the exploration of the parameter landscape. That is, use computational screening to gather sufficient data to train predictive ML models (thus minimizing the number of computationally expensive simulations), and subsequently, use the ML models to predict the properties of new systems and guide the screening toward film chemistries with desirable properties. This approach has been utilized in other studies with great success, such as developing predictive models that have errors lower than hybrid Density Functional Theory (DFT) methods, that learn and predict various protein folding events and structures, that use active learning to direct iterative optimizations and uncover optimal target like structures, and to accelerate the discovery of novel monomers and polymers for favorable macroscopic properties. Different ML models and techniques are applied, but all demonstrate a useful approach to leverage in silico data from molecular simulations and experimental data (when available). As the power of in silico data is further realized for ML models and predictive design, being able to rapidly generate, screen, learn, and predict from these data will be powerful tools for computational and experimental researchers alike.

In prior work, we developed a screening framework that enabled computational screening studies to be performed over a multitude of terminal group chemistries for contacting monolayers undergoing shear using non-equilibrium molecular dynamics (NEMD) simulations. Uniform monolayers with 16 different terminal group chemistries were examined with each monolayer terminal group chemistry independently varied, allowing key trends and several chemistry combinations that provided favorable tribological performance, i.e., both low COF and low adhesion force ($F_0$), to be identified. Furthermore, data from 100 different monolayer terminal group combinations were used to develop, train, and test ML models that allow COF and $F_0$ to be predicted with good accuracy solely from the chemistry of the terminal groups expressed as SMILES strings (discussed in detail in the section titled Methods). The success of these models, trained from a relatively small dataset, suggests that this approach can be used to prescreen computational space and accelerate the identification of films with favorable properties, assuming the dataset is diverse enough to minimize the chance of overfitting. However, to determine if tribological performance could be further improved—for example, by mixing terminal groups together within a film—several key questions regarding the use of ML in a predictive capacity for monolayer films remain. Specifically, (1) what is the minimal dataset required in order to adequately train ML models for tribological properties; (2) how transferrable are the ML models to systems with other chemistries and film compositions not included in the training data; and (3) can the models be used to prescreen the design space, and if so, what level of accuracy can be achieved with a reasonable amount of training data?

To address these questions and further probe the tribological design space of thin films and the relationships between terminal group chemistry, thin film composition, and tribology, here, we consider systems in which one monolayer consists of a single unique terminal group and the other monolayer contains a mix of two unique terminal groups, allowing the relative composition of the two groups to be varied. For these designs, a pool of 19 different terminal groups were considered resulting in 9747 unique monolayer combinations when considering the mixing ratio of terminal groups in the mixed monolayer, as will be discussed in detail in the section titled Methods and Fig. 1. In the section titled Methods, we provide an overview of the computational approach, focusing on the simulation workflow, analysis methods, and the ML model. In the section titled Results and Discussion, we present the data generated from the MD screening and identify key terminal groups and combinations associated with improved tribological performance. We then develop and evaluate the dependence of the ML models on the training set used, assessing the effectiveness of using a ML algorithm to predict properties. Finally, we utilize the ML model to demonstrate the feasibility of screening large data spaces (193 131 unique systems, created from 621 chemistries from the ChEBI library), investigating suitable strategies for utilizing ML models to guide future work. All relevant information to reproducibly generate these data and workflow is readily available and adaptable for others to use, following the principle of TRUE (Transferable, Reproducible,
The Journal of Chemical Physics

ARTICLE

FIG. 1. (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 (25:75 and 50:50), 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, fluoro phenyl, nitrophenyl, toluene, and phenol.

Useable by others, and Extensible) simulations described by Thompson et al. See the supplementary material for more details.

METHODS

Simulation models and workflow

In all cases, the simulated system consists of two opposing amorphous silica surfaces, each coated with an alkylsilane monolayer film. Specifically, each surface has dimensions of $5 \times 5 \text{ nm}^2$, constructed using the procedure outlined by Summers et al. and available as an mBuild script provided in the supplemental repository. The silica surfaces have an average surface roughness of 0.11 nm, closely approximating the more computationally intensive synthesis mimic simulation approach by Black et al. 100 alkylsilane chains are chemically bonded to each surface with an in-plane surface density of 4 chains/nm$^2$; this surface density is consistent with prior computational studies and experiments that estimate chain surface densities to be between 4.0 and 5.0 chains/nm$^2$. Each alkylsilane chain is composed of a fully saturated 17 carbon backbone that is capped with a terminal group that can be easily varied computationally [see Fig. 1(b)]. The remaining undercoordinated oxygen molecules at the surface are changed to hydroxyl groups to mimic surface oxidation. Of the two surfaces in the dual monolayer systems, the bottom surface is homogeneous (singular terminal group), while the top surface contains a mixture of two types of alkylsilane chains, differing by their terminal groups. The mixing ratios for the top monolayers considered in this study are 25:75 and 50:50. The pool of 19 different terminal group chemistries, differing by their terminal groups, was considered for each system. Of the 12,996 systems considered, composition ratios studied, the three normal loads, and three replicates considered for each system. In total, 9747 unique combinations (19 $\times$ 19 $\times$ 18 of 25:75 systems $+ \frac{1}{2} \times 19 \times 18$ of 50:50 systems) were considered. We also note that a small subset of simulations (less than 1% of the total) failed to complete due to unstable initial configurations, but in all cases, each unique system composition reported includes at least three replicates.

Each monolayer system was prepared using the MoSDeF software suite (see the supplementary material for additional information). The initialization of the monolayer structure is encapsulated as an mBuild recipe, which preserves the entire process used to construct the monolayer structure. The foyrer library was used to atom type and parameterize each system with the Optimized Potential for Liquid Simulation—All Atoms (OPLS-AA) forcefield. Parameters for the alkylsilane chains were taken from GROMACS 5.1. and those for the silica surface were taken from the work of Lorenz et al. The force field XML (extensible markup language) file used by foyrer is provided in the supplementary material and can be accessed from the supplementary repository. The project workflow as a whole was managed using the signac framework. The use of MoSDeF in addition to the signac framework ensures that all scripts and input parameters used to initialize the systems, submit the systems for simulation, and analyze the systems are TRUE (Transparent, Reproducible, Useable by Others, and Extensible). All scripts and parameter files are available in the associated GitHub repository (see the supplementary material).

MD simulations were performed using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) and GROMACS simulation engines. LAMMPS was solely used to relieve the system of initial high-energy configurations and possible overlaps. The more stable structure generated was then fed to 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 Nöse–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...
The coefficient of friction (COF), $\mu$, and adhesion force, $F_0$, were calculated from the last 5 ns of the simulation trajectory for each system under shear using the modified version of Amontons’ law of friction given by

$$F_f = F_0 + \mu \times F_N.$$  \hspace{1cm} (1)

Here, $F_f$, $F_0$, $\mu$, and $F_N$ represent the frictional force, the adhesive force, the coefficient of friction, and the normal force, respectively. \(^6\)

A linear regression of the average friction force (ordinate) vs normal load (abscissa) can be used to calculate the COF from the slope and $F_0$ from the intercept of the regression line with the ordinate axis. The friction force is calculated by summing all the forces in the direction of shear on one of the monolayers every 1 ps and averaged over the last 5 ns of the simulation.

**Calculation of tribological properties**

The coefficient of friction (COF), $\mu$, and adhesion force, $F_0$, were calculated from the last 5 ns of the simulation trajectory for each system under shear using the modified version of Amontons’ law of friction given by

$$F_f = F_0 + \mu \times F_N.$$  \hspace{1cm} (1)

Here, $F_f$, $F_0$, $\mu$, and $F_N$ represent the frictional force, the adhesive force, the coefficient of friction, and the normal force, respectively. \(^6\)

A linear regression of the average friction force (ordinate) vs normal load (abscissa) can be used to calculate the COF from the slope and $F_0$ from the intercept of the regression line with the ordinate axis. The friction force is calculated by summing all the forces in the direction of shear on one of the monolayers every 1 ps and averaged over the last 5 ns of the simulation.

**Machine learning model**

The dataset is analyzed using the random forest regressor as implemented in the scikit-learn library, consistent with our prior work in Summers et al.\(^8\)\(^5\)\(^8\)\(^9\)\(^6\)\(^0\)\(^6\)\(^1\) A training set, a set of input parameters, and expected outputs, the random forest ensemble model will create a series of decision trees, each generated from a sub-sample of the training dataset.\(^5\)\(^9\)\(^6\)\(^0\)\(^6\)\(^1\) The trained random forest model then makes predictions by averaging the predicted values from the component decision trees. Each predictive model will rank the importance of each of the input parameters based on how a given input affects the final prediction. This ranking unveils information regarding properties that play an important role in predicting the tribological properties of the monolayer (feature importance), making this method 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.\(^6\)\(^0\) In the forest, each decision tree is allowed to expand until all leaves are pure (choosing splits that decrease impurity defined by the Gini impurity). All models used the mean squared error (MAE) as the error criterion during training. Each random forest model and its subsequent decision trees are trained with 32–42 features, which are molecular descriptors calculated through RDKit.\(^6\)\(^1\) This setup is consistent with the previous study by Summers et al.,\(^6\)\(^1\) 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. An effort to optimize the parameters of the random forest model resulted in insignificant improvement in model accuracy, and thus, the original values were retained for better comparability with prior work; further details can be found in the supplementary material (see Figs. S1–S3).

The chemical and physical input parameters for the ML model are supplied by the RDKit cheminformatics library.\(^5\)\(^9\) The COF and $F_0$ calculated from the simulations are the expected outputs (i.e., targeted properties) for the random forest ensemble to predict. The training procedure of these predictive models is adapted from that described in previous work.\(^6\) Briefly, each of the systems in the training set can be represented by a set of SMILES strings,\(^6\)\(^1\) describing the terminal group chemistry. Each terminal group is represented by two SMILES strings: one of a hydrogen-capped structure and one of a methyl-capped structure. The SMILES strings are used to calculate molecular descriptors that characterize the chemical and physical properties of chemical structures via the RDKit cheminformatics library.\(^6\)\(^1\) These descriptors fit into four categories: size (e.g., molecular weight), shape (e.g., inertial shape factor), complexity (e.g., degree of branching), and charge distribution (e.g., topological polar surface area). The SMILES string of the hydrogen-capped terminal group is used to calculate descriptors relating to shape, while the SMILES string of the methyl-capped terminal group is used to calculate the remaining descriptors. While shape characteristics can be sufficiently modeled with a hydrogen-capped structure, properties that involve charge distribution among others are better represented if they mimic the actual structure the terminal groups are attached to; a methyl terminus was found to be a sufficient approximation of the alkyl chain for these measurements.\(^6\) Through this process, each chemical structure is represented by 53 descriptors, summarized in the supplementary material (see Table S1).

The molecular descriptors for the top and bottom monolayers are first calculated independently. Descriptors for the top monolayer with two terminal groups are the weighted average (by relative composition) of its component terminal groups’ descriptors, while descriptors for the bottom monolayer are the molecular descriptors of its singular terminal group. This representation can have limitations when used to describe monolayers since it does not encode information regarding connectivity of constituent chains and distribution patterns.\(^6\)\(^3\) However, since we are mainly interested in the contribution of different terminal group chemistries, making up the inter-monolayer regions/interfaces, our method of calculating the molecular descriptor “fingerprint” was found to be sufficient to encapsulate information for the region of interest with an assumption that the two terminal groups in the top monolayer are randomly distributed. These are then combined, storing the mean and minimum of each descriptor as the “molecular fingerprint” of the entire system, totaling 106 descriptors [(53 metrics) * (2 corresponding to min and mean)], which has been demonstrated in previous work to encapsulate the most important features of these systems.\(^8\) These “molecular fingerprints” later undergo a dimensionality reduction step that removes descriptors whose values have low variance and reduces highly correlated descriptors. Specifically, descriptors whose values are at least 90% correlated will be reduced to only one attribute (descriptors are sorted alphabetically), while descriptors whose variance is below 2% are also removed. This step reduced the number of descriptors (or effective fingerprints) of each system to be between 32 and 45 with a mean of 37 ± 3. The number of effective fingerprint descriptors decreases as the size of the training data increases since some correlations may not manifest with a smaller dataset. This dimensionality reduction process is
As shown in Fig. 3, the dataset can be further broken down by the composition ratio with the 50:50 and 25:75 mixing ratios considered separately, e.g., 5050-test/5050-train and 2575-test/2575-train. To examine the role of the training set size on the performance of the ML model, the total-train and 5050-train sets are further subdivided to create training sets with fewer data points. Since the COF and $F_0$ of these films have been shown to have little correlation, they are considered independently in the development of the ML models. To ensure that the entire range of COF and $F_0$ are being properly sampled (Fig. 4), each training set is binned based upon the values of the COF or $F_0$ into ten equal size quantiles. Data are then randomly selected from each bin to create training sets of varying sizes. For each training set size, five variations are created that differ only by the random seed used when sampling from the corresponding master training set, e.g., total-train and 5050-train. Distributions of individual training sets are examined to ensure that they resemble the distribution of the full dataset. Each training set is then used to train and create a predictive ML model that predicts either COF or $F_0$.

**RESULTS AND DISCUSSION**

Considering first the results of the high-throughput screening MD simulations, including those performed in the current study and in the work of Summers et al., we identify 22 monolayer designs that provide favorable frictional properties, e.g., those that have low simulated COF and $F_0$ values (see Table I and Fig. 5). This list was created by the intersection of the best 500 systems ranked from lowest to highest COF (values ranging from 0.074 to 0.114) with the best 500 systems ranked from lowest to highest $F_0$ (values ranging from 0.007 to 0.541 nN). We first note that, in general, these designs are in agreement with the conclusion obtained by Summers et al. from a study of a considerably smaller dataset, where it was noted that the COF of monolayers is primarily affected by the shape and size of the terminal group with chemistries of small sizes and simple shapes (e.g., $sp$ hybridization) exhibiting the lowest COF. Summers et al. also noted that $F_0$ is most strongly affected by charge distribution with polarity and hydrogen bonding both...
TABLE I. 22 most-favorable systems determined by the intersection of the top 500 systems ranked by their COF and the top 500 systems ranked by their $F_0$. The COF and $F_0$ mean values and standard deviation (std.) are calculated from the three replicates.

<table>
<thead>
<tr>
<th>Terminal group A</th>
<th>Terminal group B</th>
<th>Terminal group C</th>
<th>A fraction</th>
<th>B fraction</th>
<th>COF - mean</th>
<th>COF - std.</th>
<th>$F_0$ (nN) - mean</th>
<th>$F_0$ (nN) std.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyano</td>
<td>Cyano</td>
<td>Isopropyl</td>
<td>0.5</td>
<td>0.5</td>
<td>0.1032</td>
<td>0.0063</td>
<td>0.4768</td>
<td>0.1075</td>
</tr>
<tr>
<td>Cyclopropyl</td>
<td>Ethylene</td>
<td>Cyano</td>
<td>0.5</td>
<td>0.5</td>
<td>0.1086</td>
<td>0.0142</td>
<td>0.4948</td>
<td>0.2820</td>
</tr>
<tr>
<td>Difluoromethyl</td>
<td>Isopropyl</td>
<td>Cyano</td>
<td>0.5</td>
<td>0.5</td>
<td>0.1100</td>
<td>0.0098</td>
<td>0.5292</td>
<td>0.4261</td>
</tr>
<tr>
<td>Difluoromethyl</td>
<td>Methyl</td>
<td>Cyano</td>
<td>0.5</td>
<td>0.5</td>
<td>0.1128</td>
<td>0.0036</td>
<td>0.4640</td>
<td>0.1585</td>
</tr>
<tr>
<td>Ethylene</td>
<td>Isopropyl</td>
<td>Cyano</td>
<td>0.5</td>
<td>0.5</td>
<td>0.1109</td>
<td>0.0108</td>
<td>0.2161</td>
<td>0.3768</td>
</tr>
<tr>
<td>Ethylene</td>
<td>Perfluoromethyl</td>
<td>Cyano</td>
<td>0.5</td>
<td>0.5</td>
<td>0.1093</td>
<td>0.0119</td>
<td>0.3297</td>
<td>0.3124</td>
</tr>
<tr>
<td>Isopropyl</td>
<td>Methoxy</td>
<td>Cyano</td>
<td>0.5</td>
<td>0.5</td>
<td>0.1113</td>
<td>0.0144</td>
<td>0.5313</td>
<td>0.5019</td>
</tr>
<tr>
<td>Isopropyl</td>
<td>Methyl</td>
<td>Cyano</td>
<td>0.5</td>
<td>0.5</td>
<td>0.1121</td>
<td>0.0182</td>
<td>0.4041</td>
<td>0.3252</td>
</tr>
<tr>
<td>Isopropyl</td>
<td>Perfluoromethyl</td>
<td>Cyano</td>
<td>0.5</td>
<td>0.5</td>
<td>0.1030</td>
<td>0.0181</td>
<td>0.2978</td>
<td>0.1960</td>
</tr>
<tr>
<td>Difluoromethyl</td>
<td>Carboxyl</td>
<td>Isopropyl</td>
<td>0.25</td>
<td>0.75</td>
<td>0.1117</td>
<td>0.0021</td>
<td>0.1944</td>
<td>0.8319</td>
</tr>
<tr>
<td>Difluoromethyl</td>
<td>Ethylene</td>
<td>Carboxyl</td>
<td>0.25</td>
<td>0.75</td>
<td>0.1105</td>
<td>0.0078</td>
<td>0.5375</td>
<td>0.7366</td>
</tr>
<tr>
<td>Difluoromethyl</td>
<td>Methyl</td>
<td>Cyano</td>
<td>0.25</td>
<td>0.75</td>
<td>0.1126</td>
<td>0.0041</td>
<td>0.4967</td>
<td>0.0876</td>
</tr>
<tr>
<td>Ethylene</td>
<td>Isopropyl</td>
<td>Cyano</td>
<td>0.25</td>
<td>0.75</td>
<td>0.1046</td>
<td>0.0184</td>
<td>0.5122</td>
<td>0.1374</td>
</tr>
<tr>
<td>Isopropyl</td>
<td>Cyclopropyl</td>
<td>Cyano</td>
<td>0.25</td>
<td>0.75</td>
<td>0.0898</td>
<td>0.0119</td>
<td>0.4354</td>
<td>0.1522</td>
</tr>
<tr>
<td>Isopropyl</td>
<td>Perfluoromethyl</td>
<td>Cyano</td>
<td>0.25</td>
<td>0.75</td>
<td>0.1077</td>
<td>0.0073</td>
<td>0.3868</td>
<td>0.1580</td>
</tr>
<tr>
<td>Methoxy</td>
<td>Cyano</td>
<td>Ethylene</td>
<td>0.25</td>
<td>0.75</td>
<td>0.1086</td>
<td>0.0053</td>
<td>0.4346</td>
<td>0.4340</td>
</tr>
<tr>
<td>Methyl</td>
<td>Isopropyl</td>
<td>Cyano</td>
<td>0.25</td>
<td>0.75</td>
<td>0.0942</td>
<td>0.0148</td>
<td>0.4181</td>
<td>0.2017</td>
</tr>
<tr>
<td>Perfluoromethyl</td>
<td>Cyclopropyl</td>
<td>Cyano</td>
<td>0.25</td>
<td>0.75</td>
<td>0.1138</td>
<td>0.0070</td>
<td>0.3104</td>
<td>0.1799</td>
</tr>
<tr>
<td>Perfluoromethyl</td>
<td>Ethylene</td>
<td>Cyano</td>
<td>0.25</td>
<td>0.75</td>
<td>0.1067</td>
<td>0.0002</td>
<td>0.5315</td>
<td>0.1965</td>
</tr>
<tr>
<td>Perfluoromethyl</td>
<td>Isopropyl</td>
<td>Cyano</td>
<td>0.25</td>
<td>0.75</td>
<td>0.0947</td>
<td>0.0114</td>
<td>0.3566</td>
<td>0.2037</td>
</tr>
<tr>
<td>Phenyl</td>
<td>Isopropyl</td>
<td>Cyano</td>
<td>0.25</td>
<td>0.75</td>
<td>0.1098</td>
<td>0.0156</td>
<td>0.4825</td>
<td>0.3196</td>
</tr>
<tr>
<td>Toluene</td>
<td>Ethylene</td>
<td>Cyano</td>
<td>0.25</td>
<td>0.75</td>
<td>0.1119</td>
<td>0.0134</td>
<td>0.4907</td>
<td>0.6589</td>
</tr>
</tbody>
</table>

Increasing $F_0$. 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, characteristics that fit with previous work to identify chemistries that can lower the COF and $F_0$ of monolayers. We also note that most systems in Table I are made up of three components and only one system consists of two homogeneous monolayers (system 1 in Table I), which was simulated in the work of Summers et al. This result suggests a slight advantage to having mixed monolayer designs. However, we also recognize that the dataset 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.

Using the simulation data, we now explore combining ML techniques with MD simulations to perform high-throughput screening of monolayer systems. In the work of Summers et al., the random forest regressor algorithm was applied to create predictive ML models to estimate the frictional properties of alkylsilane monolayers capped with different terminal group chemistries. The ML models were trained on simulation data for homogeneous monolayers with 16 distinct terminal groups, resulting in a relatively small dataset, containing only 100 data points. The models were then applied to a test set and compared to the COF and $F_0$ results obtained for the same systems directly from MD simulation to determine the accuracy of the ML models. This comparison can be readily visualized by plotting the tribological properties obtained from the ML models against the values calculated from the MD simulations; the coefficient of determination ($R^2$) and the mean absolute percentage error (MAPE) of the plots are used to quantitatively measure the accuracy of the ML predictions. The $R^2$ is commonly used/reported to quantify the correlation between the simulated and predicted values, and MAPE provides error metrics that scale by the prediction values.
Using the additional simulation data generated herein, we can now better assess the feasibility of using ML to predict tribological properties and determine the amount of data necessary to create models that can make sufficiently precise estimations, as described in the following.

We first train a new set of ML models using 1000 data points from the 5050-train set. The models are then applied to the 5050-test set and, as described in the section titled Methods, compared to the COF and $F_0$ results obtained directly from the MD simulations in order to determine the accuracy of the ML models (see Fig. 6). Results for the ML models trained with the dataset of Summers et al. applied to the 5050-test set are also shown for comparison. When applied to the same testing set, the models of Summers et al. provide $R^2$ values of 0.472 and 0.657 for COF and $F_0$, respectively, compared to 0.822 and 0.899 for COF and $F_0$ from the 5050-train set. While the $R^2$ values are lower for the ML models of Summers et al., it is worth noting that the training dataset did not include any information regarding mixed monolayer compositions; as such, the ML models of Summers et al. still demonstrate impressive efficacy. This point is further demonstrated by their MAPE, where the models of Summers et al. could predict COF of the system with 0.056 (5.6%) error and predict $F_0$ with 0.266 (26.6%) error. These MAPE values are higher but are still comparable with those produced by the new set of models, trained with ten-fold amount of data. Clearly, our prediction of $F_0$ is less accurate in the higher adhesion regime than the lower adhesion regime, as shown in Figs. 6(b) and 6(d), as was also observed in the prior work of Summers et al. This is likely related to the challenges associated with capturing the ability of systems to form hydrogen bonds between contacting layers, although, since our primary goal in this work is to identify systems with low adhesion values, quantitative agreement in the higher adhesion regime is not required, as previously discussed in the work of Summers et al. Nonetheless, this result suggests that ML models trained with limited data could still provide meaningful estimation 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. It should be noted, however, that this relationship

FIG. 5. Distribution of simulated systems based on their COF and $F_0$ values. The 22 most-favorable systems, listed in Table I, correspond to data points confined within the red dashed box in the lower left quadrant of the figure.

FIG. 6. Predicted-versus-simulated plots for COF and $F_0$ for models trained with 100 simulation data points for uniform monolayers from the dataset of Summers et al. (a) and (b) and trained with 1000 data points randomly chosen from the 5050-train set (c) and (d). The dotted line in the middle represents perfect prediction ($y = x$). The outer two lines represent 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. For each system (data point), the predicted properties are averaged from the five predictions made by the five ML model replicates, and the simulated properties are averaged from at least three simulation replicates.
could be solely related to these monolayer systems and specifically to non-equilibrium shearing and may not be applicable to other non-equilibrium studies. The prediction deviation plots for these models are shown in Fig. 7. In general, we see 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 [see Figs. 7(a) and 7(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. This trend appears to be correlated with the size and distribution of the training set provided, with the trend becoming less apparent as the size of the training dataset is increased [see Figs. 7(c) and 7(d)]. 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 that the predictive ML models can be more confidently used to screen over potential film candidates for possible applications. We also note that while the \( R^2 \) values for COF models are substantially smaller than those of \( F_0 \) models, which might suggest the latter models outperform their COF counterparts, their MAPE values indicate the opposite, where the \( F_0 \) models exhibit significantly greater percentage errors. This disparity could be attributed to the difference in the range of these two properties; while COF values span a small range of values from roughly 0.085 to 0.2, \( F_0 \) can take values from ~0 to 8 nN (see Fig. 4), which may affect how these metrics are calculated. Hence, it is important to recognize that neither \( R^2 \) or MAPE values can directly relate to the predictive ability of the COF and \( F_0 \) models although they can still be used to compare the performance of ML models of a similar type. Comparison of the feature importance of the two models at this point can be misleading since the two sets of models are trained with feature vectors of various sizes and components, as discussed in the section titled Methods. An extensive comparison of the feature importance of different models in this study is further discussed in the supplementary material (see Figs. S4–S7).

We further examine the quality of the ML model estimations as a function of the training set size by gradually increasing the amount of data used to train the ML models. We start with only using data from the 50/50-train set; by doing so, we can later evaluate the transferability of the model to the 25:75 systems, i.e., can these predictive models provide similarly precise estimation of the frictional properties of systems with different designs. The \( R^2 \) and MAPE values for models trained on datasets of increasing sizes from the 50/50-train set are reported in Fig. 8. The results for each dataset size are calculated from five models, each differing by the random seeds used when sampling from the 50/50-train set, and the standard deviation of the predictions of the five models is represented as error bars.

FIG. 7. Deviations in predictions of COF and \( F_0 \) from ML models trained with 100 simulation data points for uniform monolayers from Summers et al.\(^8\) [(a) and (b)] and trained with 1000 simulation data points randomly chosen from the 50/50-train set [(c) and (d)].
The $R^2$ and MAPE of estimations made by the models of Summers et al. are also shown in dotted lines for reference. From Fig. 8(a), we can see that the accuracy of the ML model predictions improves rapidly as the training set size is increased with the ML predictions roughly plateauing between 1000 and 1500 data points for which the $R^2$ values for COF are $0.799 \pm 0.007$ and $0.835 \pm 0.011$ and for $F_0$ are $0.885 \pm 0.001$ and $0.907 \pm 0.007$, respectively. Similarly, in Fig. 8(b), the MAPEs of both COF and $F_0$ models also decrease gradually from $0.0546 \pm 0.003$ ($5.46\% \pm 0.3\%$) for COF and $0.259 \pm 0.016$ ($25.9\% \pm 1.6\%$) for $F_0$ at 100 data points and level-off near 1000 data points at $0.0339 \pm 0.001$ ($3.39\% \pm 0.01\%$) for COF and $0.175 \pm 0.003$ ($17.5\% \pm 0.3\%$) for $F_0$. While the predictive ability of the models does increase with the size of the training set beyond 1000 datapoints, the gains in accuracy are much less significant, suggesting one could achieve sufficiently accurate ML models even with a modest amount of data.

We now examine the transferability of the ML algorithms in terms of their ability to predict the frictional properties of systems with different designs, i.e., systems with a different mixing ratio on the top monolayer in the testing set than in the training set. Results from applying the ML models described above, trained solely with data from the 5050-train set and the dataset of Summers et al., are shown in dotted lines for reference. From Fig. 8(a), we observe that the $R^2$ values for COF and $F_0$ increase rapidly before plateauing again at a training set size of ~1000 points with values of $0.647 \pm 0.001$ and $0.848 \pm 0.007$ for COF and $F_0$, respectively. Similar trends are observed in Fig. 8(b), where the MAPEs of both COF and $F_0$ models rapidly decrease up until 1000 data points before leveling-off with an error of $0.0521 \pm 0.0$ ($5.21\% \pm 0.0\%$) for COF and $0.26 \pm 0.003$ ($26.0\% \pm 0.3\%$) for $F_0$. While the accuracy is lower, i.e., lower $R^2$ and higher MAPE, than that observed for the 5050-test set (see Fig. 8), the agreement is promising, considering the ML models were not trained with data at these composition ratios. The plateau of the accuracy of the models is important as it shows that improvements in accuracy of the models with larger datasets (as shown in Fig. 9) do not necessarily manifest themselves when the model is transferred to compositions outside of the original training set. Moreover, from Figs. 8 and 9, we notice the 5050-train models trained with 100 data points also exhibit slightly better accuracy than the model trained with the dataset of Summers et al., likely due to...
the inclusion of mixed monolayer systems in their training sets. We note that for random forest regressors, the variety of data is more important in determining the quality of the predictions compared to the amount of data beyond a certain point, which is dependent on the complexity of the systems of interest; this point is further examined for the specific systems studied herein in the supplementary material (see Figs. S8–S10). That is, there may be limited utility in using large training sets when trying to develop ML models to pre-screen systems outside of the design space of the original training set. However, generating a set of systems with well distributed properties can be challenging and hard to estimate a priori and, hence, may require more thoughtful design of the initial screening space as well as a more active learning approach to direct the screening space as the initial data are used to train the models.

The results in Fig. 9 suggest that the ML models are likely effective in predicting frictional properties of systems with design variations, i.e., despite the noticeable decline in performance, the models could likely be used as a high-level screen to sieve the parameter space. To further examine the capability of ML models in shortening the list of potential candidates to be simulated/synthesized, we examine the ability of the models to identify systems that exhibit favorable tribological performance (i.e., low COF or \( F_0 \)). To quantify the ability of the model to predict favorable solutions, we calculate the intersection of the top performing systems predicted by the ML model and those via simulation; an ideally performing ML model would be able to identify all of, or a majority of, the best performing systems determined through simulation. The ability of the ML model to accurately predict the systems with the favorable properties can be considered to be proportional to the percentage of the overlapping systems compiled from MD simulation and ML prediction. An overlap of 100\% indicates complete agreement between the two methods, i.e., ML and MD, while a low overlap value indicates lower agreement, and by extension, poorer predictive ability of the ML models. We note that this metric describes the ability of the ML model to accurately capture relative differences between systems of interest and does not necessarily require quantitative agreement between the ML model and corresponding MD simulations.

First, we consider the ability of 5050-\( \text{train} \) models in determining the best performing systems in the 5050-\( \text{test} \) set in Fig. 10(a). Systems in the set are first sorted separately by the numerical value of their COF and \( F_0 \) calculated from the simulations; the top 15% of these systems (i.e., systems with the lowest COF or \( F_0 \) values) of each set are considered, corresponding to 100 chemistries for 5050-\( \text{test} \) and 193 for 2575-\( \text{test} \). These lists are then compared to the top 15% of systems predicted by the ML models as a function of the training set size, and the overlapping percentages are calculated. For COF, as the training set size of the model increases, so too does the fraction of top performing solutions predicted, achieving 74.4\% ± 0.8\% accuracy for models trained using 1000 data points and 83.0\% ± 0.9\% accuracy for models trained with 2500 systems; adhesion shows a weaker dependence on the training set size, reaching 67.6\% ± 1.1\% at 1000 data points and 71.2\% ± 1.2\% at 2500 data points [see Fig. 10(a)]. Putting these results into perspective, if we had utilized the 5050-\( \text{train} \) model of 1000 data points to predict systems with the best performing properties from the 5050-\( \text{test} \) set and only simulated those in the top 15%, we would have reduced the total number of additional screening simulations by 85\%, while still identifying 74.4\% of the best performing systems as ranked by COF or 83\% of the best performing systems ranked by \( F_0 \). On the other hand, if we reduced the number of systems to be simulated at random, we would only expect to detect 15\% of the best performing systems, ranked by COF or \( F_0 \). In other words, this approach can significantly increase the odds of finding systems with the top performing tribological properties. We conduct the same analysis of the ability of the ML models to identify the best performing systems in the 2575-\( \text{test} \) set, i.e., focusing on the transferability of the models, and show the results in Fig. 10(b). We observe predictions with an accuracy of roughly 61\% for both the best performing systems ranked by COF or \( F_0 \), almost independent of the training set size [see Fig. 10(b)]. Even though these accuracies are not as high as the models used on the 5050-\( \text{test} \) set, they are still considerably higher than the 15\% accuracy we would have expected if selecting systems at random. This suggests reasonable efficacy of using this approach to prescreen design space. Even for models trained with limited amounts of data, i.e., models trained with 500 data points whose accuracies are 61.3\% ± 1.3\% for COF models and 62.3\% ± 2.0\% for \( F_0 \), the predictions made should still be useful enough for prescreening and provide focused guidance to perform the next round of simulations, while reducing the computational costs substantially.

![FIG. 10. Intersection between the top 15% performing systems predicted by ML models at various training set sizes and top 15% performing systems calculated by MD simulations (in silico data) of the 5050-\( \text{test} \) set (a) and the 2575-\( \text{test} \) set (b). The systems are ranked by COF (blue circles) or \( F_0 \) (red squares). The dashed, horizontal lines, colored to correspond with its respective property, show the predictive ability of the models trained using the dataset of Summers et al.](image-url)
In Fig. 10, we have compared an equipercentile of the top performers determined through different techniques, i.e., MD and ML. This approach, however, can potentially result in missing out potential candidates, e.g., a model that has an accuracy of 60% in determining the true best 15% systems will miss potentially 40% of the candidates. In practice, considering a larger list of top performing systems proposed by the ML models is likely to increase the number of top performers identified in the given parameter space. This is especially relevant for a quantity such as COF, where the overall numerical range is relatively small, e.g., for 5050-test, the top 15% of systems range from 0.0972 ± 0.016 to 0.121 ± 0.010 and the top 30% increases the upper bound very modestly to 0.129 ± 0.008. For $F_0$, the range for the top 15% is from 0.216 ± 0.377 to 0.779 ± 0.135 nN with the upper bound increasing to 1.002 ± 0.171 for the top 30%. For 2575-test, the values are similar, with the top 15% for COF ranging from 0.090 ± 0.012 to 0.122 ± 0.014, with the upper bound increasing to 0.130 ± 0.011 for the top 30%. For $F_0$, the range is from 0.085 ± 0.462 to 0.706 ± 0.410 nN for 15% with the upper bound increasing to 0.951 ± 0.436 when considering the top 30%. Figure 11 plots the overlap between the top 15% systems (MD data), ranked by their simulated tribological properties, and the top 30% performing systems predicted by the 5050-train models. This set up demonstrates a significant increase in accuracy in predicting the best performing systems in both test sets. Specifically, in Fig. 11(a), the accuracy of predicting top performing systems starts at 78.0% ± 7.5% and 92.4% ± 2.1% (for models trained with only 100 data points) to 91.6% ± 2.1% and 97.0% ± 0.6% (for models trained with 1000 data points) when predicting top systems ranked by COF and $F_0$, respectively. In Fig. 11(b), the overlap percentages start at 72.8% ± 6.7% and 85.1% ± 2.1% (for models trained with 100 data points) to 84.1% ± 1.6% and 88.5% ± 1.1% (for models trained with 1000 data points), respectively. In other words, by considering the top 30%, the ML models can reduce the number of additional systems to be considered by 70%, while being able to accurately predict the bulk of the best performing systems, regardless of the mixing ratio and even for very small training set sizes. These results also present other criteria to consider in terms of model accuracy and computational cost in terms of using ML models to prescreen a dataspace as these results show that it may be more efficient to train a model with fewer datapoints but consider a larger range of predictions from the ML model (e.g., simulating the top 30% predicted by the ML model). These results suggest a general approach to combining ML techniques with MD simulations, namely, simulating a small set of systems, e.g., about 5%–10% of systems in the design space, using the simulation results to train predictive ML models, and then utilizing the ML models to screen over a wider range of potential systems, determining systems worthy of further investigation. Such an approach could drastically minimize the number of total simulations needed, decrease the throughput time to scan the parameter space, enabling higher quality candidates to be screened much faster.

Thus far, we have considered models trained only using the 5050-train data. One would assume that accuracy could be improved by training the models on a dataset that also included those systems in the 2575-train dataset (i.e., using the total-train set). Figure 12 plots the scaling for $R^2$ [Fig. 12(a)] and MAPE [Fig. 12(b)] as a function of the training set size when sampled from the total-train set and applied to the total-test set. As seen earlier, the rapid increase in accuracy occurs as the training set size is increased to 1000 data points; the improvements in adhesion are relatively minimal beyond that point, although considerable gains are observed for COF with both measures attaining an $R^2$ value of >0.9 for a training set of size 7816 [see Fig. 12(a)]. We also observe similar trends for MAPE in Fig. 12(b), where the error quickly drops from 0.0593 ± 0.001 (5.93% ± 0.1%) for COF predictions and 0.324 ± 0.033 (32.4% ± 3.3%) for $F_0$ prediction at 100 data points to 0.023 ± 0.0 (2.30% ± 0.0%) for COF predictions and 0.179 ± 0.0 (17.9% ± 0.0%) for $F_0$ predictions at the maximum number of training data (7816). Focusing on the models trained with 1000 data points, predictions from the total-train set ML models applied to the total-test set achieve $R^2$ values of 0.701 ± 0.005 for COF and 0.888 ± 0.005 for $F_0$, which are slightly lower than the values obtained for the models trained with the 5050-train set when applied on the 5050-test set (see Fig. 8). This may appear to indicate that this set of ML models requires more data to attain a similar level of predictive ability, especially for COF models. However, it is worth noting that these evaluations are done on two test sets of differing sizes and compositions. Moreover, the 5050-test set is a strict subset of the total-test set, so the latter includes a wider range of systems and, hence, is deemed more challenging for the ML models. Thus, the relative

![FIG. 11. Intersection between the top 15% performing systems predicted by ML models at various training set sizes and top 30% performing systems calculated by MD simulations (in silico) data of the 5050-test set (a) and the 2575-test set (b). The systems are ranked by COF (blue circles) or $F_0$ (red squares). The dashed, horizontal lines, colored to correspond with its respective property, show the predictive ability of the models trained using the dataset of Summers et al.](image-url)
performance of these models could not be directly compared at this point.

To evaluate the accuracy of using the total-train models for prescreening parameter space, i.e., their ability to determine best performing systems, we again calculate the agreement between the 15% performing systems (398 total systems) from the total-test set as determined by the MD simulations and the top 15% and 30% top performing systems predicted by the ML models, plotted in Figs. 13(a) and 13(b). In Fig. 13(a), a steady improvement in the accuracy of the COF predictions is seen as nearly 8000 data points in the training set are used, achieving an overlap of 84.2 ± 0.3%; $F_0$ exhibits a similar trend observed previously in Fig. 9(b) with little dependence on the training set size beyond 1000 points, maintaining an overlap value of ~63%. Hence, increasing the number of data points to train the ML models can have a positive effect on the predictive ability for some properties, but the large amount of training data needed to improve accuracy may ultimately negate potential performance gains in terms of screening. Recall that for a training set size of 1000 data points, the 50:50-train model could predict >65% of high-performing 50:50 systems and >60% of high-performing 25:75 systems, ranked by either COF or $F_0$ when conducting similar analyses (see Fig. 10). In Fig. 13(b), we perform similar overlap analyses with an extended list of top performing systems predicted by the ML models (top 30%). When determining top COF/$F_0$ systems, we observe that the overlapping fraction rapidly increases to above 80% at $N = 500$ and subsequently plateaus, displaying minimal increase in their accuracy past this point, similar to what was previous observed in Fig. 10.

We now consider a new metric, most-favorable systems, defined as those with both low COF and $F_0$. This list of favorable systems is generated from the intersection of the top 15% [Fig. 13(a)] or 30% [Fig. 13(b)] of systems ranked by their predicted COF and $F_0$ values. This list is compared directly to those in Table I to determine their overlapping fraction, which indicates the ability of the ML models to identify the most-favorable systems. Using this metric, we can see the accuracy of the predictions from the ML models as a function of training points used. In Fig. 13(a), we can see that the overlapping percentage rapidly increases until around 1000
training data points at which point the overlapping fraction values are maintained at ~65%. Meanwhile, if we increase the intersection to 30%, we can see that the overlapping fraction can surpass 80% with as few as N = 500 data points and reach as high as 90.9% when N increases. These results reassert the feasibility of the combinatorial approach described earlier, where users can use MD to generate small sets of data necessary to build a minimally functional baseline ML model to screen over a wider set of potential candidates. The results from such baseline models can be utilized for various efforts, such as focusing on only simulating systems with the most-favorable properties as predicted from the model or building more varied and evenly sampled datasets based on the predictive deficiencies from the baseline model to further improve its robustness. Either approach can improve the quality of results obtained with finite computing resources. Depending on the specific application, complexity of the systems of interest, and computing power, one can choose an optimal strategy to employ, e.g., how many data should be collected to train ML models, and how much of the dataspace to be truncated based on suggestions by the models. However, it is worth recalling that the performance of the random forest regressor algorithm is dependent on the distribution of properties in the training data, and identifying a priori which systems to simulate to ensure appropriate distribution may be challenging. Solving this issue may require additional iterations of training ML models, where we create multiple predictive models as simulation data become available, using these ML models to suggest additional systems to ensure appropriate sampling, rather than identifying favorable candidates at this stage. Although such intermediate ML models may not have high accuracy, they can provide valuable information needed to improve the performance of subsequent ML models.

In Tables II and III, we summarize the outcome, \( R^2 \) and MAPE, of applying all of the ML models, i.e., the model of Summers et al., and 5050-train and total-train models, on all available test sets, i.e., 5050-test, 2575-test, and total-test; the summary of other common metrics, i.e., mean absolute error (MAE) and mean squared error (MSE), is also included in the supplementary material (see Tables S2 and S3). Table II directly compares the performance of the different ML models. We note that of all models trained with 100 data points, the performance of the total-train models exhibits the highest accuracy, followed by the 5050-train, and finally models trained with the data of Summers et al. The difference in performance likely results from the inclusion of mixed monolayer systems, resulting in better distributed training sets provided by total-train. However, this trend does not persist for models trained with more data. Interestingly, we note that the models trained on the 5050-train data appear to have better performance compared to the total-train models when predicting the 5050-test set since they require less data to acquire similar predictive ability. Focusing on the performance of the different models on the total-test set, which is expected to be more difficult to predict, the 5050-train models show comparable

<table>
<thead>
<tr>
<th>N</th>
<th>5050-test</th>
<th>2575-test</th>
<th>Total-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COF (^a)</td>
<td>( F_0 ) (^b)</td>
<td>COF (^a)</td>
</tr>
<tr>
<td><strong>Models of Summers et al.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.472</td>
<td>0.657</td>
<td>0.429</td>
</tr>
<tr>
<td>100</td>
<td>0.543 ± 0.063</td>
<td>0.722 ± 0.058</td>
<td>0.475 ± 0.041</td>
</tr>
<tr>
<td>200</td>
<td>0.647 ± 0.012</td>
<td>0.792 ± 0.033</td>
<td>0.557 ± 0.008</td>
</tr>
<tr>
<td>300</td>
<td>0.687 ± 0.017</td>
<td>0.826 ± 0.011</td>
<td>0.584 ± 0.011</td>
</tr>
<tr>
<td>500</td>
<td>0.734 ± 0.004</td>
<td>0.859 ± 0.008</td>
<td>0.618 ± 0.005</td>
</tr>
<tr>
<td>1000</td>
<td>0.799 ± 0.007</td>
<td>0.885 ± 0.01</td>
<td>0.647 ± 0.001</td>
</tr>
<tr>
<td>1500</td>
<td>0.835 ± 0.011</td>
<td>0.907 ± 0.007</td>
<td>0.652 ± 0.004</td>
</tr>
<tr>
<td>2000</td>
<td>0.88 ± 0.004</td>
<td>0.92 ± 0.002</td>
<td>0.656 ± 0.007</td>
</tr>
<tr>
<td>2500</td>
<td>0.915 ± 0.003</td>
<td>0.927 ± 0.001</td>
<td>0.664 ± 0.003</td>
</tr>
<tr>
<td>2680</td>
<td>0.926 ± 0.0</td>
<td>0.928 ± 0.0</td>
<td>0.668 ± 0.0</td>
</tr>
<tr>
<td>100</td>
<td>0.567 ± 0.019</td>
<td>0.701 ± 0.027</td>
<td>0.511 ± 0.019</td>
</tr>
<tr>
<td>200</td>
<td>0.614 ± 0.032</td>
<td>0.792 ± 0.008</td>
<td>0.555 ± 0.02</td>
</tr>
<tr>
<td>300</td>
<td>0.652 ± 0.021</td>
<td>0.818 ± 0.017</td>
<td>0.595 ± 0.017</td>
</tr>
<tr>
<td>500</td>
<td>0.693 ± 0.02</td>
<td>0.849 ± 0.008</td>
<td>0.631 ± 0.011</td>
</tr>
<tr>
<td>1000</td>
<td>0.738 ± 0.007</td>
<td>0.883 ± 0.007</td>
<td>0.684 ± 0.007</td>
</tr>
<tr>
<td>1500</td>
<td>0.768 ± 0.006</td>
<td>0.897 ± 0.007</td>
<td>0.715 ± 0.005</td>
</tr>
<tr>
<td>2000</td>
<td>0.784 ± 0.006</td>
<td>0.908 ± 0.005</td>
<td>0.736 ± 0.003</td>
</tr>
<tr>
<td>2500</td>
<td>0.807 ± 0.006</td>
<td>0.915 ± 0.009</td>
<td>0.756 ± 0.004</td>
</tr>
<tr>
<td>4000</td>
<td>0.851 ± 0.006</td>
<td>0.928 ± 0.006</td>
<td>0.799 ± 0.004</td>
</tr>
<tr>
<td>6000</td>
<td>0.894 ± 0.003</td>
<td>0.937 ± 0.003</td>
<td>0.853 ± 0.005</td>
</tr>
<tr>
<td>7816</td>
<td>0.925 ± 0.0</td>
<td>0.943 ± 0.0</td>
<td>0.898 ± 0.0</td>
</tr>
</tbody>
</table>

\(^a\) \( R^2 \) when predicting the coefficient of friction.  
\(^b\) \( R^2 \) when predicting adhesion force.
accuracy to the total-\textit{train} models. These observations trends are also confirmed by the MAPE values in Table III. This result suggests that as long as the models are trained on a sufficiently large dataset that captures the distribution of the population well, the models are extensible to untested regimes. That is, for significant variations, in our case composition, it may be more computationally efficient and accurate to train models with different compositions separately, rather than aggregating data into a large training set. Furthermore, all models trained on a relatively small amount of data, e.g., 1000 data points, exhibit similar performance across all test sets. This re-affirms the transferability of these ML models from a more general perspective and highlights the feasibility of utilizing ML algorithms to estimate properties of systems whose designs may be dissimilar to the training data used. Expectedly, the ML models trained with 7816 data points from the \textit{total-train} dataset (80\% of the total dataset) demonstrate the best performance across all test sets but, of course, require the highest computational cost in terms of gathering training data.

As a proof-of-concept of using ML to prescreen the design space, we perform a screening study using one of the \textit{total-train} models (model-0) for COF and $F_b$ (trained with 7816 data points each). The chemical space for this screening was constructed by querying the ChEMBL small molecule library \cite{154902-14} and identifying chemistries whose molecular weight ranges from 4 to 99 amu; this list of chemistries (981) underwent further filtering to remove those containing metallic elements and those that cannot be processed by the RDKit library, e.g., chiral or charged molecules, resulting in 621 unique terminal group chemistries (provided in the supplemental repository \cite{154902-14}). With these 621 chemistries, 193,131 unique systems can be created in which each monolayer is homogeneous (i.e., containing only one type of terminal group); mixed monolayer chemistries were not considered at the moment due to the vast amount of data that would be generated. This simple system design (dual homogeneous monolayers) was chosen to allow more unique chemistries to be considered in a reasonable time frame since the introduction of mixed monolayers would scale up the number of systems to be considered by several orders of magnitude. Descriptors for the 621 terminal groups were determined using the SMILES strings for each chemistry (as described in the section titled Method); these descriptors were then provided as input to the ML models, which, in turn, predicted tribological properties for the 193,131 unique systems. This screening process, which evaluated 385,641 systems since the duplicate systems (i.e., systems in which chemistry A was the top monolayer and chemistry B on the bottom monolayer and vice versa) were not removed, took ~24 h to predict the COF and $F_b$ values on a standard desktop computer (~0.22 s per system), which is orders of magnitudes faster than the time required to perform a single MD simulation, and without the need for expansive computational resources. The distribution of COF and $F_b$ of the systems predicted by the ML model is shown in Fig. 14. We note that the distributions differ from that of the dataset screened using MD simulations (see

\begin{table*}[h]
\centering
\small
\begin{tabular}{lcccccc}
\hline
 & \textbf{N} & \textbf{COF} & \textbf{F}_b & \textbf{COF} & \textbf{F}_b & \textbf{COF} & \textbf{F}_b \\
\hline
\textbf{Model of Summers et al.} & & & & & & & \\
100 & 0.0565 & 0.266 & 0.0653 & 0.341 & 0.0623 & 0.315 \\
100 & 0.0546 ± 0.003 & 0.259 ± 0.016 & 0.0613 ± 0.002 & 0.343 ± 0.018 & 0.061 ± 0.003 & 0.315 ± 0.016 \\
200 & 0.0478 ± 0.001 & 0.224 ± 0.011 & 0.0589 ± 0.001 & 0.312 ± 0.009 & 0.0551 ± 0.001 & 0.282 ± 0.01 \\
300 & 0.0447 ± 0.001 & 0.212 ± 0.007 & 0.0568 ± 0.001 & 0.295 ± 0.011 & 0.0527 ± 0.001 & 0.267 ± 0.009 \\
500 & 0.0407 ± 0.001 & 0.193 ± 0.004 & 0.0542 ± 0.001 & 0.275 ± 0.005 & 0.0496 ± 0.001 & 0.247 ± 0.004 \\
\hline
\textbf{5050-\textit{train}} models & & & & & & & \\
1000 & 0.0399 ± 0.001 & 0.175 ± 0.003 & 0.0521 ± 0.001 & 0.26 ± 0.003 & 0.0459 ± 0.001 & 0.231 ± 0.002 \\
1500 & 0.0301 ± 0.001 & 0.163 ± 0.001 & 0.0552 ± 0.001 & 0.253 ± 0.004 & 0.0445 ± 0.001 & 0.222 ± 0.002 \\
2000 & 0.0251 ± 0.001 & 0.155 ± 0.002 & 0.0513 ± 0.001 & 0.248 ± 0.002 & 0.0423 ± 0.001 & 0.217 ± 0.001 \\
2500 & 0.0209 ± 0.001 & 0.152 ± 0.002 & 0.0508 ± 0.001 & 0.246 ± 0.001 & 0.0406 ± 0.001 & 0.213 ± 0.001 \\
2680 & 0.0196 ± 0.001 & 0.15 ± 0.00 & 0.0505 ± 0.001 & 0.245 ± 0.000 & 0.0399 ± 0.001 & 0.213 ± 0.00 \\
100 & 0.0539 ± 0.0012 & 0.27 ± 0.022 & 0.0621 ± 0.0011 & 0.352 ± 0.04 & 0.0593 ± 0.0008 & 0.324 ± 0.033 \\
200 & 0.0504 ± 0.0025 & 0.228 ± 0.011 & 0.0592 ± 0.0016 & 0.304 ± 0.011 & 0.0562 ± 0.0016 & 0.278 ± 0.011 \\
300 & 0.0478 ± 0.0018 & 0.214 ± 0.005 & 0.0562 ± 0.0013 & 0.277 ± 0.007 & 0.0533 ± 0.0011 & 0.256 ± 0.006 \\
500 & 0.0447 ± 0.0017 & 0.197 ± 0.006 & 0.0533 ± 0.0011 & 0.262 ± 0.006 & 0.0504 ± 0.0009 & 0.24 ± 0.006 \\
1000 & 0.0408 ± 0.0005 & 0.178 ± 0.002 & 0.0486 ± 0.0008 & 0.24 ± 0.005 & 0.046 ± 0.0004 & 0.219 ± 0.003 \\
1500 & 0.0379 ± 0.0007 & 0.167 ± 0.003 & 0.0456 ± 0.0005 & 0.229 ± 0.005 & 0.043 ± 0.0004 & 0.208 ± 0.005 \\
2000 & 0.0359 ± 0.0009 & 0.162 ± 0.003 & 0.0433 ± 0.0004 & 0.225 ± 0.006 & 0.0407 ± 0.0004 & 0.204 ± 0.005 \\
2500 & 0.0336 ± 0.0005 & 0.157 ± 0.004 & 0.0411 ± 0.0004 & 0.22 ± 0.005 & 0.0385 ± 0.0002 & 0.198 ± 0.005 \\
4000 & 0.0287 ± 0.0006 & 0.148 ± 0.002 & 0.0361 ± 0.0006 & 0.21 ± 0.004 & 0.0336 ± 0.0003 & 0.189 ± 0.003 \\
6000 & 0.0236 ± 0.0004 & 0.145 ± 0.002 & 0.03 ± 0.004 & 0.204 ± 0.002 & 0.0278 ± 0.0004 & 0.184 ± 0.002 \\
7816 & 0.0198 ± 0.001 & 0.14 ± 0.00 & 0.0246 ± 0.0001 & 0.2 ± 0.0 & 0.023 ± 0.0 & 0.179 ± 0.0 \\
\hline
\end{tabular}
\caption{Summary of the performance of each ML models when predicting COF and $F_b$ for all different test sets, measured by MAPE. For each data point, the MAPE value is averaged from MAPE of individual ML models (five replicates) when applied to the test set.}
\end{table*}
Figs. 3 and 7), which is expected given the vastly expanded chemical design space. Using the first quartile of the COF distribution (0.1280) and $F_0$ distribution (0.8966 nN) obtained from MD as a reference, the dataset contains 5121 systems that can be considered to have good COF values and 10 598 systems with good $F_0$ values. To further reduce the number of systems of interest, a list of 2000 best performing systems ranked by their COF values and a list of the 2000 best systems ranked by their $F_0$ are compiled with the top 20 systems at the intersection of these lists reported in Table IV. We note that many of the same chemistries that were identified by our initial MD simulations (see Table I) are also observed in this list; specifically, systems 2, 3, and 16 have been considered in the work of Summer et al. along with several other chemistries that may be worth future consideration, such as various alkenes (allyl and propene), alkyenes (acetylene and but-2-yne), halocarbons (1,1-difluoroethyl, bromoethyl, and vinyl chloride), and nitriles (cyano, malononitrile, acrylonitrile). We note that none of the systems reported in Table IV outperform those previously identified in Table I from the MD simulations (in terms of COF and $F_0$) although this might be expected since the systems in Table IV consist of two homogeneous monolayers and, hence, do not include the benefits offered by the mixed monolayers, as we discussed earlier. Nonetheless, this highlights the feasibility of combining ML with MD screening to reduce the computational cost and identify favorable candidates for further study in particular for reducing the vast design space of mixed monolayer systems using such a database for screening.

**TABLE IV.** 20 best performing systems determined by the intersection of the top 2000 systems ranked by their COF and the top 2000 systems ranked by their $F_0$. The properties were predicted using one of the total-train models (model-0).  

<table>
<thead>
<tr>
<th>Terminal group A</th>
<th>Terminal group B</th>
<th>COF</th>
<th>$F_0$ (nN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyano</td>
<td>Propyl</td>
<td>0.1144</td>
<td>0.7257</td>
</tr>
<tr>
<td>Cyano</td>
<td>Cyclopentyl</td>
<td>0.1151</td>
<td>0.4631</td>
</tr>
<tr>
<td>Methyl</td>
<td>Cyano</td>
<td>0.1153</td>
<td>0.5532</td>
</tr>
<tr>
<td>Acetylene</td>
<td>1,1-difluoroethyl</td>
<td>0.118</td>
<td>0.7699</td>
</tr>
<tr>
<td>Cyano</td>
<td>Ethyl</td>
<td>0.1206</td>
<td>0.649</td>
</tr>
<tr>
<td>Fulminic acid</td>
<td>Cyclopentyl</td>
<td>0.1236</td>
<td>0.7117</td>
</tr>
<tr>
<td>Ethylene</td>
<td>1,1-difluoroethyl</td>
<td>0.1244</td>
<td>0.7341</td>
</tr>
<tr>
<td>Bromoethyl</td>
<td>1,2-diformylhydrazine</td>
<td>0.125</td>
<td>0.7695</td>
</tr>
<tr>
<td>Methyl</td>
<td>Fulminic acid</td>
<td>0.126</td>
<td>0.7704</td>
</tr>
<tr>
<td>Cyano</td>
<td>Difluoroethyl</td>
<td>0.1265</td>
<td>0.7279</td>
</tr>
<tr>
<td>Bromoethyl</td>
<td>Malononitrile</td>
<td>0.1269</td>
<td>0.7098</td>
</tr>
<tr>
<td>Acetylene</td>
<td>Ethyl</td>
<td>0.127</td>
<td>0.7254</td>
</tr>
<tr>
<td>1,1-difluoroethane</td>
<td>Propene</td>
<td>0.1271</td>
<td>0.729</td>
</tr>
<tr>
<td>Propyl</td>
<td>2,2-difluoroacetamide</td>
<td>0.128</td>
<td>0.7423</td>
</tr>
<tr>
<td>Acetylene</td>
<td>Propyl</td>
<td>0.1281</td>
<td>0.7777</td>
</tr>
<tr>
<td>Methyl</td>
<td>Acetylene</td>
<td>0.1281</td>
<td>0.7405</td>
</tr>
<tr>
<td>Bromoethyl</td>
<td>1,2-dicyanoethyl</td>
<td>0.1282</td>
<td>0.7737</td>
</tr>
<tr>
<td>Fulminic acid</td>
<td>Ethyl</td>
<td>0.1283</td>
<td>0.7725</td>
</tr>
<tr>
<td>Cyclopentyl</td>
<td>Acrylonitrile</td>
<td>0.1283</td>
<td>0.7152</td>
</tr>
<tr>
<td>Allyl</td>
<td>But-2-yne</td>
<td>0.1283</td>
<td>0.7546</td>
</tr>
</tbody>
</table>

**CONCLUSION**

Utilizing the MoSDeF software suite with the signac framework, a workflow to initialize, parameterize, generate MD engine inputs, perform MD simulations, calculate the tribological properties of interest, and then train a predictive ML model for these soft matter monolayer systems, was developed and extended from a previous smaller scale study. The tribological properties of nearly 10 000 unique system designs have been screened. From the MD simulations, systems that exhibited both low COF and $F_0$ were identified. The bulk of these systems consisted of a cyano group monolayer contacting a heterogeneous monolayer, suggesting mixed monolayer systems may provide a viable route for further tuning tribological performance. However, we note that no clear trends were observed for different mixing ratios in the monolayer, suggesting that their effects strongly depend on the chemistries involved. Although using high-throughput screening, we were able to much more readily determine systems with favorable properties than could be accomplished through experiment, the process still requires a significant amount of time and computing resources. However, coupling MD simulations with machine learning can guide the screening process and reduce the simulations needed in order to optimize system designs. Using this approach, we have assessed the dependence of the random forest regressor’s predictive ability based on the training set provided. The results suggest a positive correlation between the performance of machine learning models with the size of the training set along with factors such as the distribution of the data; however, performance typically plateaus once a modest dataset size is reached. For the type of
systems considered in this study, a training set of 1000 data points is found to be sufficient to train an efficient predictive model. We note that at small training set sizes, i.e., fewer than 500 data points, the machine learning models were still quite successful in determining the best performing systems. Moreover, the models were shown to have high transferability when applied to predict properties of dissimilar systems and that improvements in accuracy seen for larger training sets often do not necessarily equate to improved performance when models are transferred to systems outside of the training set. These findings suggest a synergistic approach of using MD simulations and machine learning to build high quality predictive models and minimize computing resources needed: 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 narrow the parameter space. The performance of the baseline model is dependent on the distribution of training data provided; however, the accuracy of the model can be improved via a few iterations of training using earlier, less accurate, models to guide simulations toward creating well distributed training data. In addition, the baseline model could help confirm/provide insight into the connection between chemical intuition and properties of interest. We also note that care must be taken to ensure that the dataset is not overly biased and that the further trained models are not overfit to the provided data. This work follows guidelines suggested by the TRUE standard, emphasizing the reproducibility and extensibility of the study; accordingly, the supplementary material contains all the information needed to reproduce the simulations and machine learning models described in this work. The code and data are distributed via GitHub.

SUPPLEMENTARY MATERIAL

See the supplementary material for instructions to access the supplemental GitHub repository containing data and analysis codes, additional forcefield details, and further discussion regarding the ML models utilized in this work.

ACKNOWLEDGMENTS

Funding for this work was provided by the National Science Foundation (NSF) through Grant No. OAC-1835874. A.M.-H. also acknowledges support from the National Science Foundation through Grant No. 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 No. DE-AC05-000R22725 via an award from the INCITE program and the National Energy Research Scientific Computing Center (NERSC) operated under Contract No. DE-AC02-05CH11231.

AUTHOR DECLARATIONS

Conflict of Interest

The authors declare no conflicts of interests.

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its supplementary material.

REFERENCES

12 See https://mosdef.org for Molecular Simulation Design Framework (MoSDeF).


See https://github.com/mosdefhub/mbuild for mBuild Github Repository (accessed 17 August 2018).


See https://signac.io/ for Signac Framework.


See https://github.com/mosdefhub/mbuild for mBuild Github Repository (accessed 17 August 2018).