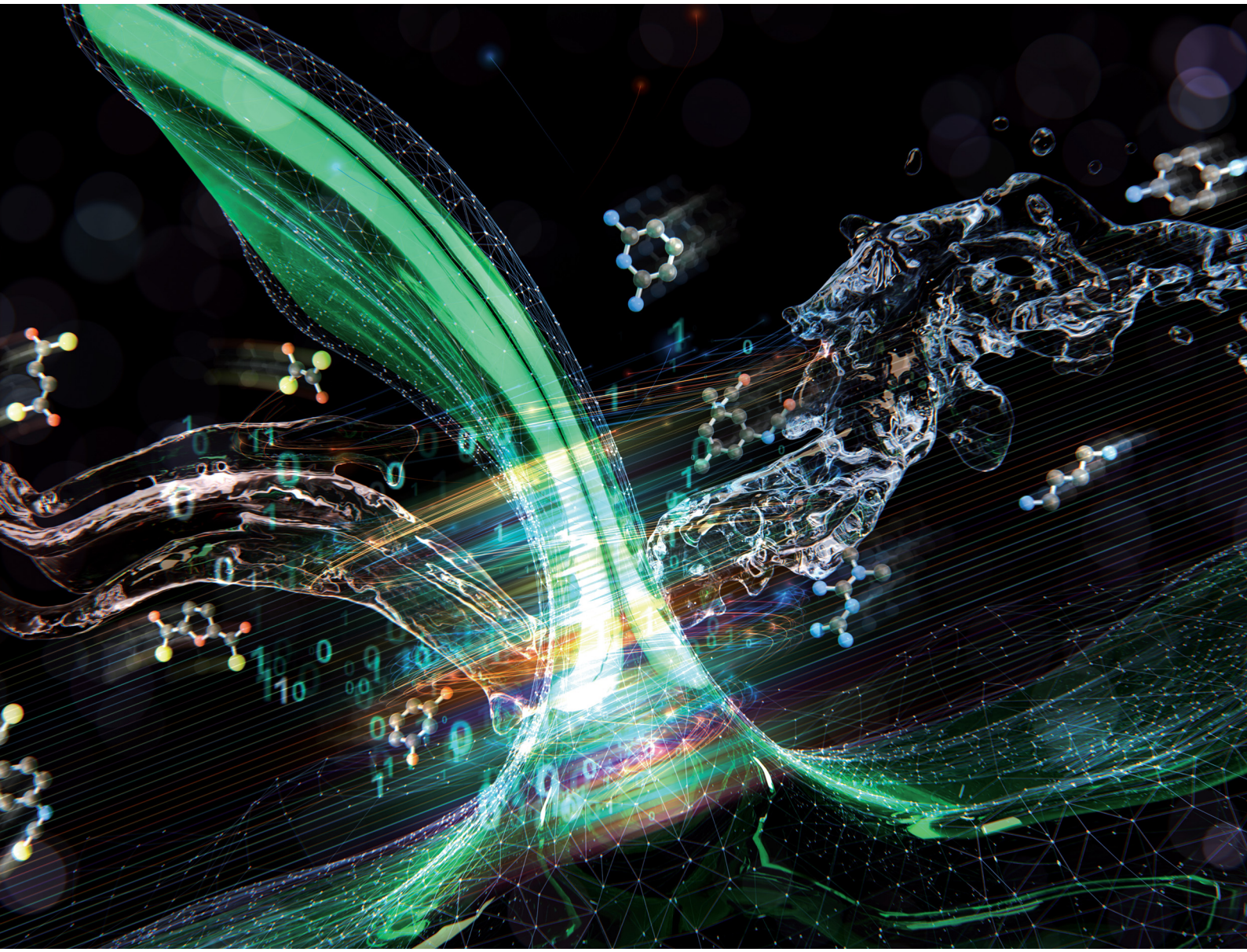
Volume 12
Number 21
7 November 2025
Pages 8759–9320

# Materials Horizons

rsc.li/materials-horizons



ISSN 2051-6347

ROYAL SOCIETY OF CHEMISTRY

**COMMUNICATION**

Gyorgy Szekely *et al.*

A data-driven approach to interfacial polymerization exploiting machine learning for predicting thin-film composite membrane formation

Materials Horizons

ROYAL SOCIETY OF CHEMISTRY

COMMUNICATION

View Article Online
View Journal | View Issue



Cite this: *Mater. Horiz.*, 2025, 12, 9009

Received 24th July 2025,
Accepted 15th September 2025

DOI: 10.1039/d5mh01420d

rsc.li/materials-horizons



# A data-driven approach to interfacial polymerization exploiting machine learning for predicting thin-film composite membrane formation

Gergo Ignacz,[a] Muhammad Irshad Baig,[a] Karuppasamy Gopalsamy,[a] Andres Villa,[b] Suzana Nunes,[a] Bernard Ghanem,[b] Tejus Shastry,[c] Sanat K. Kumar[c] and Gyorgy Szekely *[a]

**Polymeric thin-film membranes prepared by interfacial polymerization are the cornerstone of liquid separation, with the potential to reduce industrial waste and energy consumption. However, the limited diversity of monomers may hinder further development by restricting the accessible chemical space. To address this, we propose a divide & conquer approach for the interfacial polymerization membrane development pipeline. We constructed a dataset using 18 organic- and 73 water-phase monomers, conducting 1246 interfacial reactions and analyzing membranes *via* AFM and optical microscopy. This unprecedentedly large and open access dataset marks a considerable step toward data-driven thin-film membrane development. We trained five machine learning models on molecular structures and density functional theory calculations to study film formation parameters and their binary outcomes. The results indicate that film formation can be predicted directly from monomers, facilitating the potential of data-driven membrane development. Our work shifts the focus from performance prediction to the fundamental step of thin-film formation, offering a new perspective in data-driven membrane research.**

### New concepts

We demonstrate a paradigm-shifting divide & conquer concept for thin-film composite membrane fabrication. This approach focuses on the early phase of material development. We synthesized more free-standing thin-films than all previously reported in the literature combined, meticulously analyzed the formed membranes, and also report negative results, which were previously absent from the literature, hindering the implementation of machine learning in the field. The release of an open-access dataset enables us to build the first structure–activity relationship machine learning models to predict free-standing thin-film formation. We demonstrate remarkable leave-one-out performance, allowing the models to be used in high-throughput virtual screening in the future, an unprecedented advancement in the field. Furthermore, we report the first image classification models trained on optical images. These models can be further employed in high-throughput material fabrication.

## 1 Introduction

Polymer thin-films are at the forefront of separation technologies, sensor applications, supercapacitors, batteries, and polymer-capsule delivery.[1,2] Common techniques for the fabrication of thin-films encompass a variety of physical, chemical, and hybrid methods such as interfacial,[3] vapor and electrochemical depositions,[4] dip and spray coatings, or printing techniques.[5] These techniques allow for precise control over film thickness, composition, and microstructure. Interfacial polymerization is the state-of-the-art fabrication technique for the preparation of thin-film composite (TFC) membranes. During interfacial polymerization, the thin-film forms at the interface of two immiscible phases containing the reactive monomers. The properties of the resulting thin-films vary over a wide range of topological and chemical properties such as anisotropic shapes or hollow cores.[1,6] Through careful selection and design of the monomer chemical structure, interfacial polymerization prepared membranes and thin-films can have a commensurately wide range of applications, from separating small molecules and ions to playing a pivotal role in redox flow batteries,[7] hydrogen purification,[8] impurity removal,[9] catalyst recovery,[10] solute concentration, and solvent recycling.[11]

TFC membrane materials research and development ranges from an early idea for the materials to an application concept (Fig. 1A). This pipeline includes several fabrication and analysis steps, resulting in a generally slow and labor intensive research and development process.[12] We hypothesize an alternative methodology for polymer thin-film fabrication. Instead of

[a] *Advanced Membranes and Porous Materials Center, Chemical Engineering Program, Physical Science and Engineering Division (PSE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. E-mail: gyorgy.szekely@kaust.edu.sa; Web: www.SzekelyGroup.com*
[b] *Computer, Electrical and Mathematical Science and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia*
[c] *Department of Chemical Engineering, Columbia University, New York 10027, USA*

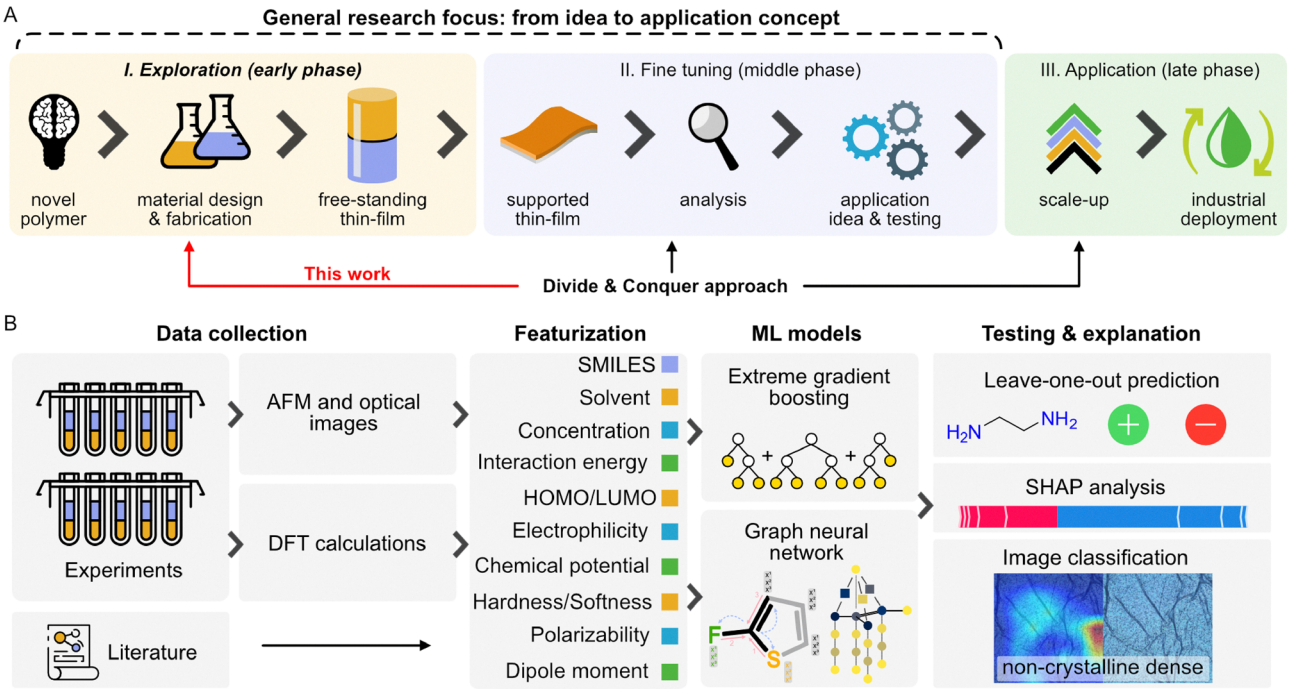



Fig. 1 Schematic representation of membrane development cycle and our workflow. (A) The nine stages of thin-film material development with the three highlighted phases representing the divide & conquer approach; (B) workflow of the study discussed herein including experimentation, calculations, featurization, machine learning models and explanations.

considering the whole idea-to-application pipeline, we direct our attention to the early exploration phase of the workflow (Fig. 1A). Narrowing this design space would allow us to take advantage of previously unexplored approaches, such as building a structure–activity relationship between the chemistry of the monomers and thin-film polymerization. The outcome of the interfacial polymerization depends on several input parameters such as monomer type, concentration, diffusion rate, solubility, and properties of the interface between the two phases. For example, a common scenario occurs when a reaction between two monomers does not result in film formation because of slow reaction rates, high diffusion rates, or powder-like precipitation. Most of the reported studies are characterized by a low diversity in chemical structure among thin-film monomers.[13,14] This low monomer diversity, in conjunction with an absence of negative results, hinders the understanding of the relationship between the input parameters and the result of interfacial polymerization. Predictive models built solely on positive-outcome data will be heavily biased and cannot be used to explore new out-of-band reactions and chemical structures. Recently, these data-driven approaches have emerged to aid and speed up the process of thin-film material design. These works usually use data from the curated literature to build machine learning algorithms, which are then applied to large virtual datasets to identify outstanding monomer candidates in a high-throughput fashion.[14–19] However, approaches inherently assume that a thin-film can be fabricated from the suggested 'lead-like' monomers, ignoring limitations in solubility, reactivity, and film-forming ability. Therefore, their success in exploring new monomeric structures and reactions is limited. Furthermore, datasets used to train machine learning model to predict the expected flux and rejection for membranes are not suitable to predict thin-film formation, as they are trained only on positive results,[14–18] and are also limited to only membrane and separation applications. Similarly, reaction prediction tools are also available,[20] but they only provide information on whether a reaction occurs in a single phase but might not be applicable for interfacial polymerization.

Therefore, we opt to create and provide an open-access reaction dataset with a diverse set of monomers, focusing on a specific set of reactive groups by tackling the first stage of the divide & conquer approach. We also report negative results, as they allow us to develop a classification algorithm to predict whether a film can be formed given the input parameters. Having a stable, defect-free thin-film at the interface is an essential prerequisite, albeit not the sole criterion for any successful polymer thin-film development. Assessing whether a film can be formed is crucial in the early stages of development. Recognizing the foundational importance of thin-film formation in material development, we conduct a comprehensive, data-driven study to illuminate this critical initial step. We perform interfacial polymerization reactions between 18 organic-phase and 73 water-phase monomers, assembling an extensive and chemically diverse dataset to date, greatly exceeding the overall combined scope of the previous literature.[13,14,21] Rather than focusing only on monomer diversification, we aim to explore whether film formation can be predicted from the initial monomer chemical structure and reaction parameters, such as concentration. Using our dataset, we develop machine

learning models that predict thin-film formation and classify optical images into distinct morphological categories. We incorporate features from density functional theory (DFT) calculations and apply advanced techniques, including leave-one-out learning and feature importance extraction for interpretability (Fig. 1B). Furthermore, we present the first comprehensive optical image collection dedicated to thin-film classification. To the best of our knowledge, this work represents the first data-driven exploration of thin-film formation from a chemical structural point of view. We aim to initiate a paradigm shift in the development of polymer thin-film materials by leveraging a divide & conquer approach. By focusing on the crucial first step of film formation, our approach aligns with the anticipated data-driven future of membrane science and integrates well with the emerging inverse design methodologies.

# 2 Results

## 2.1 Interfacial polymerization

Our proposed divide & conquer approach (Fig. 1A) clusters the historical TFC membrane materials pipeline into three major steps: early-stage exploration; fine tuning as a middle stage; and the final application step as the late stage. The rationale behind this separation lies in the current slow idea-to-application approach, which is not compatible with rising data-driven methodologies.[12] The output of the exploration stage is a large number of free-standing films with material characterizations such as AFM, SEM, and infrared spectra. The second stage of the pipeline is fine tuning, which builds upon the positive results of the exploration stage and outputs TFC membranes on support, with performance characterization results such as rejection and permeance. At the end of the second stage, an application concept can be proposed based on the preliminary performance results. For example, whether the membranes would be better suited for reverse osmosis, nano-, ultra-, or microfiltration applications. The last stage is the application phase, which involves scale-up and pilot deployment and sources the membranes for narrow applications addressing a given industrial problem. We tackled the first stage of our divide & conquer pipeline by selecting 18 organic-phase and 73 water-phase monomers and performing pairwise interfacial polymerization reactions between them. Selection criteria were the sufficient solubility in the phases, stability and chemical diversity. The water-phase monomers contained bis- and tris-hydroxy and amine functionalities, while the organic-phase monomers contained acyl chlorides, benzyl bromides, isocyanates, isothiocyanates, and sulfonyl chlorides as bis and tris functionalities (Fig. 2A). Except one non-reactive amine, all selected monomer pairs were expected to polymerize based on their reactivity, forming either a linear polymer or a network polymer. Although linear polymer thin-film materials are atypical, they have been reported before.[22]

Of the 1246 reactions performed, 190 thin-films were formed (Fig. 2A–C), resulting in an approximate 13% hit ratio (hit ratio = number of positive reactions/total reactions, Fig. 2B). In particular, the hit ratio for amines and alcohols with acyl chlorides was 51% and 20%, respectively, and notably lower for isocyanates and isothiocyanates (9%) and sulfonyl chlorides (20%). Under our reaction conditions, benzyl bromides did not form a film with any of the water-phase monomers. Given that we focus more on the exploration of monomers, optimization of every reaction is out of the scope of our investigation, and furthermore, negative results contribute equally in the scoping out the design space. Fig. 2D shows the chemical space between our reaction monomers and the ones from the literature. We performed 6.5 times more thin-film polymerization reaction that were reported in the literature. Moreover, the chemical diversity of our dataset is larger than of the literature visualized by the spreadingness in the latent space.

We characterized the thin-films *via* optical microscopy and atomic force microscopy (AFM) imaging (Fig. 2E–G). Based on the characteristics of the optical images, five classes of polymers were identified: dendritic, dense-gel-like, crystalline, non-crystalline with defects, and non-crystalline dense. These classifications are based on visual examinations and must not be confused with the actual crystallinity determined *via* other techniques like X-ray crystallography. The optical images were classified purely based on visual pattern observation. The AFM height measurements revealed a skewed count distribution of average thickness across all thin-films (Fig. 2G). The average thin-film thickness measured by AFM was 819 nm, with a maximum of 16.9 μm. The formation of films is a diffusion-limited process, where the diffusivity of monomers across the interface plays a crucial role in determining the morphology and structure of the film. Some monomers react to form a homogeneous film, while other monomer combinations form precipitates (*e.g.*, no film formation). The outcome depends on the monomer type and its chemical properties. The diffusivity is controlled by the type of monomer and the solvents used.[23] For example, the diffusivity of the commonly used water-phase monomer *m*-phenylenediamine (MPD) is higher in *n*-hexane compared to *n*-heptane and cyclohexane, which aligns with the solvent viscosity model.[23] A solvent with lower viscosity and lower surface tension tends to improve the diffusion of amine monomers, ultimately leading to thinner polyamide films with higher permeance.[24]

Surprisingly, most of the minor classes originated from water-phase monomers containing the hydroxyl group, and only a few were assigned to amines (Fig. 3A). The four example images are shown in Fig. 3B–E. More than 75% of the thin-films belong to the noncrystalline dense class, and the rest of the images are almost evenly distributed among the remaining classes. The number of hydroxyl and amine monomers was balanced, with 36 and 37 monomers. Fig. 3F show rough and smooth example surfaces for two different thin-films from different reaction types. For a comprehensive list of optical and AFM images, refer to Fig. S1–S122. However, diffusion limitations do not explain why some monomers did not form a film. We also observed film formation for pyridine-2,6-diamine with fumaryl chloride and sebacoyl chloride, but not with succinyl chloride, although the three organic monomers

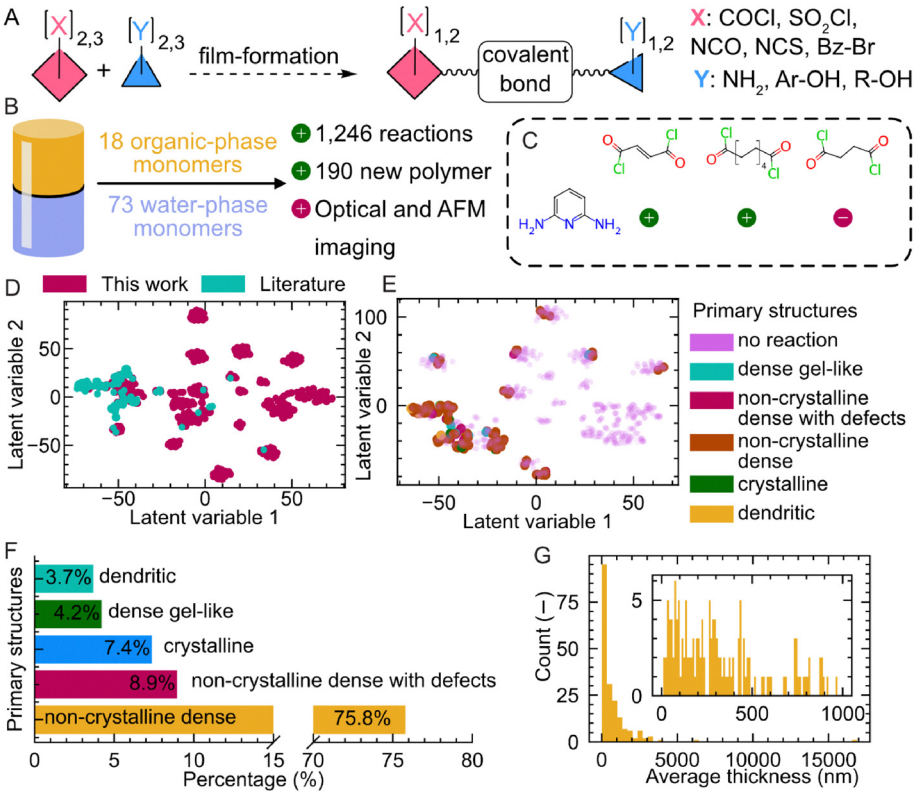



Fig. 2 Thin-film formation reaction results. (A) Thin-film formation reaction scheme and the functional groups used in this study; (B) schematics on the thin-film formation experiments and the summary of the positive and negative results; (C) outcome of the example of three interfacial polymerization reaction of pyridine-2,6-diamine with three different acyl chlorides. The positive and negative signs refer to positive and negative film-formation experiments; (D) principal component analysis (PCA) coupled t-distributed stochastic neighborhood embedding (t-SNE) plot to visualize the chemical diversity and spreadness between this work and the current literature; (E) primary structures identified for this works dataset using a PCA-tSNE plot; (F) percentage distribution of the five primary structure classes determined from the optical images; (G) count distribution of film-thickness.

are all open-chain aliphatic carbonyl chlorides with two reactive functional groups (Fig. 3C). We attribute this phenomenon to polymer chain entanglement, which could sufficiently stabilize the films. Similar linear polymer thin-films have been reported before.[22]

The release of our thin-film formation dataset marks a milestone for data-driven TFC membrane development using interfacial polymerization. Although seemingly less valuable on the surface level, negative results are just as vital as positive results, in that they deepen our understanding of the design/variable space. These negative results could help us to determine when the model is operating out-of-boundary or within training parameters. Furthermore, this dataset has ramifications for eventual reinforcement learning or autonomous experiments. Therefore, we urge the polymer and membrane community to report negative results to help better understand the film formation process and facilitate the implementation of machine learning in the field.

### 2.2 Predicting thin-film formation

Using our film formation dataset combined with available literature data, we created three datasets and trained five machine learning models on them (Table 1). The D1 dataset contains 1719 datapoints with our film formation data and the aggregated literature data. This literature data contained duplicated monomer-pairs but at different concentration and in different solvents. The D2 dataset contained our data with only the chemical structures and D3 contained our data with the chemical structures and the additional calculated DFT features. The D4 dataset only contained the optical images as input features. The raw datasets contain the thin-film reaction parameters, such as solvent type of the phases and the monomer concentrations and the chemical structures of the monomers stored as SMILES strings. Before training the machine learning models, the data were preprocessed. All datasets were split into training, validation, and test sets using a five-fold cross-validation with split ratios of 0.6/0.2/0.2, respectively[25] and the SMILES strings were converted to either molecular graphs or Morgan fingerprints.

The task of the machine learning models were to predict the binary outcome of the film-formation. This task is more suitable for early phase research compared to the currently available machine learning approaches, which are mainly focused on membrane performance prediction.[14–18] The five machine learning approaches used were logistic regression, support vector machine (SVM), extreme gradient boosting tree with linear (SVM-linear) and Gaussian kernel trick (SVM-rbf), Naive Bayes and a graph neural network (GNN). The logistic

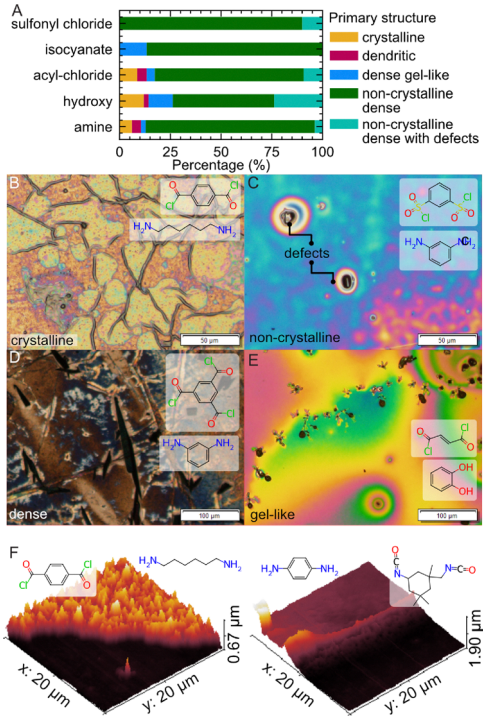



Fig. 3 Thin-film formation reaction results. (A) Percentage distribution of the five primary structure classes for the different monomers; crystalline (B); non-crystalline (C); dense (D); and gel-like (E) optical images of different monomer pairs. Optical images for all thin-films are shown in Fig. S1–S61; (F) example AFM surface images of different monomer pairs. All AFM images are shown in the Fig. S62–S122.

Table 1 The four datasets used in this work. The p/n ratio denotes the positive/negative ratio in the dataset. The positive examples are those where film formation were observed

| Dataset name | Total size | Train size | Validation size | Test size | Features | Film/no-film ratio | Contains |
|---|---|---|---|---|---|---|---|
| D1 | 1719 | 1031 | 344 | 344 | 6 | 0.34 | Our data & literature |
| D2 | 1246 | 747 | 249 | 250 | 6 | 0.13 | Our data |
| D3 | 1246 | 747 | 249 | 250 | 47 | 0.13 | D2 & DFT |
| D4 | 486 | 389 | 97 | — | — | 1.00 | Optical images |

regression model and the SVM-linear are linear models that establish a linear relationship between the input features and the predictions. The other models are considered non-linear. The task of each model was to predict the binary outcome of the film-formation reaction being either negative or positive. The input data and features are the same for all the models except the GNN. The GNN model uses molecular graphs as input, with an option to add extra features, such as concentration or DFT calculation energies to be concatenated to the molecular graph. The logistic regression, SVM-linear, SVM-rbf, XGB, Naive Bayes models uses tabulated data where the chemical structural information is first converted to Morgan fingerprints (extended connectivity fingerprints with radius of 2 or length of 128) and used as input.

Table 2 shows the test validation scores for four metrics: Matthews correlation coefficient, accuracy, receiver operating characteristic area under the curve (ROC-AUC) and the F1 scores. These metrics are widely used in binary classification and the MCC is considered the most reliable single-metric.[26] A comprehensive overview of the metrics are detailed in the Methods section. On D1, all models except Naive Bayes perform similarly (MCC range between 0.869–0.853), with overlapping confidence intervals. Logistic regression scores the highest MCC and ROC-AUC, while GNN scores the highest F1. For the D2 dataset, the MCC scores are lower for all models, ranging from 0.412 for the Naive Bayes (Gaussian) up to 0.742 for the XGB due to the more challenging, imbalanced setting (lower p/n ratio). With class-weighting, XGB has the highest MCC and accuracy, GNN has the highest F1, and logistic regression has the highest ROC-AUC. However, the test results differences across the top three models remain modest and within overlapping confidence intervals. However, the XGB model performed the best on the main MCC metric on the D2 dataset.

Similarly to the D1 results, the Naive Bayes models performed the worst. On the D3 dataeset, adding DFT features yields small but model-dependent changes. Logistic regression has the highest MCC and ROC-AUC scores, while GNN has the highest accuracy and F1 scores with overlapping confidence intervals. Overall, linear and non-linear models capture the main structure of the problem under random splits, with non-linear models (notably XGB) benefitting more from class reweighting on the imbalanced datasets. Removing the molecular fingerprints from D3, we observed a drastic drop in the MCC performance, indicating that the DFT and the process parameters are not enough for accurate film formation prediction. Except the Gaussian Naive Bayes, all models show higher scores for the D3 dataset than for the D2 dataset. Table S7 in the SI displays the non-balanced results.

We observed a noticeable deviation from the baseline score for the different functional groups for the D1 dataset for three models on the test sets: XGB, GNN and logistic regression (Table 3). The models performed very poorly for isocyanates but had similar performance scores for the other functional groups. This negative shift for the isocyantes for all three models is an expected behavior for our dataset where the positive film formation class is very low; thus the classifier defaults back to predicting the majority negative film formation class. Consistently, the F1 score decreases from 0.92 to 0.68 for the XGB model, indicating poor recall on the minority class. For isocyanates, XGB and GNN models show very low ROC-AUC scores of 0.452 and 0.19, respectively. The low ROC-AUC scores show that the ranking of positives *versus* negatives is close to non-informative, meaning that the score distribution itself does not separate the outcome classes. A ROC-AUC substantially below 0.5 suggests a systematic misranking of isocyanate instances by the learned representation for this functional group, meaning that the positive instances are counter-intuitive for the model. Sulfonyl chlorides behave similarly to isocyanates but with lesser deviation from the baseline scores. These functional group effects are also model-consistent, suggesting that both the descriptor and graph-based representation might lack critical features to describe

Table 2 Mean performance metrics for different models and datasets using random splits. Errors are represented as the standard deviation from the mean value for a 5-fold cross-validation. Higher score the better. Bold highlights represents the highest scores for a given dataset and metric. Data are given as mean average values and the error represents the standard deviation of the cross-validation test sets. SVM: support vector machine, rbf: radial basis function, XGB: extreme gradient boosting, GNN: graph neural network. MCC: Matthews correlation coefficient. ROC-AUC: receiver operating characteristic area under the curve. Bold indicates the highest mean per dataset metric. Table S7 in the SI displays the non-balanced results

| Model | Dataset | MCC | Accuracy | F1 | ROC-AUC |
|---|---|---|---|---|---|
| Logistic regression | D1 | **0.869 ± 0.021** | **0.938 ± 0.01** | 0.92 ± 0.014 | **0.986 ± 0.003** |
| SVM-rbf | D1 | 0.858 ± 0.03 | 0.933 ± 0.014 | 0.911 ± 0.02 | 0.971 ± 0.011 |
| XGB | D1 | 0.860 ± 0.020 | 0.934 ± 0.009 | 0.914 ± 0.015 | 0.931 ± 0.009 |
| Naive Bayes (Gaussian) | D1 | 0.774 ± 0.025 | 0.888 ± 0.016 | 0.829 ± 0.022 | 0.963 ± 0.011 |
| GNN | D1 | 0.853 ± 0.025 | 0.935 ± 0.011 | **0.922 ± 0.013** | 0.901 ± 0.017 |
| Logistic regression | D2 | 0.702 ± 0.039 | 0.905 ± 0.011 | 0.74 ± 0.042 | **0.968 ± 0.006** |
| SVM-linear | D2 | 0.647 ± 0.06 | 0.913 ± 0.019 | 0.681 ± 0.05 | 0.954 ± 0.011 |
| SVM-rbf | D2 | 0.61 ± 0.043 | 0.905 ± 0.005 | 0.66 ± 0.044 | 0.942 ± 0.012 |
| XGB | D2 | **0.741 ± 0.041** | **0.924 ± 0.019** | 0.774 ± 0.055 | 0.896 ± 0.033 |
| Naive Bayes (Gaussian) | D2 | 0.412 ± 0.026 | 0.668 ± 0.052 | 0.468 ± 0.044 | 0.787 ± 0.012 |
| Naive Bayes (Bernoulli) | D2 | 0.508 ± 0.052 | 0.793 ± 0.042 | 0.567 ± 0.049 | 0.911 ± 0.017 |
| GNN | D2 | 0.597 ± 0.076 | 0.914 ± 0.017 | **0.782 ± 0.056** | 0.638 ± 0.078 |
| Logistic regression | D3 | **0.72 ± 0.039** | 0.912 ± 0.012 | 0.756 ± 0.041 | **0.969 ± 0.005** |
| Logistic regression | D3 w/o fingerprints | 0.556 ± 0.047 | 0.914 ± 0.008 | 0.578 ± 0.044 | 0.917 ± 0.019 |
| SVM-linear | D3 | 0.635 ± 0.05 | 0.912 ± 0.016 | 0.671 ± 0.05 | 0.952 ± 0.005 |
| SVM-rbf | D3 | 0.667 ± 0.047 | 0.918 ± 0.014 | 0.706 ± 0.046 | 0.944 ± 0.013 |
| XGB | D3 | 0.687 ± 0.055 | 0.906 ± 0.017 | 0.732 ± 0.050 | 0.882 ± 0.027 |
| Naive Bayes (Gaussian) | D3 | 0.482 ± 0.059 | 0.757 ± 0.057 | 0.539 ± 0.055 | 0.853 ± 0.031 |
| GNN | D3 | 0.625 ± 0.02 | **0.919 ± 0.009** | **0.793 ± 0.026** | 0.662 ± 0.015 |

Table 3 Functional group results for the XGB and GNN models for the D1 dataset. Signed percentage values represent the change from the overall values. XGB: extreme gradient boosting, GNN: graph neural network. MCC: Matthews correlation coefficient. ROC-AUC: receiver operating characteristic area under the curve. The higher score the better. Values in parenthesis represents the deviation from the overall value and only test results are given

| Model | Functional group | MCC | Accuracy | F1 | ROC-AUC |
|---|---|---|---|---|---|
| XGB | Overall | 0.869 | 0.938 | 0.92 | 0.986 |
| | Acyl chloride | 0.776 (−10.70%) | 0.904 (−3.62%) | 0.879 (−4.46%) | 0.931 (−5.58%) |
| | Isocyanate | 0.434 (−50.06%) | 0.954 (+1.71%) | 0.686 (−25.43%) | 0.452 (−54.16%) |
| | Alcohol | 0.842 (−3.11%) | 0.941 (+0.32%) | 0.926 (+0.65%) | 0.88 (−10.75%) |
| | Amine | 0.86 (−1.04%) | 0.93 (−0.85%) | 0.93 (+1.09%) | 0.931 (−5.58%) |
| | Sulfonyl chloride | 0.674 (−22.44%) | 0.893 (−4.80%) | 0.8 (−13.04%) | 0.727 (−26.27%) |
| GNN | Overall | 0.853 | 0.938 | 0.986 | 0.92 |
| | Acyl chloride | 0.748 (−12.31%) | 0.89 (−5.12%) | 0.878 (−10.95%) | 0.918 (−0.22%) |
| | Isocyanate | 0.262 (−69.28%) | 0.963 (+2.67%) | 0.554 (−43.81%) | 0.19 (−79.35%) |
| | Alcohol | 0.814 (−4.57%) | 0.942 (+0.43%) | 0.904 (−8.32%) | 0.849 (−7.72%) |
| | Amine | 0.855 (+0.23%) | 0.927 (−1.17%) | 0.927 (−5.98%) | 0.924 (+0.43%) |
| | Sulfonyl chloride | 0.713 (−16.41%) | 0.893 (−4.80%) | 0.875 (−11.26%) | 0.78 (−15.22%) |
| Logistic regression | Overall | 0.869 | 0.938 | 0.92 | 0.986 |
| | Acyl chloride | 0.801 (−7.83%) | 0.914 (−2.56%) | 0.893 (−2.93%) | 0.938 (−4.87%) |
| | Isocyanate | 0.527 (−39.36%) | 0.959 (+2.24%) | 0.741 (−19.46%) | 0.545 (−44.73%) |
| | Alcohol | 0.86 (−1.04%) | 0.949 (+1.17%) | 0.932 (+1.30%) | 0.894 (−9.33%) |
| | Amine | 0.87 (+0.12%) | 0.935 (−0.32%) | 0.935 (+1.63%) | 0.936 (−5.07%) |
| | Sulfonyl chloride | 0.758 (−12.77%) | 0.917 (−2.24%) | 0.872 (−5.22%) | 0.811 (−17.75%) |

the underlying chemistry. Nonetheless, the highest MCC and accuracy scores were reached by the amine and alcohol functional groups, the two main reactive functionalities of the water-phase monomers. The prediction of isocyanates and sulfonyl chlorides is challenging for the models because of the uneven distribution of the film-formation outcomes in the training data. For example, the outcome of the reaction is more dependent on the water-phase monomer than on the type of isocyanate. However, the result of film formation is more dependent on the type of sulfonyl chloride than the water-phase. Fig. 4 shows this monomer dependence as a reaction outcome table. Compared to isocyanates (Fig. 4C), the film-formation outcome is more driven by the organic-phase monomers for sulfonyl chlorides (Fig. 4B). In case of acyl chlorides (Fig. 4A) shows structured dependencies depending on the monomers in the water- or organic-phase. The two Naive Bayes models with a Gaussian and Bernoulli distribution trained on the D2 dataset with only fingerprints (Table 2) also indicate that the data sets are better modeled using the Bernoulli distribution rather than a pure random event.

### 2.3 Leave-one-out predictions

The test scores in Fig. 4 were reported by random data splitting, which means that the original dataset was randomly split into training, validation, and test sets. This random splitting is

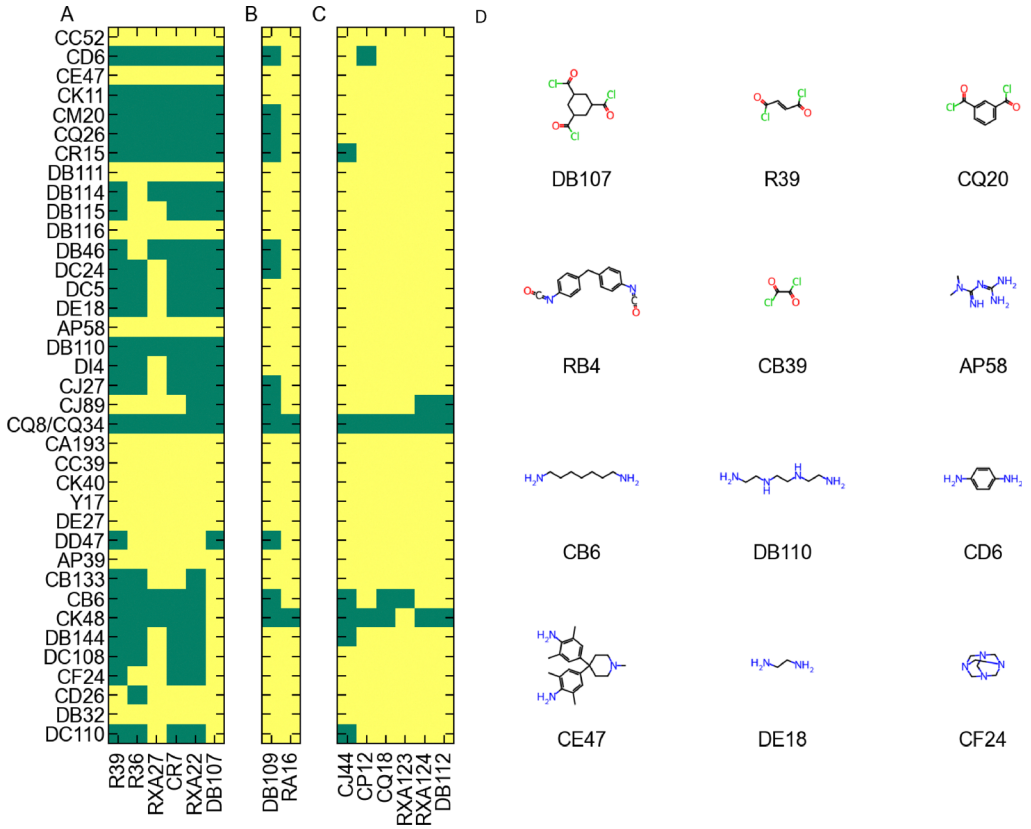



Fig. 4 Film-formation reaction outcome table of amines. (A) Acyl chlorides, (B) sulfonyl chlorides, (C) isocyanates. The code-resolutions are detailed in Table S2. Green color denotes positive film-formation outcomes, yellow color denotes no film-formation outcomes. Each cell represents a film-formation reaction outcome between the water-phase amine monomers (*y*-axis) and the organic-phase monomers (*x*-axis); (D) example structures of the monomers with their respective labels.

representative for testing when the monomer pairs were already seen by the dataset separately but not together. However, in real world applications, we want the model to predict the outcome of unseen molecules. Therefore, in a leave-one-out testing approach, we evaluated the performance of the GNN, XGB, and the logistic regression models by selectively removing each monomer instance and using it as a test set. In our case, leave-one-out learning refers to the evaluation method when the model only sees the monomer in the test time and not during training. Fig. 5A shows that the XGB model outperforms the GNN model in leave-one-out applications, with a difference of 0.016 points in the MCC score for the D1 dataset (water only). However, both the XGB and the GNN models outperformed the logistic regression model by a large margin. This difference is substantially larger than the average difference during random sampling (Fig. 4A). This result indicates that the logistic regression struggle to extrapolate from the chemical space of the training set and the XGB model is best suited to explore new monomers. The leave-one-out test by the XGB model showed higher true positives (expected film formation) and lower false negatives (expected no film formation). The comparison of Fig. 5B and C underscores the better expressiveness of the XGB model. The GNN models in Fig. 5B and C have higher standard deviations in the 1-MCC scores. Compared to the GNN, the XGB models showed narrower 1-MCC scores (Fig. 5B and C).

Fig. 5D shows an example of a water-phase monomer, octane-1,8-diamine, with six organic-phase monomers. In the wet-lab experiments, all monomers formed a film with the exception of isocyanate. Although the GNN model incorrectly predicts no film formation in all instances, the XGB model correctly predicts four out of six instances of film formation. The logistic regression model predicts film-formation for all instances. Fig. 5E shows an example of an organic-phase monomer with six water-phase monomers. In wet-lab experiments, excluding the two phenols, all amine water-phase monomers reacted with the acyl chloride. Similarly to Fig. 5D, the GNN only predicts two out of six examples correctly, while the XGB model correctly predicts all instances (Fig. 5E). The logistic regression model correctly predicts four out the six test examples. The better generalization capabilities of the XGB model in our case make it better suited for future inverse design or high-throughput virtual searches. The low performance of the logistic regression in leave-one-out predictions could be explained by that the logistic regression model is a linear model, while the XGB and the GNN are non-linear. The seemingly lower performance of the GNN model compared to the regression in Fig. 5D and E are attributed to cherry-picking.

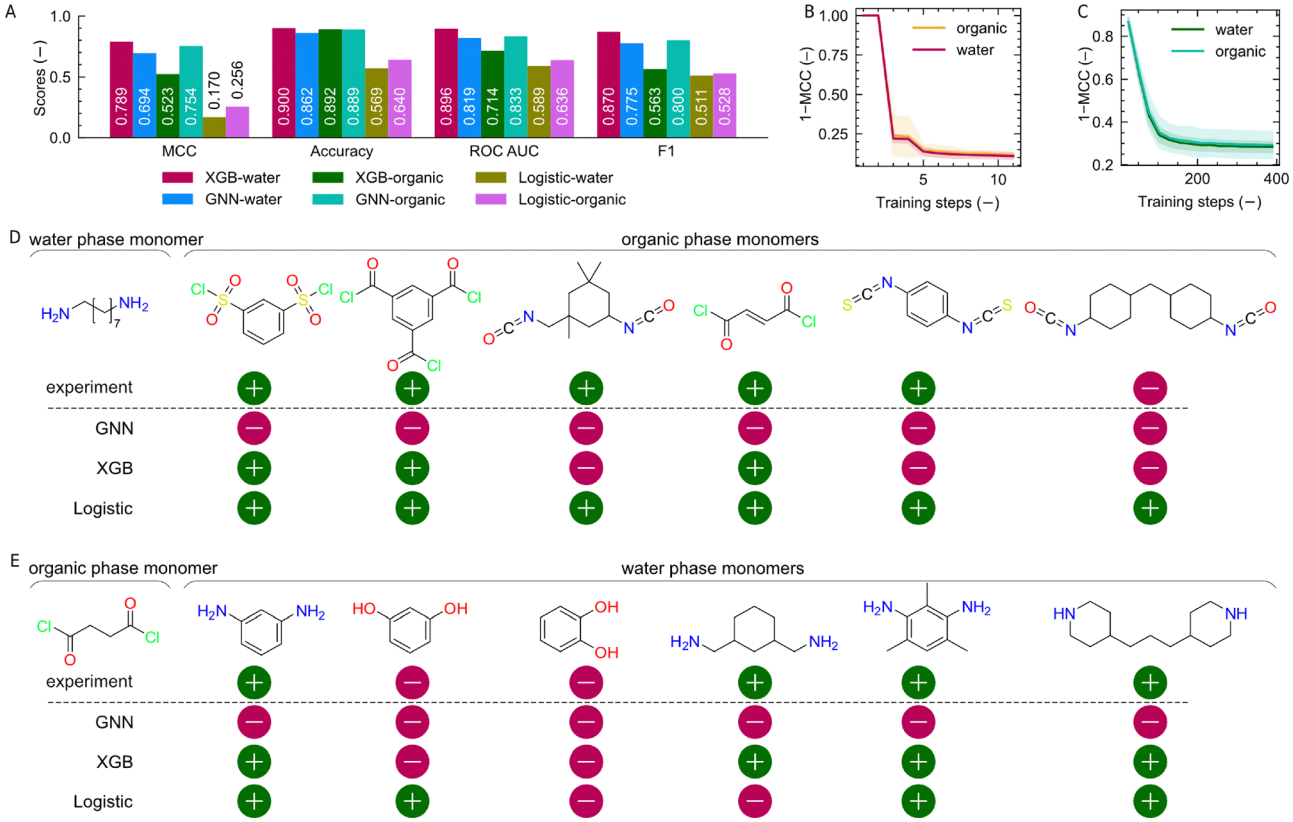



Fig. 5 Leave-one-out prediction results. (A) Average evaluation score of the XGB, GNN and logistic regression models on the water and organic leave-one-out tests. Average MCC score for the XGB (B) and the GNN (C) model on the water and organic leave-one-out test. Data are presented as moving average values; the shading represents the standard deviation. Example water- (D) and organic-phase (E) monomer test with six organic- and water-phase monomers, respectively. The plus sign represents positive outcome for the wet-lab experiment, GNN, XGB and logistic regression models.

We speculate that chemical-structure extrapolation might depend on non-linear feature correlations, which the logistic regression cannot capture.

### 2.4 Importance of DFT features

The test scores between the D2 and D3 datasets revealed that the XGB model performs better when DFT calculations are included (Table 2). These DFT features resulted from a systematic study of the non-covalent interactions between both monomers and the two solvents, water and *n*-heptane. Each entry contained the results of four relaxed structure pairs, each with 11 electronic indices. These 44 additional features were used during the training of D3, while D2 did not contain these features. Fig. 6A–D show the relaxed structure of a water-phase and an organic-phase monomer with both water and *n*-heptane (Fig. 6A and B). The calculated features include, for example, HOMO/LUMO energies, electrophilicity, and ionization energies. The full list of calculated features can be found in the Methods section. We performed a SHAP and feature importance analysis on the XGB and logistic regression results, respectively, to better understand the impact of the DFT features on the prediction performance. The SHAP analysis highlighted that the XGB model extensively uses the DFT features during prediction (Fig. 6E) for the D3 dataset. For the D2 dataset, where the XGB model can only use Morgan fingerprints for predictions, the test scores are lower (Fig. 4B). These indicate that the calculated DFT features could slightly improve the prediction performance in all instances, but the prediction does not explicitly depend on them. For the XGB SHAP results, the 10 best performing features have a negative score, meaning that their DFT value or the presence or absence of fingerprints drives the model for the prediction of negative film formation. This observation is expected due to the low positive hit ratio (13%).

In L1-regularized logistic regression, a smaller subset of features can be identified that model the underlying pattern in the data reasonably well.[27] The weight of the features can then be directly associated with their relative importance during prediction given the linear nature of the model. Fig. 6G and H shows the average feature importance of the 10 best performing features by their absolute values for the logistic regression model for the D3 and D2 datasets respectively. Similarly to the XGB-SHAP results, additional DFT features are included in the top performing features (Fig. 6G); however, to fewer instances compared to Fig. 6E. HOMO energies, ionization energy, electrophilicities, dipole moment and polarizatbility had the highest feature importance for both XGB and the logistic regression.

Feature importance values for the D2 dataset (Fig. 6F and H) share similarities with the D3 results (Fig. 6E and G), suggesting that the addition of DFT features did not affect the predictions significantly. These observations are in line with the minor differences in the D2 and D3 test scores (Table 1).

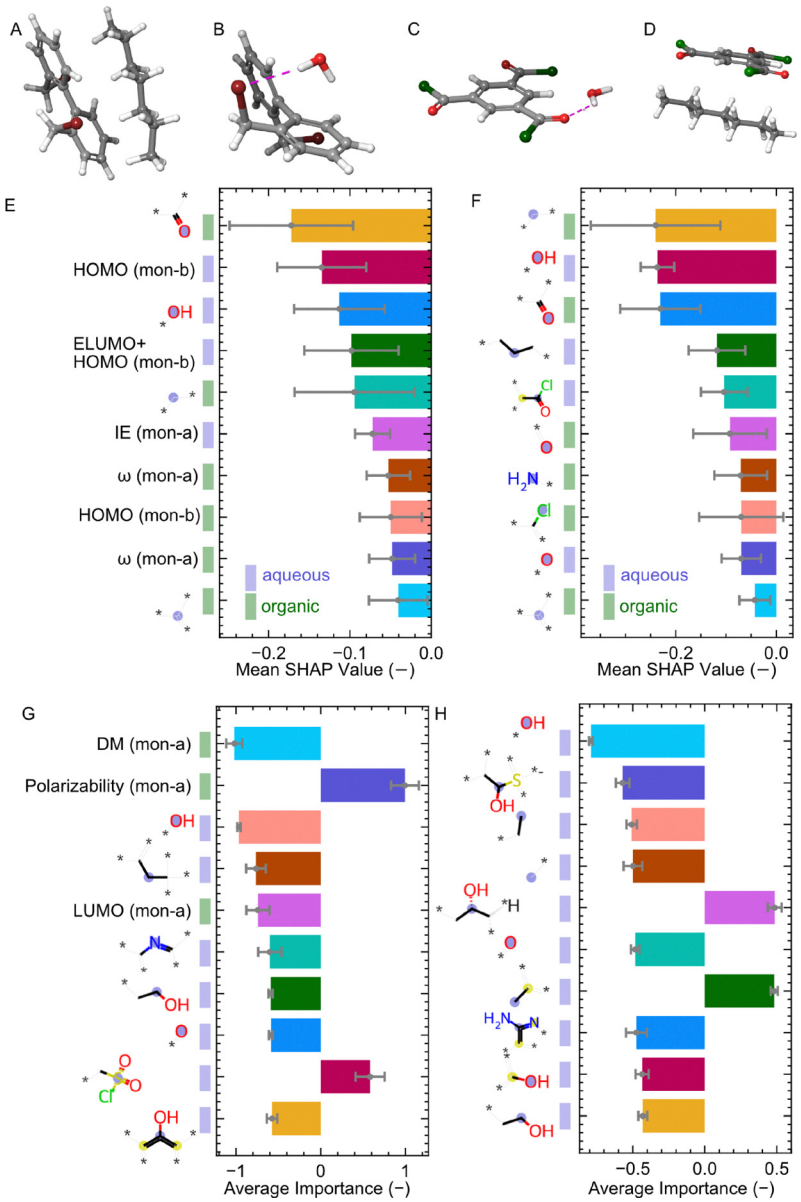



Fig. 6 Geometry optimization and SHAP analysis. Relaxed structure between a water-phase monomer with *n*-heptane (A) and water (B). Relaxed structure between an organic-phase monomer with *n*-heptane (C) and water (D). SHapley Additive exPlanations (SHAP) results of the XGB models for all five folds of the validation for (E) D3 and (F) D2 datasets. Average feature importance of the logistic regression models for the (G) D3 and (H) D2 datasets. "mon. a" and "mon. b" denote the water and organic-phase monomers, respectively. The blue-highlighted atoms of the Morgan-bit structures denotes non-aromatic atoms. Data are presented as average values. Error bars represents standard deviation. $\omega$ denotes electrophilicity; DM denotes dipole moment, HOMO/LUMO denotes highest/lowest occupied/unoccupied molecular orbital in eV, ELUMO denotes the electronic lowest unoccupied molecular orbital in eV.

The top performing fingerprints are similar across the D2 and D3 datasets and the XGB and the logistic regression model. These fingerprints are most commonly simple structures, such as hydroxy groups, secondary or tertiary aliphatic carbon atoms with negative average importance scores. Only the carbonyl and the imine-like Morgan-bits showed positive average importance scores (Fig. 6H).

### 2.5 Image classification

We designed a deep-learning model for classifying films based on optical images. As depicted in Fig. 7A, the model comprises a visual encoder that extracts salient features from the images and a linear classifier that assigns a class label based on these features. The model was trained using a fully supervised approach. Given a labeled dataset $\mathcal{D} = (x_i, y_i)_{i=1}^{n}$, the optimal model parameters $f$ were learned by minimizing the error between the model's predictions and the ground truth labels, following the optimization function: $\mathcal{L}(f(x_i), y_i)$.

To identify the most effective visual encoder, we experimented with state-of-the-art architectures, including the ResNet family (ResNet18, ResNet34, ResNet50) and the EfficientNetv2 family (EfficientNetv2-s, EfficientNetv2-m).[28,29]

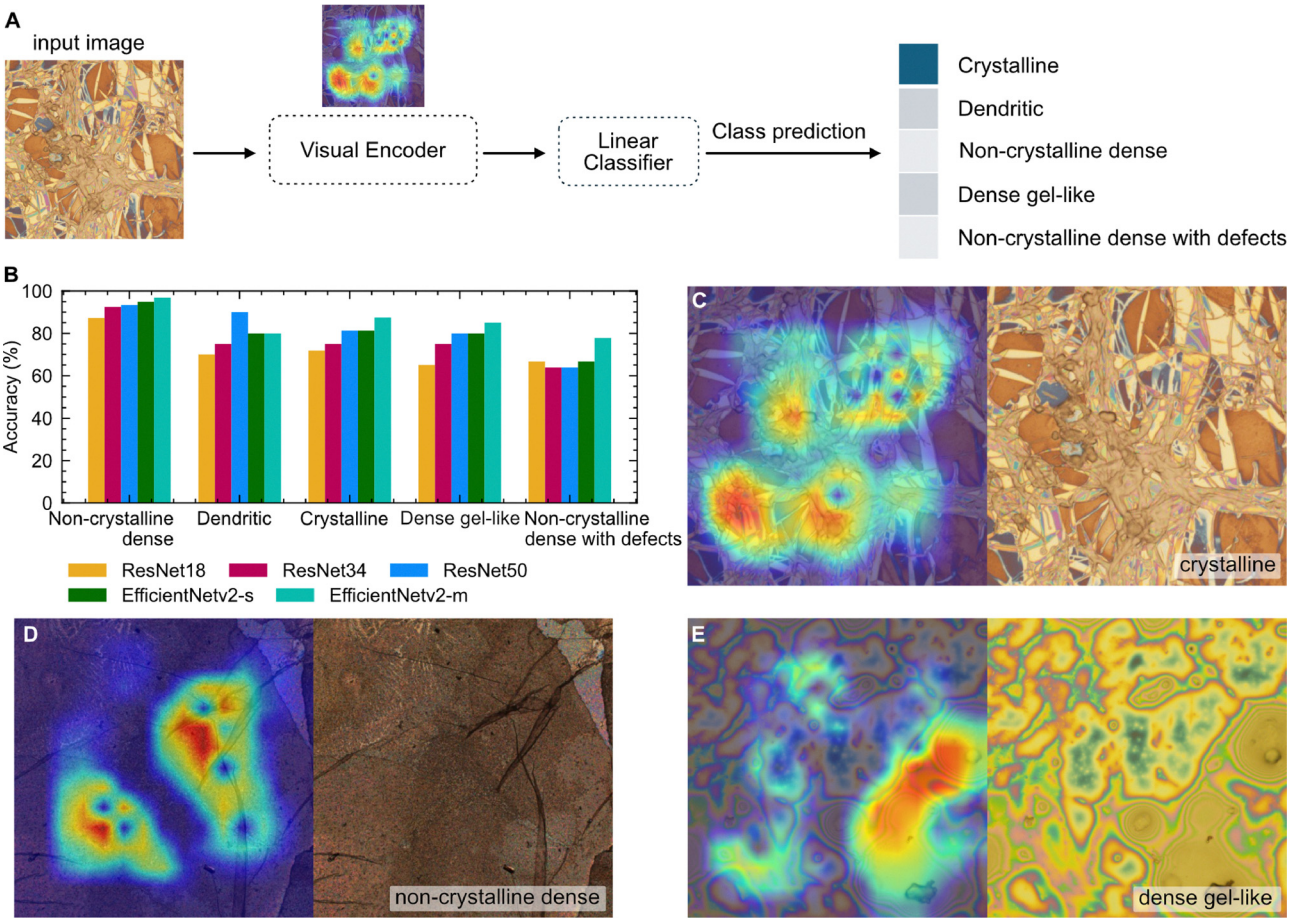



Fig. 7 Image classification model outline and results. (A) Our model leverages a pretrained visual encoder on ImageNet-1k and a linear classifier to predict the film type from the optical images. (B) We analyze the accuracy per class. It is worth noting that EfficientNetv2-m outperforms all the other visual encoders in most classes and obtains a performance of over 78% in all classes. Grad-CAM visualizations for crystalline (C), non-crystalline dense (D) and dense gel-like structures (E). We leverage Grad-CAM to explain the model's decision. It generates heatmaps highlighting the regions of the original images that the model considers to classify them. All displayed Grad-CAM images were predicted correctly by the model.

Given the limited and imbalanced nature of our dataset, we leveraged pretrained models on ImageNet-1K[30] to mitigate the risk of overfitting. Additionally, to address the class imbalance, we applied a weighted sampling strategy that increases the probability of selecting instances from underrepresented film classes. This approach helps to prevent the model from becoming biased towards the dominant "non-crystalline dense" class, which constitutes the majority of the dataset, as shown in Fig. 2F.

In Table 4, we report the average accuracy, which accounts for the per-class accuracies presented in Fig. 7B, thereby avoiding bias towards the most prevalent class. Our model, employing EfficientNetv2-m as the visual encoder, achieved an accuracy of 85%, outperforming all other tested models by over 5%. Furthermore, Fig. 7B demonstrates that this model consistently performs well across various classes, including those with limited data, such as the "crystalline" class, which represents only 8% of the dataset.

To further investigate the performance of our model with EfficientNetv2-m as the visual encoder, we utilized gradient-weighted class activation mapping (Grad-CAM).[31] This technique highlights the image regions most influential in the model's decision-making process. Fig. 7C and D illustrates two examples of correctly classified instances from different classes, revealing that our model effectively captures the relevant features necessary for accurate classification.

Table 4 Image classification results by model type. We report the average of the accuracies, F1-scores, ROC-AUC and MCC scores per class to handle the data imbalance. It is important to note that the EfficientNetv2-m, the biggest method we evaluated, outperforms the others by a large margin

| Model | Model size | Num. output features | Avg Acc. | Avg. F1-score | Avg. ROC AUC | Avg. MCC |
|---|---|---|---|---|---|---|
| ResNet18 | 11.2 M | 512 | 72% | 83% | 95% | 63% |
| ResNet34 | 21.2 M | 512 | 76% | 87% | 96% | 70% |
| ResNet50 | 23.5 M | 2048 | 81% | 89% | 97% | 75% |
| EfficientNetv2-s | 20.2 M | 1280 | 80% | 90% | 97% | 77% |
| EfficientNetv2-m | 52.9 M | 1280 | 85% | 93% | 99% | 84% |

For instance, the model pays more attention to the edges and clusters on the optical images Fig. 7C–E.

# 3 Discussion

Our study provides a paradigm-shifting approach for the early phase of TFC membrane development. Compared to previous approaches, our focus on only the film-formation data allowed us to develop predictive models to aid this early phase of discovery. We proposed a divide & conquer pipeline to reverse the old application-focused paradigm to be more aligned with modern data-driven approaches. Separately focusing on the first exploration stage alleviates the pressure to report only those use cases where the monomer-pair dissolved in the solvents; formed free-standing films; successfully attached to a support; analyzed; and proposed an application for an industrially relevant use case (Fig. 1A). We only tackle the first stage of the divide & conquer challenge by selecting a large number of water- and organic-phase monomers to perform interfacial polymerization reactions across them. The rationale behind our proposed pipeline is that current TFC research only employs limited number of monomers, mainly focusing on a selected few reactive acyl chlorides and diamines, which were discovered and developed in the early 70's.[32] Our divide & conquer shifts the focus towards data-centric membrane development. The low hit ratio of approximately 13% in the 1246 reactions highlights the challenges in achieving successful thin-film formation, as well as the importance of reporting negative data. Performing high-throughput reactions with high hit-ratio would minimize material needs. This low success rate aligns with the inherent complexity of the interfacial polymerization process, where multiple variables such as monomer reactivity, diffusion rates, concentration, and interface properties come into effect. Our study focused on diversifying both the organic- and the water-phase monomers. Despite all monomers having reactive functional groups, some monomer pairs did not form free-standing films, which indicates that factors beyond simple reactivity, such as polymer chain entanglement, solvent interactions, and oligomer solubility, might play important roles. However, quantitative exploration of these underlying factors has not been determined, and it remains a topic of debate. Nonetheless, our dataset containing an extensive amount of negative examples is a valuable resource on its own for future research. Our dataset contains both more organic- and water-phase monomers than the combined available literature.[13,14,17,33] We also showed several instances in which film formation can be realized using monomers with two reactive functional groups, forming linear polymers. However, thin-film membranes with linear polymer chains have been described before.[22] We also incorporated thin-film polymers that are less utilized, such as polysulfonamides. The observed structured dependency across the film-formation outcomes highlights that if monomers tend to form films, they usually form films with all other monomers. We hypothesize that the underlying reactivity and diffusion of each monomer result in this highly structured binomial distribution. The structured distribution suggests the existence of one or more molecular or process features that the outcome is highly dependent for both the water- and organic-phase monomers. This underscores the potential of the dataset in the development of thin-film materials, including artificial intelligence and ML applications. Our dataset serves as the first step towards closing the loop[34] in materials development for porous materials[35] and polymers.[36]

Our optical microscopy analyses revealed that the majority of the films were noncrystalline and dense, suggesting that our selected monomers tend to favor the formation of amorphous polymers. The skewed distribution of thicknesses observed in the AFM measurements further supports the idea that diffusion-controlled processes dominate film formation, leading to substantial variability in film morphology. The identification of different structural classes based on optical imaging provides a framework for categorizing thin-film morphologies, which could be valuable for future studies aiming to optimize film properties or to study their structure–property relationships. These findings could facilitate the development of targeted materials by helping select the most promising monomer pairs.

The machine learning models developed in this study demonstrated promising predictive capabilities, particularly the XGB model, which outperformed all models in both random and the GNN in leave-one-out tests. Boosting trees usually outperform neural networks on tabular data, where the input features are individually meaningful and lack strong multi-scale temporal or spatial structures[37] and they are usually heterogeneous and noisy with high cardinality and different scales.[38] Boosting trees can efficiently determine the decision space using the hyperplane-like boundaries in the tabular data. However, the GNN makes use of the node-neighborhood and node features of the molecular structures. Based on the SHAP analysis results from Fig. 6E, the additional tabulated features, such as electronic properties play a key role in determining film formation. These tabulated features are more suited for the XGB and likely causes the higher performance compared to the GNN. The marginal difference between linear models (logistic regression, SVM-linear) and the other non-linear models suggest that the structured binomial distribution of the data can be modeled using log-linear models. The high predictive performance of the logistic regression in the random splitting test suggest that data can be efficiently explained with linear models, while nonlinear models add only a minor improvement. Introducing class-weighting to address the strong class imbalance produces predictable shifts in the four metrics. Balanced training moves the decision boundary toward the minority class, typically increasing recall (and thus F1 and often MCC) while also modifying ranking, which resulted in the ROC-AUC score changes. Consistent with this, after reweighting we observe modest increases for XGB on D2 and a trade-off for GNN (higher F1 with a lower MCC). These changes are quantitatively small, mostly within cross-validation uncertainty. While the inclusion of DFT-calculated features only slightly improved the predictive accuracy of the XGB model, it suggests that additional calculated features could improve the overall predictive power. However, the calculated

DFT features alone are not enough for an accurate film formation prediction. Thus, molecular features play a crucial role in determining the film formation outcome directly or indirectly. As expected, the low positive film formation ratio, the most important molecular Morgan-bits have negative mean importance values for both the XGB and the logistic regression model. The most important molecular Morgan-bits were simple structural parts. An indirect effect could be the altered reaction or diffusion rate caused by a particular molecular moiety. The ability of the XGB model to generalize to unseen monomers in the leave-one-out tests can be a useful tool in the design of new monomers for thin-film applications. In the leave-one-out test, the logistic regression scored notably lower than both the GNN and the XGB, suggesting that extrapolation to new monomer pairs requires nonlinear models. This result is in strong contrast with the results from random splitting, where logistic regression outperformed XGB and GNN. The XGB model's extrapolation capability is particularly important for the inverse design of novel materials, where the goal is to identify monomer candidates with a high likelihood of forming stable films. However, more comprehensive data including variations on the concentration, solvent, and temperature of the film formation are necessary. Our study is the first to address the importance of film-formation and develop models to predict polymer thin-film formation. The strong linearity of the model further suggests that if a single example is provided to the model for a new monomer, further film formation can predicted with high accuracy. This result would help researchers streamline and narrow wet-lab experimentation.

The machine learning models developed in this study demonstrates the potential of artificial intelligence to advance materials characterization and accelerate the early phase of the development cycle of TFC membrane materials. By enabling accurate classification of membrane formation from monomers, the model makes the workflow more efficient in materials research. The optical image classification lays the groundwork for future studies aiming to establish stronger correlations between morphological features and material properties, ultimately aiding in the rational design of thin-films for targeted applications.

Overall, this study provides a robust dataset and predictive tools that can be leveraged for the rational design of new materials. Our approach is a paradigm shift from the conventional polymer thin-film works. High-throughput and automated robotic applications are on the rise.[39,40] The release of our thin-film formation dataset can be used to close the loop in material synthesis applications.[41,42] By encouraging the sharing of both positive and negative results, we aim to foster a more comprehensive understanding of the factors that govern successful thin-film formation, ultimately accelerating the development of next-generation polymer thin-film materials.

# 4 Methodology

## 4.1 Thin-film formation reactions

All chemicals were purchased from commercial suppliers without further purification. The thin-film formation reaction were performed according to literature data.[43] In summary, the monomers were dissolved at given concentrations in water (2.0 wt/vol%) and heptane (0.15 wt/vol%) and then poured in this order into a small (5 ml) flat bottom screw-cap vial. The reaction left for running between 3 minutes for amines and 18 minutes for alcohols. In total, 1246 reactions were performed and 190 films were analyzed. For a detailed description on the reaction parameters and preparation, please refer to the Supplementary Methods. Except DB32 (Table S2), all organic- and water-phase monomers were expected to undergo chemical reaction with their respective pair. The hard-negative example was selected to monitor and eliminate false positive film formation reaction (human misclassification). Usual preparation time included around three hours solution and vial preparation, while setting up the reaction took around half an hour while cleanup took another 10–15 minutes. The reactions were done in batch and parallel and the approximated productivity was around 20 reactions per day calculating by 7–8 hours laboratory time per day.

## 4.2 Optical and AFM height images

The optical images of the free-standing thin-films were taken on Olympus Material Microscope BX61 (OLYMPUS, Japan) at three magnifications. The height profiles of all the samples were estimated by conducting atomic force microscopy (AFM) using Dimension Icon SPM (Bruker, USA). A scan area of 20 μm by 20 μm was scanned in tapping mode in air using a RTESPA probe (Bruker). The AFM images were further processed in Gwyddion software to extract the height profiles. All optical and AFM images are presented in Fig. S1–S122.

## 4.3 Optical image classification

The human expert classification of TFC membrane morphologies was based solely on their appearance under an optical microscope. Given the limited availability of optical images of TFC membranes in the literature, drawing direct analogies with similar polymeric systems is challenging. However, the closest resemblance was found with the microstructures of metallic alloys, which often display distinct crystalline and dendritic patterns and are well-documented in the literature. Membranes with crystalline morphology exhibited crystallite-like microstructures, resembling those commonly observed in metallic alloys such as steel or brass.[44] The membranes featured similar dark regions similar in appearance to alloy grains, hence the term "crystalline." Membranes with non-crystalline dense category lacked the distinct domain structures observed in crystalline morphologies. Instead, they appeared homogeneous and uniform under the microscope, with no visible crystalline features. Dense gel-like membranes displayed a sticky, gel-like texture and appearance, which was also apparent during handling. The dense and cohesive nature of the morphology led to the label "dense gel-like." Membranes with dendritic category included membranes with branched, tree-like patterns resembling dendritic microstructures frequently reported in metallic alloys such as 304 stainless steel. Membranes displaying this distinctive pattern were classified as dendritic.[39,45]

The non-crystalline dense with defects membranes were morphologically similar to the ''non-crystalline dense'' type but featured prominent dark regions that interrupted the otherwise homogeneous background. These darker zones were interpreted as defects, distinguishing them from the defect-free dense membranes.

### 4.4 Density functional theory calculations

Geometry optimizations and subsequent vibrational frequency analyses were carried out at the M06-2X/6-31++G** ref. 46–48 level of theory and no imaginary frequencies were obtained for both the complexes and the individual monomers. M06-2X is one of the widely accepted global hybrid functional developed by Zhao and Truhlar and the performance of the functional well tested for noncovalent interactions, thermochemistry, and kinetics. All electronic structure calculations were carried out with the computational chemistry software package Gaussian 09[49] and results were analyzed with the help of the GaussView 6.0 program.[50]

The conceptual DFT reactivity indices such as chemical potential ($\mu$), hardness ($\eta$), softness ($S$), electrophilicity ($\omega$), dipole moment and polarizability ($\alpha$) could be vital to understand the chemical reactivities of the system and whose analytical explanations are defined as follows:

$$\mu = \frac{E_{\mathrm{LUMO}} + E_{\mathrm{HOMO}}}{2} \tag{1}$$

$$\eta = \frac{E_{\mathrm{LUMO}} - E_{\mathrm{HOMO}}}{2} \tag{2}$$

$$S = \frac{1}{2\eta} \tag{3}$$

$$\omega = \frac{\mu^2}{2\eta} \tag{4}$$

$$\langle\alpha\rangle = \frac{1}{3}\left(\alpha_{xx} + \alpha_{yy} + \alpha_{zz}\right) \tag{5}$$

where $E_{\mathrm{HOMO}}$ (HOMO-highest occupied molecular orbital) and $E_{\mathrm{LUMO}}$ (LUMO-lowest unoccupied molecular orbital) is the frontier molecular orbital energies, $\langle\alpha\rangle$ is the mean of diagonal components ($\alpha_{xx}$, $\alpha_{yy}$, $\alpha_{zz}$) of the polarizability tensor.

The interaction energies (IEs) for the complexes with water and *n*-heptane molecules adsorbed were calculated using a supermolecule approach and corrected for basis set superposition error (BSSE) using the counterpoise (CP) procedure suggested by Boys and Bernardi.[51]

$$\mathrm{IE} = E_{\mathrm{complex}} - \left(E_{\mathrm{host}} + E_{\mathrm{guest}}\right) \tag{6}$$

Where $E_{\mathrm{complex}}$ is the total energy of the complex formed between the monomers and water/*n*-heptane. Both $E_{\mathrm{host}}$ and $E_{\mathrm{guest}}$ represent the energies of the monomer and water/*n*-heptane, respectively.

### 4.5 Literature data collection

The literature data was collected building upon previous works[17,33] and from the open membrane database.[21] We aggregated data from 163 peer reviewed research articles and removed those duplicates, where the two phases monomers and the concentration were pairwise the same. The final literature data contained 477 entries with.

### 4.6 Datasets

We created four different datasets from our reaction data, literature, and image data (Table 1). We used a five-fold cross validation test dataset on D1–D3 to get a more average sense of model performance. The D1 dataset was the combination of the literature data with our in-house measurements, providing 1719 trainable examples. D1 contained different solvents, concentrations, and a variety of monomers, but all were positive examples (film formed). The D2 dataset contained only our work without DFT calculations. D3 contained all of the D2 dataset as well as the DFT calculation results. The total trainable examples were 1246 for both D2 and D3. The split ratios for D1–D3 were 80%/20% between train and the test. The train set was further split into train and validation set by 75%/25% ratio. The model parameters were fit on the training set, while hyperparameters were fit using the validation set. After model training, the test set was used asses the final trained model's performance. D4 contained 486 images, three images per film using different magnifications at different spatial positions.

### 4.7 Computational methods

The graph neural network used herein was built *via* chemprop[52] using a directed message passing neural network architecture.[53,54] The XGB model was trained using the xgboost python library.[55] Chemical data preprocessing and visualization were done by the rdkit python package.[56]

The binary Matthews correlation coefficient (MCC) is defined as:

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{7}$$

where: TP is the true positive count, TN is the true negative count, FP is the false positive count and FN is the false negative count. We chose binary MCC over the traditional ROC-AUC score because the binary MCC involves positive and negative prediction values and generally has better specificity and sensitivity.[57] MCC is also advantageous against other metrics, such as F1 score and accuracy[26] because it considers all four elements of the confusion matrix (TP, TN, FP, FN), ensuring a high MCC only when true positives, true negatives, positive predictive value, and negative predictive value are all high. This is unique, as other metrics might score high despite misclassifications in one of these areas. MCC incorporates dataset prevalence and classifier bias, making it robust against distortions that class imbalance might introduce. This is important in our case, because all datasets have a strong class imbalance. A high

MCC indicates that the model is simultaneously good at identifying both film-forming and non-forming monomer pairs, which mirrors the experimental need to balance discovery of new films against avoiding dead-end reactions.

$$\text{ROC-AUC} = \int_{-\infty}^{\infty} \text{TPR}(t)\,\text{d}(\text{FPR}(t)) \tag{8}$$

where: TPR($t$) is the true positive rate at threshold $t$ and FPR($t$) is the false positive rate at threshold $t$. ROC-AUC measures how well the model ranks monomer pairs by their predicted probability of film formation. A high ROC-AUC means true film-forming reactions tend to receive higher scores than non-forming ones to prioritize candidates for follow-up testing. We can choose a ROC-AUC threshold to balance the acceptable false positive rate against the desired hit rate.

The F1 score is the harmonic mean of precision and recall:

$$\text{F1 score} = 2 \cdot \frac{\text{Precision} \cdot \text{recall}}{\text{Precision} + \text{recall}} \tag{9}$$

where:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

and TP, FP, and FN are the numbers of true positives, false positives, and false negatives, respectively. In our reaction screening context, a high precision minimizes wasted effort on false positives (monomer pairs that did not form a free-standing membrane), while a high recall ensures that the model does not overlook promising film-forming reactions. The F1 score is the trade-off between experimental cost and discovery rate.

The accuracy is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{12}$$

where TN is the number of true negatives. Accuracy simply defines in an absolute value, how many film formations (either positive or negative) were predicted correctly. Accuracy in our case is prone to be overestimated because of the imbalanced dataset.

The binary cross-entropy (BCE) loss is defined as:

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N} [\, y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{13}$$

where $y_i$ is the true label and $p_i$ is the predicted probability for the $i$-th sample. The BCE was used to calculate the loss term in both the GNN and the XGB models' training. All the other metrics (F1, ROC AUC, MCC) were monitored during the trainin but were not used as a loss function.

### 4.8 Graph neural network

The directed message passing graph convolution neural network (GNN) was based on the Chemprop v2 python package.[52]

Table 5 Optimal hyperparemeters for each fold for the GNN model after hyperparameter optimization

| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
|---|---|---|---|---|---|
| Depth | 5 | 3 | 3 | 5 | 5 |
| Message_hidden_dim | 900 | 600 | 500 | 800 | 800 |
| Dropout | 0 | 0 | 0 | 0 | 0 |
| Activation | RELU | LEAKYRELU | RELU | LEAKYRELU | RELU |
| ffn_num_layers | 1 | 2 | 2 | 2 | 2 |
| ffn_hidden_dim | 900 | 1100 | 1000 | 500 | 900 |

The inputs were the monomer SMILES which are internally transformed into RDKit molecular graphs by Chemprop. Monomer concentration, one-hot encoded solvent types, and the DFT results were added as additional inputs. The initial atom (atomic number, number of bonds to other atoms, formal charge, hybridization, aromaticity, *etc.*) and bond features (conjugation, in-a-ring, stereochemical information) are passed to the directed message passing algorithm as detailed in the original publication.[52] After hyperparemeter optimization for each five folds separately, the best settings for each fold are detailed in Table 5. All other hyperparameters were set to default values. The Chemprop v2 built-in optuna was used for hyperparameter optimization. Class imbalance was handled by providing the network with an auxiliary dictionary of instance-level class weights, computed from the positive-to-negative ratio.

### 4.9 Extreme gradient boosting

The XGB package was used to perform the model optimization in python.[55] The same cross-validation folds were used as in the GNN model. XGB is an ensemble technique which trains several weak learners sequentially to predict the output label and the final prediction is the weighted sum of these weak learners. The weak learners are decision trees and the boosting refers to the sequential training in XGB: the subsequent learners are aimed to reduce the error of the previous learners. XGB cannot handle molecular-graph objects natively; therefore, the corresponding 128-length Morgan fingerprint vectors were used as input. All other inputs, such as the DFT, one-hot encoded solvent type, and concentration values were concatenated to the Morgan fingerprints and used as inputs. The task was binary classification to predict film formation or no film formation from the given input features. XGB is an ensemble technique which trains several weak learners sequentially to predict the output label and the final prediction is the weighted sum of these weak learners. The weak learners are decision trees and the boosting refers to the sequential training in XGB: the subsequent learners are aimed to reduce the error of the previous learners. We used a bayesian optimization method implemented using optuna[58] to find the best hyperparameters for the XGB model. The hyperparameters are detailed in Table S1. Class imbalance was handled using the built-in 'scale_pos_weight' parameter, set based on the positive-to-negative class ratio.

### 4.10 Logistic regression

A logistic regression model was built using scikit-learn (v1.5.2).[59] The data was scaled using $z_i = (x_i - \mu)/\sigma$ where $z_i$

and $x_i$ are the scaled and original features, $\mu$ is the sample mean and $\sigma$ is the standard deviation. The fitting run for a 1000 iterations with a 'liblinear' solvent in scikit-learn. The output values were scored using BMMC, accuracy, F1-score and ROC-AUC, similarly to the other models. In logistic regression, the model optimizes the probability function to predict the film formation outcome ($P(Y = 1|X)$):

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\mathbf{x}^T\mathbf{X})}}$$

where **w** and **X** are the model weights and features, respectively. Logistic regression is a linear model, and the logits of the prediction probabilities are the linear combinations of the weights and features. This results in highly explainable model by examining the relation of $e^{w_i}$ to 1. We used L1 penalty during the training. The average feature importance was calculated between the 5 fold cross validation test results. Class imbalance was mitigated by adjusting the loss contribution of each class according to the ratio of positive to negative samples.

### 4.11 Support vector machine

The SVM model was built using scikit-learn (v1.5.2).[59] The SVM is a supervised model that is tasked to find the maximal marginal hyperplane between two clusters of datapoints.[60] We used bayesian optimization, implemented in optuna[58] to search for the optimal hyperparameters ($C$ and gamma) separately for a linear SVM (SVM-linear) and one with a radial basis function kernel trick SVM (SVM-rbf). The linear SVM tries to minimize the function $\gamma$ given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$:

$$\gamma = \min_{i=1,\ldots,m} \gamma^{(i)} \left( \left( \frac{w}{\|w\|} \right)^T x^{(i)} + \frac{b}{\|w\|} \right) \quad (14)$$

where $w$ are the weights and $x^i$ are the features. For the kernel trick, we used the Gaussian kernel:

$$K(x, z) = \exp\left( -\frac{\|x - z\|^2}{2\sigma^2} \right) \quad (15)$$

where $\sigma$ is a hyperparameter. Class imbalance was handled *via* automatically adjusting the $C$ penalty value for the minority class (positive film formation).

### 4.12 Naive Bayes classifier

The Naive Bayes classifier was used to predict the outcome of the film formation. The model used a Bernoulli Naive Bayes-based approach to assign a class $\hat{y} = C_k$ for any $k$:

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \quad (16)$$

where $n$ is the number of samples. The probability ($p$) of the Gaussian Naive Bayes classifier is modeled using:

$$p(x = v|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} \quad (17)$$

where $\sigma$ and $\mu$ are the variance and mean, respectively.

The probabilty of the Bernoulli Naive Bayes classifier is modeled using:

$$p(\mathbf{x}|C_k) = \prod_{i=1}^{n} p_{ki}^{x_i} (1 - p_{k_i})^{(1-x_i)}. \quad (18)$$

Class imbalance was mitigated by adjusting the prior probability term for the minority class.

### 4.13 Image classification data

We conducted our experiments using the limited and imbalanced dataset D4, as detailed in Table 1. This dataset comprises 486 optical images representing five distinct film types: non-crystalline dense, dendritic, crystalline, dense gel-like, and non-crystalline dense with defects. As illustrated in Fig. 2F, there is a notable imbalance, with the non-crystalline dense class substantially outweighing the other classes. To ensure a thorough and rigorous evaluation, we employed a cross-validation strategy with $K = 4$ folds, carefully ensuring that the validation sets from each fold were mutually exclusive. For each fold, the dataset was split into training and validation sets, maintaining an 80% to 20% ratio. To address the pronounced class imbalance, we preserved the original distribution of classes in both the training and validation sets by applying the respective class proportions.

## Author contributions

GI designed the methodology, built the ML models for film prediction, analyzed the data, wrote the original manuscript and edited the revisions. MIB performed all chemical reactions and material characterizations. KG performed the DFT calculations. AV performed image classifications and analyzed the results. TS trained ML models and performed feature importance calculations; edited the manuscript. SN, BG, SK and GS conceptualized and supervised the project, administered the data, secured funding and co-wrote the manuscript.

## Conflicts of interest

Authors declare that they have no competing interests.

## Data availability

Supplementary information: AFM images with height data, optical images, the complete monomer list used in the thin-film formation reactions and the film-formation data table are available in the supplementary materials (SI). See DOI: **https://doi.org/10.1039/d5mh01420d**.

D1–D4 datasets can also be downloaded from **https://www.osndatabase.com/datasets** under the section "TFC Film Formation Datasets" (fully open access).

## Acknowledgements

The research reported in this publication was supported by funding from the King Abdullah University of Science and Technology (KAUST) under Near Term Grand Challenge – AI under Award No. ORA-2022-5240, Center of Excellence for Generative AI under Award No. 5940, and BAS/1/1401-01-01. We acknowledge the KAUST Supercomputing Laboratory for providing computational resources of the Ibex Cluster and Shaheen II.

## Notes and references

1 F. Zhang, J. Fan and S. Wang, Interfacial Polymerization: From Chemistry to Functional Materials, *Angew. Chem., Int. Ed.*, 2020, **59**(49), 21840–21856, DOI: **10.1002/anie.201916473**.

2 Y. Song, J.-B. Fan and S. Wang, Recent progress in interfacial polymerization, *Mater. Chem. Front.*, 2017, **1**(6), 1028–1040, DOI: **10.1039/c6qm00325g**.

3 M. F. Jimenez-Solomon, Q. Song, K. E. Jelfs, M. Munoz-Ibanez and A. G. Livingston, Polymer nanofilms with enhanced microporosity by interfacial polymerization, *Nat. Mater.*, 2016, **15**(7), 760–767, DOI: **10.1038/nmat4638**.

4 M. E. Alf, *et al.*, Chemical Vapor Deposition of Conformal, Functional, and Responsive Polymer Films, *Adv. Mater.*, 2010, **22**(18), 1993–2027, DOI: **10.1002/adma.200902765**.

5 D. Lee, M. F. Rubner and R. E. Cohen, All-Nanoparticle Thin-Film Coatings, *Nano Lett.*, 2006, **6**(10), 2305–2312, DOI: **10.1021/nl061776m**.

6 L. Lin, R. Lopez, G. Z. Ramon and O. Coronell, Investigating the void structure of the polyamide active layers of thin-film composite membranes, *J. Membr. Sci.*, 2016, **497**, 365–376, DOI: **10.1016/j.memsci.2015.09.020**.

7 R. Tan, *et al.*, Thin Film Composite Membranes with Regulated Crossover and Water Migration for Long-Life Aqueous Redox Flow Batteries, *Adv. Sci.*, 2023, **10**(20), 2206888, DOI: **10.1002/advs.202206888**.

8 T. H. Lee, *et al.*, Hyperaging-induced H2-selective thin-film composite membranes with enhanced submicroporosity toward green hydrogen supply, *J. Membr. Sci.*, 2023, **672**, 121438, DOI: **10.1016/j.memsci.2023.121438**.

9 B. Alhazmi, *et al.*, Ultraselective Macrocycle Membranes for Pharmaceutical Ingredients Separation in Organic Solvents, *Nat. Commun.*, 2024, **15**, 7151, DOI: **10.1038/s41467-024-51548-7**.

10 Z. Wen, D. Pintossi, M. Nuño and T. Noël, Membrane-based TBADT recovery as a strategy to increase the sustainability of continuous-flow photocatalytic HAT transformations, *Nat. Commun.*, 2022, **13**, 6147, DOI: **10.1038/s41467-022-33821-9**.

11 C. Jin, *et al.*, Fabrication of Coffee-Ring Nanostructured Membranes for Organic Solvent Nanofiltration, *Angew. Chem.*, 2024, **63**, e202405891, DOI: **10.1002/ange.202405891**.

12 G. Ignacz, A. K. Beke and G. Szekely, Data-driven investigation of process solvent and membrane material on organic solvent nanofiltration, *J. Membr. Sci.*, 2023, **674**, 121519, DOI: **10.1016/j.memsci.2023.121519**.

13 J. Hu, G. Ignacz, R. Hardian and G. Szekely, Triazinane-based thin-film composite membrane fabrication *via* heterocycle network formation from formaldehyde and amines for organic solvent nanofiltration, *J. Membr. Sci.*, 2023, **679**, 121701, DOI: **10.1016/j.memsci.2023.121701**.

14 M. Fetanat, *et al.*, Machine learning for design of thin-film nanocomposite membranes, *Sep. Purif. Technol.*, 2021, **270**, 118383, DOI: **10.1016/j.seppur.2021.118383**.

15 D. Rall, *et al.*, Rational design of ion separation membranes, *J. Membr. Sci.*, 2019, **569**, 209–219, DOI: **10.1016/j.memsci.2018.10.013**.

16 D. Rall, *et al.*, Simultaneous rational design of ion separation membranes and processes, *J. Membr. Sci.*, 2020, **600**, 117860, DOI: **10.1016/j.memsci.2020.117860**.

17 M. Wang, G. M. Shi, D. Zhao, X. Liu and J. Jiang, Machine Learning-Assisted Design of Thin-Film Composite Membranes for Solvent Recovery, *Environ. Sci. Technol.*, 2023, **57**(42), 15914–15924, DOI: **10.1021/acs.est.3c04773**.

18 C. Wang, L. Wang, A. Soo, N. Bansidhar Pathak and H. Kyong Shon, Machine learning based prediction and optimization of thin film nanocomposite membranes for organic solvent nanofiltration, *Sep. Purif. Technol.*, 2023, **304**, 122328, DOI: **10.1016/j.seppur.2022.122328**.

19 B. Sutariya, P. Sarkar, P. D. Indurkar and S. Karan, Machine learning-assisted performance prediction from the synthesis conditions of nanofiltration membranes, *Sep. Purif. Technol.*, 2025, **354**, 128960, DOI: **10.1016/j.seppur.2024.128960**.

20 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, Prediction of Organic Reaction Outcomes Using Machine Learning, *ACS Cent. Sci.*, 2017, **3**(5), 434–443, DOI: **10.1021/acscentsci.7b00064**.

21 S. Van Buggenhout, *et al.*, Open and FAIR data for nanofiltration in organic media: A unified approach, *J. Membr. Sci.*, 2025, **713**, 123356, DOI: **10.1016/j.memsci.2024.123356**.

22 Q.-F. An, *et al.*, Microstructural characterization and evaluation of pervaporation performance of thin-film composite membranes fabricated through interfacial polymerization on hydrolyzed polyacrylonitrile substrate, *J. Membr. Sci.*, 2019, **583**, 31–39, DOI: **10.1016/j.memsci.2019.04.050**.

23 A. K. Ghosh, B.-H. Jeong, X. Huang and E. M. Hoek, Impacts of reaction and curing conditions on polyamide composite reverse osmosis membrane properties, *J. Membr. Sci.*, 2008, **311**(1–2), 34–45, DOI: **10.1016/j.memsci.2007.11.038**.

24 S. H. Kim, S.-Y. Kwak and T. Suzuki, Positron Annihilation Spectroscopic Evidence to Demonstrate the Flux-Enhancement Mechanism in Morphology-Controlled Thin-Film-Composite (TFC) Membrane, *Environ. Sci. Technol.*, 2005, **39**(6), 1764–1770, DOI: **10.1021/es049453k**.

25 A. Rácz, D. Bajusz and K. Héberger, Modelling methods and cross-validation variants in QSAR: a multi-level analysis, *SAR QSAR Environ. Res.*, 2018, **29**(9), 661–674, DOI: **10.1080/1062936x.2018.1505778**.

26 D. Chicco, N. Tötsch and G. Jurman, The Matthews correlation coefficient (MCC) is more reliable than balanced

accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation, *BioData Min.*, 2021, **14**, 13, DOI: 10.1186/s13040-021-00244-z.

27 A. Y. Ng, *Feature selection, L1 vs. L2 regularization, and rotational invariance*, in *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, Association for Computing Machinery, New York, NY, USA, 2004, p. 78, DOI: 10.1145/1015330.1015435.

28 K. He, X. Zhang, S. Ren and J. Sun, *Deep Residual Learning for Image Recognition*, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, DOI: 10.1109/CVPR.2016.90.

29 M. Tan and Q. V. Le, *EfficientNetV2: Smaller Models and Faster Training*. CoRR abs/2104.00298, 2021.

30 J. Deng, *et al.*, *ImageNet: A large-scale hierarchical image database*, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255, DOI: 10.1109/CVPR.2009.5206848.

31 R. R. Selvaraju, *et al.*, *via*. CoRR abs/1610.02391, 2016.

32 J. E. Cadotte, *Interfacially synthesized reverse osmosis membrane*, 1981.

33 H. Gao, *et al.*, Revolutionizing Membrane Design Using Machine Learning-Bayesian Optimization, *Environ. Sci. Technol.*, 2021, **56**(4), 2572–2581, DOI: 10.1021/acs.est.1c04373.

34 A. Zunger, Inverse design in search of materials with target functionalities, *Nat. Rev. Chem.*, 2018, **2**, 0121, DOI: 10.1038/s41570-018-0121.

35 B. Kim, S. Lee and J. Kim, Inverse design of porous materials using artificial neural networks, *Sci. Adv.*, 2020, **6**, eaax9324, DOI: 10.1126/sciadv.aax9324.

36 K. Sattari, Y. Xie and J. Lin, Data-driven algorithms for inverse design of polymers, *Soft Matter*, 2021, **17**(33), 7607–7622, DOI: 10.1039/d1sm00725d.

37 S. M. Lundberg, *et al.*, From local explanations to global understanding with explainable AI for trees, *Nat. Mach. Intell.*, 2020, **2**(1), 56–67, DOI: 10.1038/s42256-019-0138-9.

38 S. Ivanov and L. Prokhorenkova, *Boost then Convolve: Gradient Boosting Meets Graph Neural Networks*, *arXiv*, 2021, preprint, arXiv:2101.08543, DOI: 10.48550/arXiv.2101.08543.

39 H. Zhao, *et al.*, A robotic platform for the synthesis of colloidal nanocrystals, *Nat. Synth.*, 2023, **2**(6), 505–514, DOI: 10.1038/s44160-023-00250-5.

40 D. Caramelli, *et al.*, Discovering New Chemistry with an Autonomous Robotic Platform Driven by a Reactivity-Seeking Neural Network, *ACS Cent. Sci.*, 2021, **7**(11), 1821–1830, DOI: 10.1021/acscentsci.1c00435.

41 J. Park, *et al.*, Closed-loop optimization of nanoparticle synthesis enabled by robotics and machine learning, *Matter*, 2023, **6**(3), 677–690, DOI: 10.1016/j.matt.2023.01.018.

42 J. Noh, *et al.*, Inverse Design of Solid-State Materials *via* a Continuous Representation, *Matter*, 2019, **1**(5), 1370–1384, DOI: 10.1016/j.matt.2019.08.017.

43 J. Liu, *et al.*, Smart covalent organic networks (CONs) with "on-off-on" light-switchable pores for molecular separation, *Sci. Adv.*, 2020, **6**, eabb3188, DOI: 10.1126/sciadv.abb3188.

44 I. Volokitina, A. Volokitin and E. Panin, Gradient microstructure formation in carbon steel bars, *J. Mater. Res. Technol.*, 2024, **31**, 2985–2993, DOI: 10.1016/j.jmrt.2024.07.038.

45 B. Ma, *et al.*, Hot deformation behavior of GH4169 superalloy with high proportion of recycled material addition and initial dendrite structure, *J. Alloys Compd.*, 2024, **1007**, 176352, DOI: 10.1016/j.jallcom.2024.176352.

46 Y. Zhao and D. G. Truhlar, The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals, *Theor. Chem. Acc.*, 2007, **120**(1–3), 215–241, DOI: 10.1007/s00214-007-0310-x.

47 Y. Zhao and D. G. Truhlar, Density Functionals with Broad Applicability in Chemistry, *Acc. Chem. Res.*, 2008, **41**(2), 157–167, DOI: 10.1021/ar700111a.

48 R. Krishnan, J. S. Binkley, R. Seeger and J. A. Pople, Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions, *J. Chem. Phys.*, 1980, **72**(1), 650–654, DOI: 10.1063/1.438955.

49 M. J. Frisch, *et al.*, *Gaussian ~ 16 Revision C.01*, Gaussian Inc., Wallingford CT, 2016.

50 R. Dennington, T. A. Keith and J. M. Millam, *GaussView Version 6*, Semichem Inc., Shawnee Mission KS, 2019.

51 S. Boys and F. Bernardi, The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors, *Mol. Phys.*, 1970, **19**(4), 553–566, DOI: 10.1080/00268977000101561.

52 E. Heid, *et al.*, Chemprop: A Machine Learning Package for Chemical Property Prediction, *J. Chem. Inf. Model.*, 2023, **64**(1), 9–17, DOI: 10.1021/acs.jcim.3c01250.

53 K. Yang, *et al.*, Analyzing Learned Molecular Representations for Property Prediction, *J. Chem. Inf. Model.*, 2019, **59**(8), 3370–3388, DOI: 10.1021/acs.jcim.9b00237.

54 W. Jin, R. Barzilay and T. Jaakkola, *Multi-Objective Molecule Generation using Interpretable Substructures*, 2020.

55 T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16 (ACM), 2016, DOI: 10.1145/2939672.2939785.

56 G. Landrum, *et al.*, *rdkit/rdkit: 2020_03_1 (Q1 2020) Release*, 2020, DOI: 10.5281/zenodo.3732262.

57 D. Chicco and G. Jurman, The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification, *BioData Min.*, 2023, **16**, 4, DOI: 10.1186/s13040-023-00322-4.

58 T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, *Optuna: A Next-generation Hyperparameter Optimization Framework*, 2019.

59 F. Pedregosa, *et al.*, Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

60 V. N. Vapnik, *The Support Vector method*, Springer, Berlin Heidelberg, 1997, p. 261–271, DOI: 10.1007/bfb0020166.