Virtual High-Throughput Screening of Vapor-Deposited Amphiphilic Polymers for Inhibiting Biofilm Formation

Zhihao Feng[1]†, *Yifan Cheng*[1]†, *Alexandra Khlyustova*[1]†, *Aasim Wani*[1], *Trevor Franklin*[1], *Jeffrey D. Varner*[1]*, *Andrew L. Hook*[2]*, *Rong Yang*[1]*

[1] Robert Frederick Smith School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, NY 14853, U.S.A.
[2] Advanced Materials and Healthcare Technologies, University of Nottingham, Nottingham, NG7 2RD, U.K.

*Communication authors contact information:
Prof. Rong Yang, ryang@cornell.edu
Prof. Andrew L. Hook, andrew.hook@nottingham.ac.uk
Prof. Jeffrey D. Varner, jdv27@cornell.edu
†These authors contributed equally to this work

Keywords

Amphiphilic copolymers, high-throughput screening, machine learning, iCVD, antibiofilm

Abstract

Amphiphilic copolymers (AP) represent a class of novel anti-biofouling materials whose chemistry and composition can be tuned to optimize their performance. However, the enormous chemistry-composition design space associated with AP makes their performance optimization laborious; it is not experimentally feasible to assess and validate all possible AP compositions even with the use of rapid screening methodologies. To address this constraint, we report a robust model development paradigm, yielding a versatile machine learning approach that accurately predicts biofilm formation by Pseudomonas aeruginosa on a library of AP. The model excels in extracting underlying patterns in a “pooled” dataset from various experimental sources, thereby expanding the design space accessible to the model to a much larger selection of AP chemistries and compositions. The model is used to screen virtual libraries of AP for identification of best-performing candidates for experimental validation. Initiated chemical vapor deposition (iCVD) was used for the precision synthesis of the model-selected AP chemistries and compositions for validation at solid-liquid interface (often used in conventional antifouling studies), as well as the air-liquid-solid triple interface. Despite the vastly different growth conditions, the model successfully identified the best-performing AP for biofilm inhibition at the triple interface.

1. Introduction

Biofilms formed on the surface of indwelling medical devices are the root cause of many life-threatening nosocomial infections.[1] Compared to their planktonic counterparts, bacteria in these surface-bound biofilm 'fortresses' can be up to 1000-fold more resistant to antibiotics and host defenses.[2] Furthermore, biofilms can facilitate the dissemination of drug resistance genes across multiple co-inhabiting species via horizontal gene transfer.[3] As such, biofilms further challenge humanity's capability to stay ahead in the race against emerging bacterial pathogens that already resist multiple last-line antibiotics.[4]

Polymer coatings that inherently resist bacterial attachment and subsequent biofilm formation, namely anti-biofouling coatings, have been brought to the forefront of combating the formation of antibiotic-resistant biofilms. Rather than killing bacteria, as in the case of antibiotics, which inevitably breed resistant strains over time, these novel coatings prevent biofilm by disrupting bacterial adhesion to a solid surface, thus imposing minimal selective pressure on bacteria.[5,6] Successful examples of such materials include poly(ethylene glycol) brushes,[7] zwitterionic polymers,[8] liquid-infused surfaces,[9] and amphiphilic copolymers (AP).[10–15] Here, we focus on AP for their capability in mitigating bacterial attachment and biofilm formation at both solid-liquid interfaces and solid-liquid-gas triple interfaces, the latter of which has received far less attention despite their implication in nosocomial material-associated infections.[12]

While the anti-biofouling performance of most polymer coatings has been attributed to hydrophilicity, i.e., the increased enthalpic barrier for foulant adhesion, the fundamental mechanism for AP's fouling resistance is not well understood. The hypotheses that have received most attention include Baier curve minimum,[16] dynamic surface reconstruction,[12,17] and nanoscale or molecular heterogeneity.[10,14,18,19] Based on these theories, comonomers from the opposite ends of the surface energy spectrum (e.g., pairing of hydrophilic zwitterionic monomers with hydrophobic fluorinated monomers)[12] have been selected to form AP, which create high surface energy mismatch and thus high thermodynamic penalty upon bacterial contact. Indeed, these AP coatings have demonstrated superior anti-biofilm efficacy than zwitterionic coatings.[12]

Despite their great promise, the discovery of novel and effective AP is limited by the lack of a set of guiding principles, which constitutes a major barrier due to the immense chemistry-composition design space associated with AP. There are countless combinations of hydrophilic and hydrophobic monomer pairs and compositional variations,[19,20] making a comprehensive assessment extremely laborious. To address that challenge, we resort to data-driven machine learning (ML) models, which have demonstrated considerable potential in predicting biological responses of materials, including biofilm formation.[21] To generate a sufficiently large dataset for ML models, we adopted a high-throughput screening platform reported by Hook et al., which has been leveraged to discover anti-biofouling copolymers that are unanticipated based on existing theoretical or empirical knowledge.[6,22] More recently, that platform has enabled the quantitative prediction of the attachment by multiple pathogenic bacteria on a variety of polymers, pointing to the discovery of materials with broad-spectrum efficacy.[23] However, despite their proven advantage in rapid synthesis and testing, the high-throughput platform relies on solution-phase polymerization, which limits the discovery of AP with high-contrast and molecular-scale heterogeneity. Due to the lack of a common solvent and the tendency for microphase separation for monomer pairs with contrasting surface energies, solution-based synthesis is considered inappropriate to address the detailed molecular design to advance anti-biofouling AP.

To enable the precision synthesis of AP using monomer pairs with contrasting surface energies, we leverage an all-dry synthesis approach, namely initiated chemical vapor deposition (iCVD). It has been used for AP synthesis due to its solvent-free nature,[24,25] a high retention rate of functional moieties borne in the monomers,[26] and its capability to deposit conformal nanolayers over nanostructured substrates.[27,28] These unique advantages of iCVD enable the synthesis of desired AP chemistries and their one-step synthesis and application on essentially any substrate while maintaining the beneficial bulk properties (e.g., mechanical properties) and desired surface topographical features (e.g., nanostructures).[24,29] Despite its many advantages, it is currently challenging to sift through the vast chemistry-composition design space associated with AP using iCVD alone.[19,20]

This begs the question: can we combine the best of both worlds, whereby solution-phase high-throughput synthesis provides a sufficiently large dataset for training a general ML model, which is also applicable to the vapor-deposited AP materials? Moreover, since AP holds great promise in reducing biofilm formation at both solid-liquid interfaces and solid-liquid-air triple interfaces, can a general ML model transfer what is learned from solid-liquid interfaces to complete similar tasks at the triple interfaces?

Here, we demonstrate that a general support vector regression (SVR), when trained on a diverse polymer library, captures the fundamental quantitative structure-activity relationship (QSAR) underlying both vapor- and solution-synthesized copolymers, thus allowing transfer of the polymer chemistry-biofilm performance knowledge across polymer synthesis methods. We further demonstrate that transfer is effective even across different growth environments for biofilm formation (i.e., liquid-solid versus triple interface). Compared with the state-of-the-art,[23] the model reported here expands the accessible number of molecular fingerprints by over twofold, that of monomers by sixfold, and that of unique polymers by over fivefold, while maintaining a desired accuracy of predicting their antibiofilm performance. To demonstrate the potential impact of the model, we show how QSAR can be leveraged to shed light on the structure-performance correlation (by associating low bacterial attachment with specific monomers and molecular fingerprints borne by polymers) and to guide the subsequent synthesis and testing of promising candidates. The model development and implementation procedures demonstrated here can be directly adopted by others in the field to rapidly discover anti-biofilm chemistries, synthesized using solution- or all-dry method, which are highly effective under different bacteria growth conditions.

2. Methodologies

The capability of a general ML to learn fundamental polymer chemistry-biofilm relationships from numerous and diverse examples is vital to achieving robust predictive performance across polymer synthesis methods and types of interfaces at which biofilm forms. General ML models are characterized by their capability to learn 'deeper' patterns concealed in the high-dimensional feature space with proper abstraction and generalization (e.g., dimension reduction) and subsequently apply such learning to achieve accurate predictions of material properties even when new examples appear different from the original training domain. In this work, we pooled together 2,240 polymers and their fluorescence intensities against *Pseudomonas aeruginosa* (PAO1) (labeled as F_{PA}) from seven microarrays, including the dataset previously published by Hook et al.[6,22] as well as unpublished samples from the same lab. Instead of emphasizing the model's generalization at a range of pathogens similar to previous studies,[6,30] we aim to investigate a model's ability to generalize with data obtained from independent experimental sources and possible transferability of learned patterns to a distinct polymerization strategy. Thus, fluorescence intensities against PAO1 are the only model learning target included in the polymer library. This library included homo- and co-polymers with a broad range of pendant groups, including hydrophilic, hydrophobic, aromatic, cyclic, and branched functional groups. There was an emphasis on copolymers with weakly amphiphilic pendant groups that were a part of the class of poly(meth)acrylates previously shown to prevent bacterial biofilm formation.[31] The polymer synthesis procedure and biofilm formation measurement are illustrated in "Experimental Data Collection" of **Figure 1**. Also, the corresponding details are described in Experimental Section. We cannot exclude that specific copolymer architectures were generated on the microarray as a result of the polymerization strategy used that would limit the ability to maintain the biological properties measured on the microdots upon scale-up. However, this is unlikely to be a dominant effect, as materials selected from the microarray have been successfully scaled up whilst maintaining antibiofilm properties using a range of polymerization strategies.[22,32]

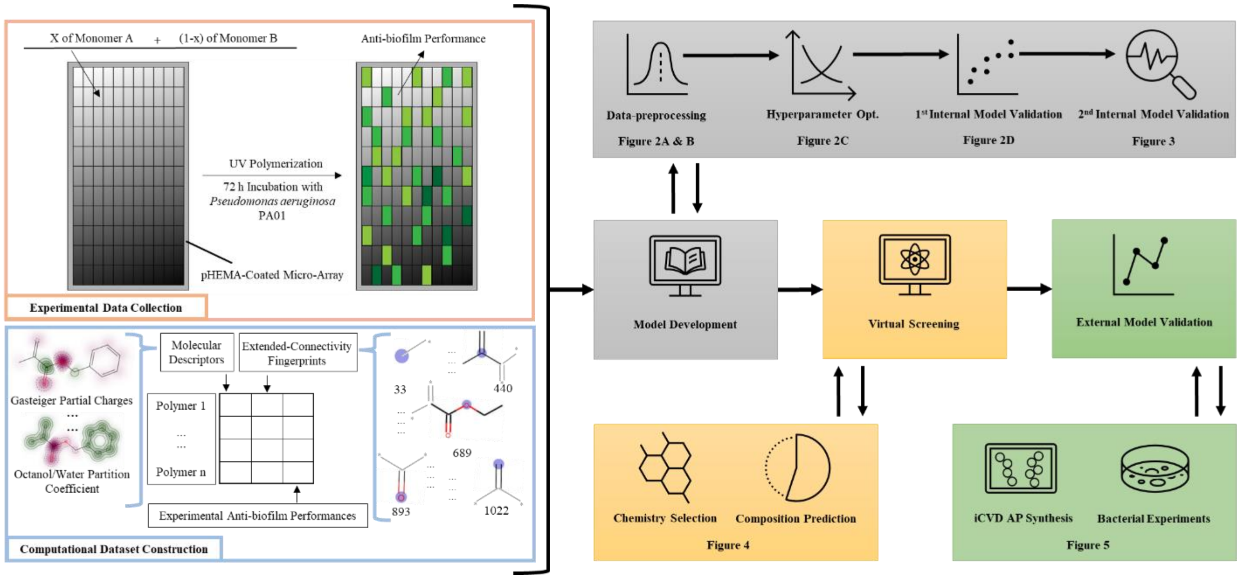

Figure 1. Machine Learning Model Development Pipeline. This pipeline is divided into five sections with a color code. Referenced figures are indicated correspondingly. During "Experimental Data Collection" (red frame), a polymer that contained component A and B was achieved by dispensing x (volume fraction) of monomer A and (1-x) of B in a mixture onto glass slides to form microarrays. Upon incubation of microarrays with PAO1 for 72 hours, the microarrays' fluorescence intensities were measured to quantify the biofilm formation on 2,240 polymers (after removing the data below the limit of detection) The impact of applying a limit of detection on the dataset has been described previously.[22] With the experimental data, "Computational Dataset Construction" (blue frame) was performed with molecular descriptors

(e.g., Gasteiger partial charges, octanol/water partition coefficient, and other 95 descriptors) and 1,024 extended-connectivity fingerprints with a linear summation of two corresponding monomer components' features for every copolymer's computational feature entry against its experimental antibiofilm performance. In "Model Development" (gray blocks), we preprocessed the training set followed by hyperparameter optimizations of support vector regressors. Subsequently, we generated an ensemble model that was trained under various autoencoder-encoded feature sets for the internal model validation (e.g., at the solid-liquid interfaces) with the hold-out test set and previously observed AP. In "Virtual Screening" (yellow blocks), the ensemble model virtually screened a newly built AP dataset for antibiofilm performance with ranked candidates. Among the top-ranked candidates, compositional optimization was performed using the ensemble model. Finally, in "External Model Validation" (green blocks), we synthesized the model-predicted AP using iCVD and tested their antibiofilm performance at liquid-solid and triple interfaces.

Due to the structural complexity of the pooled polymers, we computed each polymer's features with a linear combination of two corresponding monomer components' features used for this polymer. Therefore, the following steps were carried out: (i) we converted 137 unique monomers (used to derive all polymers) into their corresponding simplified molecular-input line-entry (SMILE) strings for generating computational features. These features include Gasteiger partial charges, octanol/water partition coefficient, and other 95 descriptors, as well as 1,024 extended-connectivity fingerprints[33] with RDKit. (ii) we linearly combined these monomers' features pairwise to derive all polymers' features. The dataset's structure is presented in "Computational Dataset Construction" in **Figure 1.**

Previously, the Bayesian neural network has learned from a polymer microarray comprising 404 unique homo- and co-polymers derived from 22 monomers.[23] We expanded the number of individual monomers included in training from the existing record by six-fold (from 22 to137), unique molecular fingerprints by more than two-fold (from 183 to 423), and unique homopolymers and copolymers by more than five-fold (from 404 to 2,240).[23] This considerable expansion in the size and diversity of the training set could bring the following potential challenges to the development of a general model: (i) microarray results obtained from multiple independently performed experiments could introduce variables that were not explicitly controlled (e.g., slight differences in experimenter's operation); (ii) diversity in monomer chemistry could make accurate prediction more difficult, as unbalanced sample distribution might be embedded into this dataset. To tackle these challenges, a generalizable algorithm was desired. Therefore, we decided to build a radial-basis-function-kernelized support vector regression (SVR) model, which is presumably more generalizable due to its ε-insensitive region.[34]

Therefore, we aim to (i) train an SVR ensemble model generalized enough to predict the antibiofilm performances of unseen copolymers comprising diverse monomer chemistries and compositions, making use of the bacterial attachment dataset obtained through high-throughput microarray experimentation carried out at the solid-liquid interface ("Model Development" in Figure 1); (ii) apply this trained ensemble model to virtually screen new hydrophilic-hydrophobic monomer pairs for top-ranked amphiphilic copolymers based on their predicted performance and optimize the hydrophilic-to-hydrophobic ratio of selected comonomer pairs ("Virtual Screening" in Figure 1); (iii) test model transferability across synthetic methods by using iCVD-synthesized AP and transferability across interfacial environments for biofilm formation by predicting biofilms formed on AP at solid-liquid-gas interfaces using a model trained at solid-liquid interfaces ("External Model Validation" in Figure 1). In particular, we categorize model validations at the solid-liquid interfaces as the internal validation and the solid-liquid-gas interface as the external validation. Also, we focus on each of the last three sections sequentially in the rest of the paper.

3. Model Development and Internal Model Validation at the Solid-Liquid Interface

Upon completing the computational dataset construction for the polymers, we performed data preprocessing and hyperparameter optimization. Besides the details mentioned in Experimental Section (e.g., the training-test set split, feature rescaling, etc.), some critical components taken to validate the ensemble model internally are highlighted here. (i) In **Figure 2**A**,** log-transformation was applied to the fluorescence intensity distribution (labeled as $\mathrm{Log}(F_{\mathrm{PA}})$) due to the left-skewed issue in the original distribution. (ii) Three identical autoencoders were trained at each number of reduced features with the signature architecture (**Figure 2**B) to tackle the instabilities caused by the autoencoders' local optima and the number of reduced features. (iii) To trade off the bias and variance issue, the range of reduced features (e.g., between 45-49) was selected, as shown in **Figure 2**C. The ensemble model consisted of separately trained 15 SVR models, which were developed corresponding to three autoencoders times five reduced feature numbers.

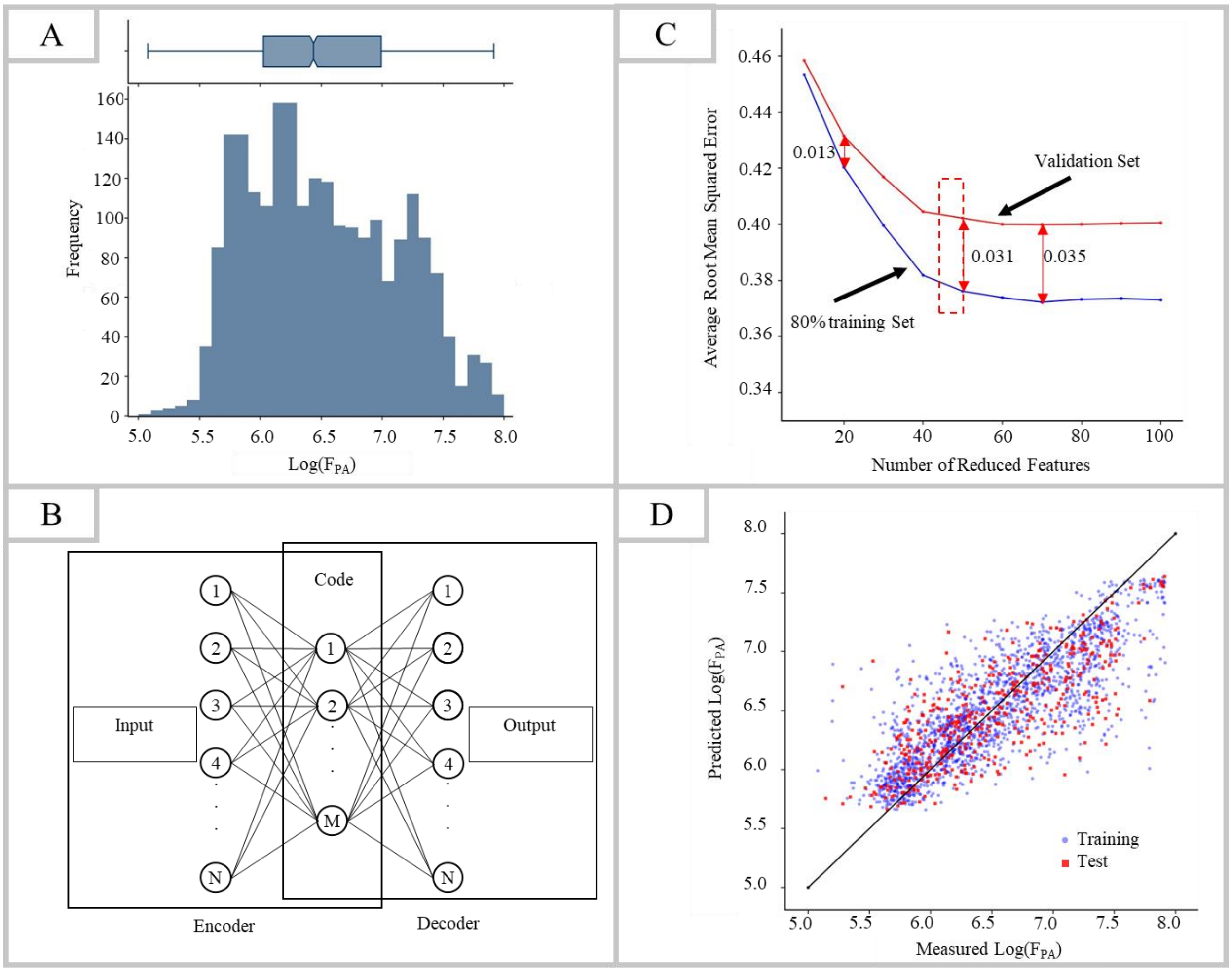

Figure 2. Key Components of Model Development. A. The distribution of the experimentally measured fluorescence intensities after logarithmic transformation. B. Architecture of the three-layer autoencoder, where the number of nodes in the input and output layers is equal to the original feature dimensions (N=1,121). The code layer has M nodes (equal to the desired number of reduced features) chosen based on a bias-variance tradeoff in five-fold cross-validation over the autoencoder-encoded training set. C. A variance-bias tradeoff plot indicated an acceptable range for the number of reduced features (45-49) for balancing the variance and bias issue. D. The ensemble model's internal validation over the holdout test set. In each hyperparameter optimization step over the autoencoder-coded training set, the model was assessed through the averaged root mean squared error (RMSE), the averaged coefficient of determination (R^2), and their standard deviations as the evaluation metrics with the five-fold cross-validation (CV). To

demonstrate the results of the bias-variance tradeoff in the five-fold CV, Table 1 presents a typical model benchmark, using 47 reduced features and one of the three autoencoders as an example. The difference between the averaged RMSE in the 80% training, and validation sets fell within the desired range (e.g., ~ 0.03 $\mathrm{Log}(F_{\mathrm{PA}})$) to minimize the variance issue while maintaining acceptable errors in both sets. For instance, this model could cause up to 5% to 7% errors when predicting a polymer's antibiofilm performance with an average $\mathrm{Log}(F_{\mathrm{PA}})$ (~ 6.0) in the collected dataset. Although this model showed a lower R^2 compared to the previous model's benchmark (e.g., $R^2 > 0.85$),[23] we hypothesized that our model could perform better over the polymers out of the training domain since it was developed over a much more diverse training space as stated earlier.

Table 1: Benchmark of Five-Fold CV (± standard deviation)

Metrics	Benchmark
Average RMSE for 80% training [$\mathrm{Log}(F_{\mathrm{PA}})$]	0.3611 (± 0.0069)
Average R^2 for 80 % training	0.6467 (± 0.0110)
Average RMSE for Validation [$\mathrm{Log}(F_{\mathrm{PA}})$]	0.3915 (± 0.0273)
Average R^2 for Validation	0.5826 (± 0.0448)

Upon confirming the bias-variance tradeoff in each model's five-fold CV benchmark, we validated the ensemble model over the holdout test set, namely the first internal validation. These results, as shown in Table 2, are very comparable to the individual model's benchmark in the five-fold CV, which demonstrated the necessity of applying five-fold to foresee the model's performance over the holdout test data. In **Figure 2**D, an encouraging agreement between the measured and predicted quantities is showing in the training $\mathrm{Log}(F_{\mathrm{PA}})$ range. However, errors appear over the ranges (e.g., approaching both ends of the training range) with lack of training samples more significantly compared to the sample-rich range. Thus, we shall be cautious of applying this model to predict a polymer with extreme $\mathrm{Log}(F_{\mathrm{PA}})$.

Table 2: Benchmark of Ensemble Model

Metrics	Benchmark
RMSE for Training [Log(F_{PA})]	0.3567
R^2 for Training	0.6553
RMSE for Holdout Test [Log(F_{PA})]	0.3835
R^2 for Holdout Test	0.5959

Upon passing the first internal model validation using the holdout test set, we retrained the ensemble model with all accessible data to expand its knowledge base and predict other unseen datasets in later model validations. Here, we report that the ensemble model over the entire dataset has RMSE and R^2 of 0.3504 Log(F_{PA}) and 0.6664, respectively.

A second internal model validation was performed using the above ensemble model and a previously reported dataset that measures biofilm formation on AP at the solid-liquid interface.[22] We first defined AP as a copolymer comprising a hydrophobic monomer (with an octanol/water partition coefficient, Log P >2) and a hydrophilic monomer (with Log P < 0). In the reported dataset,[22] one hydrophilic monomer (labeled "A" for consistency with the previous labeling system) paired with four hydrophobic monomers (labeled "9", "14", "11", "12") fulfilled this criterion as shown in **Figure 3**A. The four AP series (i.e., 9A, 14A, 11A, and 12A) each features six compositions, represented by the volume percentage of A that was 0%, 10%, 15%, 20%, 25%, and 30%. These AP series were reserved for model validation and thus excluded from the training dataset, i.e., they represent unseen data to the ensemble model.

As shown in **Figures 3**B and **3C**, the ML model successfully predicted the effect of monomer chemistries on biofilm formation at the solid-liquid interface. The values of Log(F_{PA}) predicted by the model (solid lines in **Figure 3**B) matched closely to the experimental values (individual data points in **Figure 3**B). The prediction benchmarks, i.e., an R^2 value of 0.87 and an RMSE value of 0.16 Log(F_{PA}), also pointed to the excellent quality of the model predictions. Moreover, the model correctly identified the surprising effect of the hydrophilic monomer on biofilm formation, which increased with an increasing amount of A in the AP. In contrast, conventional wisdom considers hydrophilicity desirable for antifouling purposes.[35] Interestingly, the model did not accentuate the fluctuations in $\mathrm{Log}(F_{\mathrm{PA}})$ among neighboring compositions, which implied that the model was generalized based on the structure-activity relationship uncovered from the training set. Taken together, the validation results presented here indicate that the ML model provides quantitatively accurate predictions of the effect of monomer hydrophilicity and composition on biofilm formation at the solid-liquid interface. This capability is at the core of the virtual screening of high-performing AP using a diverse polymer library, as described below.

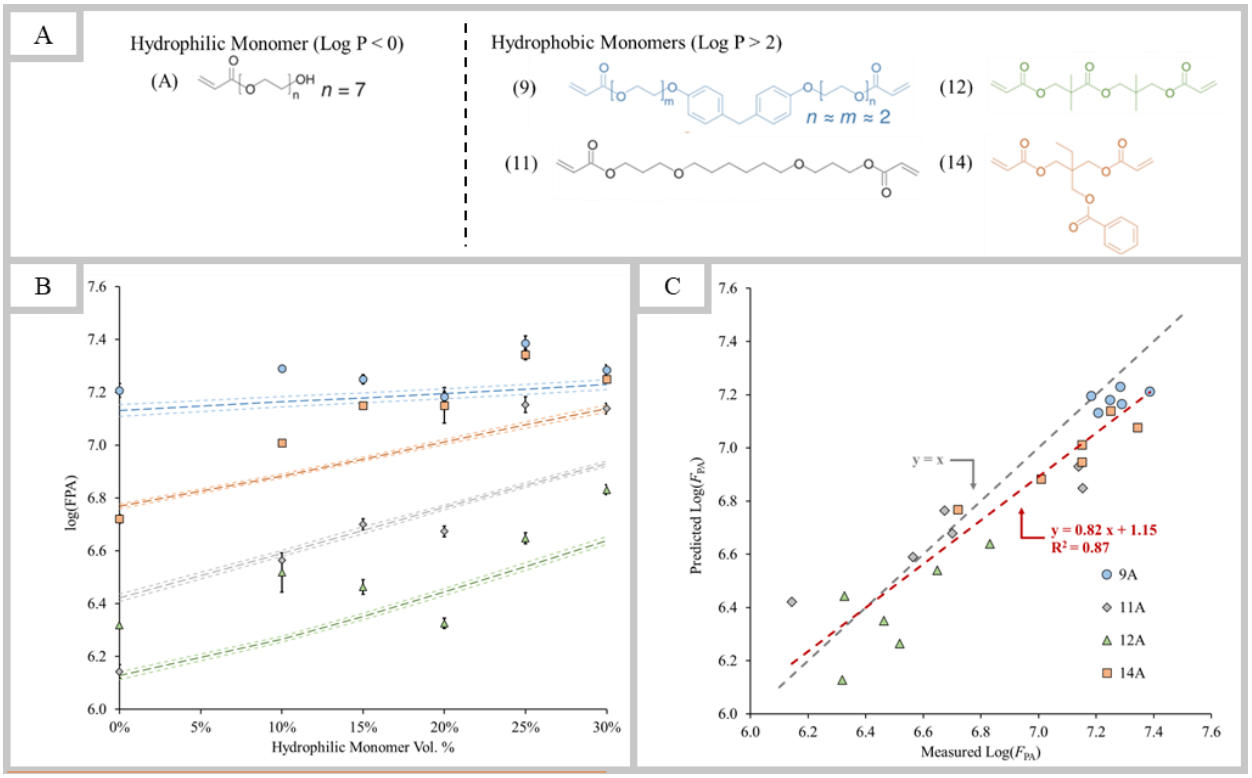

Figure 3. **Experimental validation of ML-predicted biofilm formation on various AP at the solid-liquid interface.** A. Molecular structures of the hydrophilic and hydrophobic monomers used in the validation;[22] B, Comparison of the ML predications (lines) and experimental results (discrete data points) regarding biofilm formation at different film compositions for the four series of AP. The dashed and dotted lines represent the ML-predicted $\mathrm{Log}(F_{\mathrm{PA}})$ values and the standard errors, respectively; the discrete data points with error bars represent the experimental mean $\mathrm{Log}(F_{\mathrm{PA}})$ and standard errors, respectively. C, Comparison of the ML predictions (y-axis) and experimental results (x-axis). The predicted linear line (e.g., R^2 of 0.87) is drew with the red dashed line, while the gray dashed line is the diagonal line (e.g., $y = x$). The four series of AP have the same hydrophilic monomer ("A") and different hydrophobic monomers ("9","11","12", or "14").

4. Virtual Screening

Upon confirming the prediction accuracy of the ensemble model over the internal unseen dataset, we virtually screened all possible AP derived from the 137 unique monomers with various compositions. We first separated these monomers into the hydrophilic (Log P<0) and the hydrophobic (Log P>2) groups, respectively, yielding 61 hydrophobic and five hydrophilic monomers. We subsequently combined one hydrophobic monomer's features with one hydrophilic monomer's features at varying compositions from 10 vol.% to 90 vol.% with an increment of 10 vol.%. The increment was chosen to test the ensemble model’s ability to differentiate the performance of AP within a relatively narrow range of compositions, while maintaining a low computational cost. This choice was unlikely to affect the accuracy of prediction because the experimental uncertainty (e.g., during the quantification of AP composition and biofilm formation) was on the order of 10 vol.%.[22] Even smaller compositional increments (e.g., 1 vol.% or less) are also compatible the model if needed, albeit a drastic increase in computational cost. Therefore, the entire dataset contained 305 AP series (i.e., 61 hydrophobic monomers × 5 hydrophilic monomers) and each AP series included nine compositions with the same combination of hydrophobic-hydrophilic monomers.

We applied the ensemble model to virtually screen the 50 vol.% AP first, anticipating that to be a top-ranked composition (after preprocessing using the recallable scaler and corresponding autoencoders). As a result, the ensemble model predicted that poly(2-hydroxyethyl methacrylate-*co*-hexafluorobutyl acrylate) with 50 vol. % HEMA (abbreviated as pHEMA-*co*-HFBA, **Figure 4**A) would result in antibiofilm performance that ranks among the top 5% out of the 305 AP candidates. This monomer pair was selected for the external validation detailed below also because of their appropriate volatility required for iCVD.[36] We then performed virtual compositional optimization of pHEMA-*co*-HFBA. The predicted Log(F_{PA})-composition correlation appeared to depend on autoencoder seed selection, with an example shown in **Figure 4**B for 45 reduced features, as well as the number of reduced features, as shown in **Figure 4**C. These observations further supported the need for the ensemble model and our approach of constructing the model to improve stability and generalizability. The model predicted that optimum composition of pHEMA-*co*-HFBA for anti-biofilm-formation lied in the range of 60-80 vol.% HEMA. Note that this prediction was made based on training using bacterial quantification data measured at the solid-liquid interfaces from polymer microarray samples.

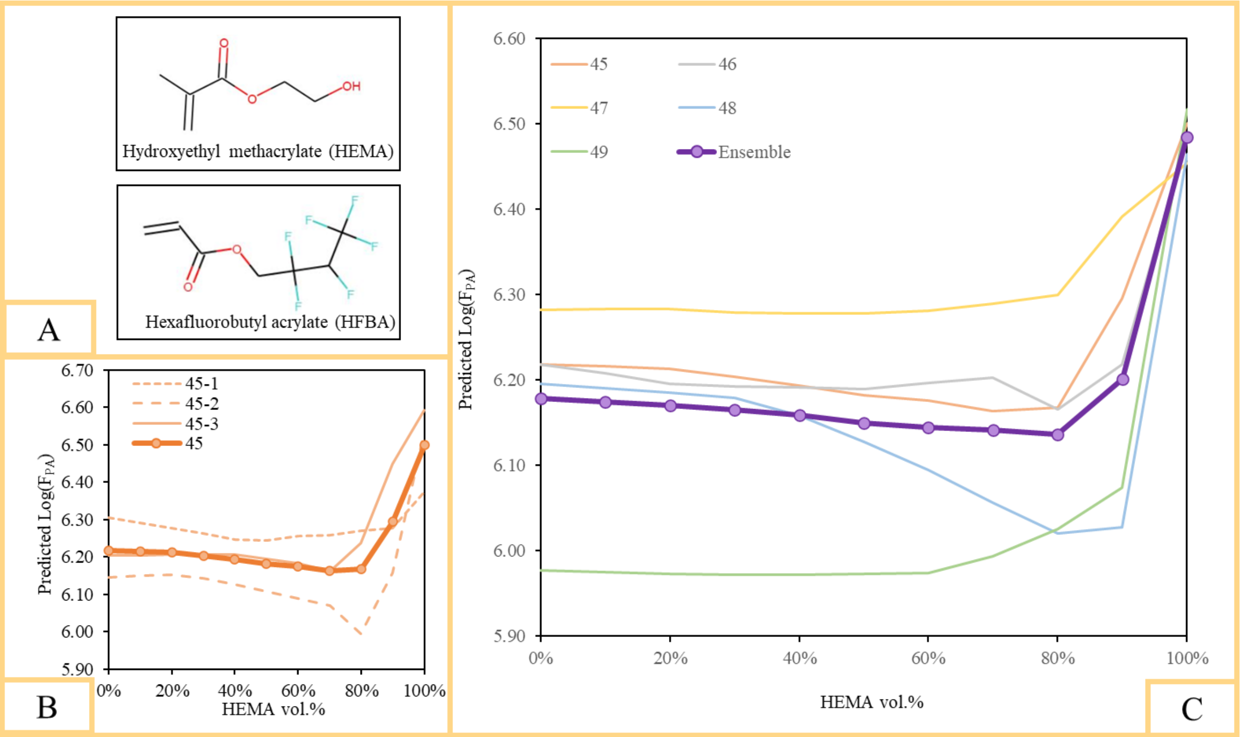

Figure 4. Virtual screening results of the best-performing AP and its antibiofilm performances at various compositions. A. The ensemble model predicted that poly(2-hydroxyethyl methacrylate-*co*-hexafluorobutyl acrylate) (or pHEMA-*co*-HFBA) was a top performer, with an antibiofilm performance among the top 5% out of the 305 AP candidates screened. B. The prediction of antibiofilm performance depends on the autoencoder seeds used for the SVR models. The thick line represents the averaged prediction for the three seeds tested. C. The prediction of antibiofilm performance depends on the number of reduced features. The thick purple line shows the ensemble model prediction, and the colored thin lines represent the model predictions using different numbers of reduced features (averaged over three SVR models). Therefore, the ensemble model improves stability and generalizability of the predictions.

5. External Model Validation at the Solid-Liquid-Air Interface

5.1. iCVD synthesis of AP

The ability to combine the virtual screening of AP using a generalized ML model with the high-fidelity synthesis of AP coatings using iCVD would considerably accelerate the discovery of high-performing AP in various antibiofilm contexts. Further, here we test whether a generalized ML model that performs well at its native solid-liquid interfaces can also provide useful synthesis guides at unseen solid-liquid-air interfaces.

To this end, we synthesized the pHEMA-*co*-HFBA copolymer thin films with vol.% of HEMA ranging from 0% (i.e., pHFBA) to 100% (i.e., pHEMA) using iCVD (see Table S1 for iCVD synthesis conditions). iCVD is an all-dry deposition technique that produces polymer thin films via free radical polymerization[20,24] in which initiator and monomer(s) are introduced simultaneously into a reaction chamber, as depicted in **Figure 5**A. Due to the solubility of pHEMA homopolymer in an aqueous environment, we introduced a small amount (~15% in final composition) of a hydrophilic crosslinker, namely ethylene glycol dimethacrylate (EGDMA), to improve its stability during biofilm formation tests. The choice of EGDMA as the crosslinker was based on its molecular structure, which resembles that of HEMA. The inclusion of EGDMA rendered the copolymers insoluble during prolonged exposure to an aqueous environment (Figure S6). The introduction of EGDMA to hydrophilic polymer coatings has been shown to have minimal effect on the antifouling activity of the coatings in previous studies. The successful synthesis of pHEMA-*co*-HFBA thin films at various compositions and the pHEMA-*co*-EGDMA thin film was confirmed by Fourier transform infrared spectroscopy (FTIR, **Figure S**1) and X-ray Photoelectron Spectroscopy (XPS, **Table S2**).

Complete polymerization for all polymers was confirmed by the absence of the C=C peak at around 1640 cm^{-1}. There was a slight shift in the peak position of C=O, i.e., from 1757 cm^{-1} (for pHFBA) to 1753 cm^{-1} (for pHEMA-*co*-HFBA), confirming successful copolymerization reaction rather than physical mixing of the homopolymers.[37] The composition of pHEMA-*co*-EGDMA was determined using a previously established method.[38] The composition of pHEMA-*co*-HFBA was calculated by performing peak deconvolution on the C=O group: 1753 cm^{-1} for HFBA and 1728 cm^{-1} for HEMA. Note that pHEMA-*co*-HFBA with 76 vol.% HEMA was soluble in water (**Figure S**2) and thus not suitable for the antibiofilm experimentation. To further confirm the chemical composition of pHEMA-*co*-HFBA, XPS survey scans were collected, and the composition was calculated and shown in **Table S2**. The difference between the compositions calculated using XPS and that using FTIR was up to 5%. This discrepancy, though relatively small, might suggest possible differences between the film surface and bulk compositions. We chose to use the FTIR results in the compositional analysis of the iCVD-synthesized AP films because they represent the film bulk composition (as the data were collected based on IR signals that penetrated the entire thickness of the films). Whereas XPS is known to be a surface-sensitive technique, probing the composition within very top layer (~10 nm) of a given sample,[39] and is thus sensitive to surface contamination (e.g., adventitious carbon). Surface chain reorientation (e.g., with hydrophobic moieties concentrated in the top layer) has been reported in AP copolymers,[12] which may lead to an overestimate of the hydrophobic content when XPS results were used (as shown in Table S2). Furthermore, the ML model was trained on microarray polymer libraries within which copolymer compositions were calculated using the volume ratios of the constituent monomer feeds, which were more characteristic of average bulk compositions than surface compositions. For this reason, the bulk compositions derived from FTIR should provide a closer match with the compositional descriptor used in the training data and hence were expected to be more suited for testing the ML model with respect to the iCVD-synthesized AP coatings. To further minimize possible errors due to differences in compositional units, AP compositions were shown in vol.% (rather than mol.%) to be consistent with training (see Experimental Section).

Molecular-scale heterogeneity is one of the key attributes that underpin the antibiofilm performance of AP. To confirm that the pHEMA-*co*-HFBA synthesized using iCVD was indeed free of microphase separation, we performed Fineman-Ross copolymerization analysis (**Figure S3**) to determine reactivity ratios of HEMA (r_{HEMA}) and HFBA (r_{HFBA}), respectively (see the experimental section for details). r_{HEMA} and r_{HFBA} were determined to be 1.05 ($R^2 = 0.97$) and 0.65 ($R^2 = 0.97$), i.e., $r_{\mathrm{HEMA}} > 1$ and $0 < r_{\mathrm{HEMA}} \cdot r_{\mathrm{HFBA}} < 1$, indicating pHEMA-*co*-HFBA exhibited random copolymerization with a small preference for HEMA addition to a growing chain end. Fineman-Ross plots using XPS-derived HEMA compositions yielded poorer linearity (**Figure S4** and **S5**), which was likely caused by XPS being a surface analysis technique. Hence, iCVD enabled the synthesis of random AP, which is not possible in solution-phase free radical polymerization where AP tend to derive from block copolymers with microphase separation.[40]

Lastly, to confirm the stability of pHEMA-*co*-HFBA, the as-synthesized thin films were immersed in deionized (DI) water over a period of 24 hours at room temperature. The film thickness was measured before and after the incubation (**Figure S**6), which showed decreases less than 10% for pHEMA-*co*-HFBA thin films, and less than 17% for pHEMA-*co*-EGDMA, which were within the range anticipated due to chain reorganization and compression. We further corroborated that attribution through the measurement of refractive index (RI) before and after 24 hours of incubation in water (Table S3), as RI is known to reflect the packing density (free volume) of a polymer.[41] After the 24-hour incubation, we observed an increase in the film RI, implying that the films became more compact with less free volume and a greater density. Furthermore, a previous study has shown that with polymer chain dissolution or coating delamination, iCVD films would disappear in 10 hours.[42] As such, we believe the observed reduction in thickness after soaking was likely due to chain reorganization and compression.

5.2. Performance comparison between model prediction and iCVD-enabled AP thin films

The amount of biofilms formed on each iCVD thin film, at solid-liquid-air triple interfaces, was quantified using a crystal violet biofilm assay, following the steps shown in **Figure 5**B. The triple interface was chosen as the focus of our performance testing because biofilms of motile bacteria, including PAO1, form predominantly at the triple interfaces.[12] Despite the prevalence and the importance of biofilms formed at the triple interface in materials-associated nosocomial infections, antibiofilm performance at the triple interface is not well understood. As such, this focus provides important insight into the materials design principles, while allowing us to explore the transferability of the generalized model that was trained using biofilm data collected at the solid-liquid interfaces to predicting performance at the practically important triple interface.

The ensemble model successfully predicted the existence of a minimum in the experimentally obtained biofilm quantities, which resided between 60 and 80 vol.% HEMA (**Figure 5**C). Here, we performed max-min standardization on the $\mathrm{Log}(F_{\mathrm{PA}})$ predicted by the ensemble model (see Eq. (4)) such that the maximum and minimum of the ML predicted means match those of the experimentally acquired means at the triple interface. The experimental means of the amounts of biofilm formed on pHEMA-*co*-HFBA at the triple interface, as indicated by $\mathrm{Log}(A^{*}_{\mathrm{norm}})$, traced the trend predicted by the ML model reasonably well, except for that of the hydrophobic pHFBA homopolymer. Specifically, the ML model predicted the hydrophobic homopolymer pHFBA to be more antibiofilm than the hydrophilic homopolymer pHEMA, whereas the biofilms formed at the triple interface demonstrated no significant difference between pHFBA and pHEMA. That discrepancy could be attributed to the high sensitivity of the performance of hydrophobic pHFBA to the presence of an air phase at the interface of interest. Indeed, we demonstrated experimentally that the biofilms formed on the pHFBA at the liquid-solid interface exhibited a mean value of

Log(A^{*}_{norm}) close to that predicted by the ML model (p = 0.34, **Figure 5**D). That sensitivity of the hydrophobic homopolymer to the presence of an air phase is likely a result of the disparate antifouling mechanisms at play at the different interfaces. For example, air entrapment by a hydrophobic surface has been considered as an effective antifouling mechanism at the liquid-solid interface,[43] which may lose its effectiveness due to the rapid air exchange at the triple contact line. In summary, the ML model trained using the biofilm data at the solid-liquid interface successfully captured the dependence of AP's antibiofilm performance on the polymer composition. Nevertheless, transfer of the ML model to the triple interface should be carried out with caution, especially in the case of hydrophobic homopolymers.

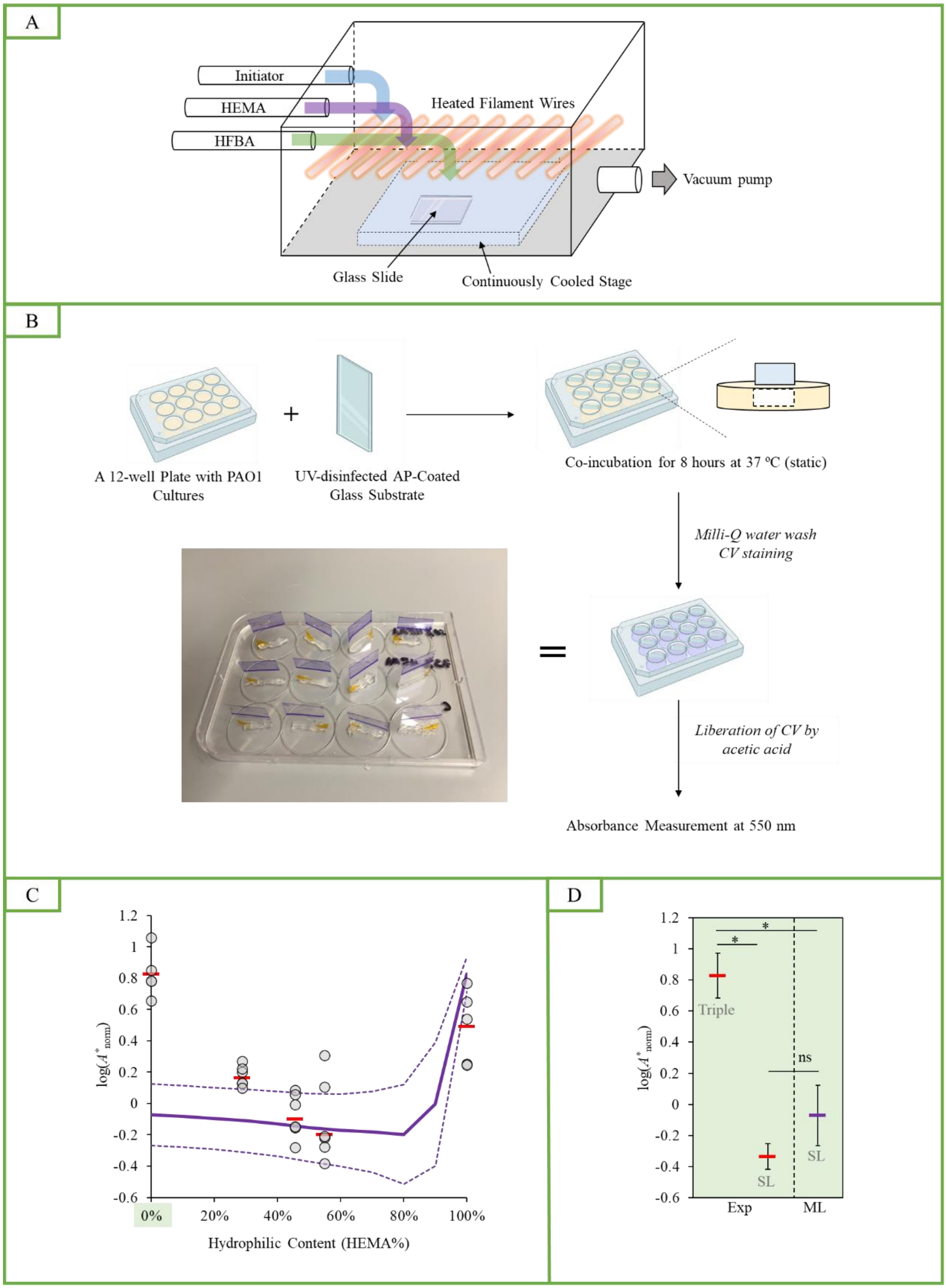
A
Initiator
HEMA
HFBA
Heated Filament Wires
Vacuum pump
Glass Slide
Continuously Cooled Stage
B
A 12-well Plate with PAO1 Cultures
+
UV-disinfected AP-Coated Glass Substrate
Co-incubation for 8 hours at 37 ºC (static)
Milli-Q water wash
CV staining
=
Liberation of CV by acetic acid
Absorbance Measurement at 550 nm
C
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
log(A*norm)
0%
20%
40%
60%
80%
100%
Hydrophilic Content (HEMA%)
D
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
log(A*norm)
*
*
ns
Triple
SL
SL
Exp
ML

Figure 5. Model validation at using iCVD-synthesized AP for their antibiofilm performance at the triple interface. A. A schematic of the iCVD synthesis, during which polymerization occurs on a substrate placed on a continuously cooled stage. B. Quantification of the amount of biofilms

formed at the triple interface on a series of pHEMA-*co*-HFBA with varying compositions. C. Comparison of ML-predicted and experimentally obtained biofilm quantities on the AP series. The solid purple line represents the averaged predictions by the ensemble model trained using data obtained at the solid-liquid (SL) interface and dashed purple lines represent 95% confidence intervals (CI); the red dashes and grey circles represent the experimentally obtained means and individual data points, respectively. D. On the hydrophobic homopolymer pHFBA, although the amounts of biofilms formed at the triple interface (red and "triple") differed from the ML prediction (purple), that formed at the SL interface (red and "SL") was consistent with the ML predictions. Error bars represent 95%CI. *** indicates $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and ns $p > 0.05$.

6. Discussion and Conclusion

In this study, we constructed and applied an ensemble model to achieve high-throughput screening of a large virtual library of AP to search for antibiofilm polymer chemistries in the vast chemical and compositional space. The model provided accurate predictions of antibiofilm performance of AP synthesized using a precision synthesis approach, i.e., iCVD, based on training datasets obtained using a high-throughput approach, i.e., microarrays. It predicted the quantity of solid-liquid biofilms reasonably well (R^2=0.60, RMSE=0.38 $\mathrm{Log}(F_{\mathrm{PA}})$) involving polymers over a vast chemistry-composition design space. Intriguingly, the accuracy of the model improved further (R^2=0.87, RMSE=0.16 $\mathrm{Log}(F_{\mathrm{PA}})$) when tested on the hold-out set comprising only AP. This is quite encouraging because the model successfully extrapolated the behavior of AP from a training set without any examples of amphiphilic copolymers. Moreover, the prediction accuracy AP even surpassed that obtained during training for all copolymers (R^2=0.66, RMSE=0.36 $\mathrm{Log}(F_{\mathrm{PA}})$) by an appreciable margin. One possible explanation might be that polymers in the training set featured disparate antifouling mechanisms (or QSAR), which the model captured with varying degrees of success, and that the QSAR for AP might belong to the category that the model captured to a greater degree.

Furthermore, for biofilms formed at the air-liquid-solid triple interface, the ensemble model successfully predicted how antibiofilm performance varies with the polymer composition, identifying an optimum at the HEMA composition of 60-80 vol%, which was consistent with the experimental results. As such, we demonstrated the ensemble model as a reliable tool for virtual screening of high-performance amphiphilic chemistries and for optimizing their compositional design. The model also points to the possibility of transferring the knowledge obtained from biofilm data obtained under one condition (i.e., liquid-solid interface) to identify best-performing AP under other conditions (e.g., the triple interface), and more broadly, the generalizability of the ensemble model demonstrated here.

That generalizability was enabled by the robust model development framework we established for generating ML models with reasonable bias-variance tradeoff and generalizability. The model was achieved through the implementation of rigorous data-preprocessing steps (e.g., log-transformation, feature rescaling, and dimensionality reduction with autoencoders), of hyperparameter optimization in a five-fold CV, and of an ensemble model. The model demonstrated the ability to captured broad diversity in monomer chemistry (137) and unique molecular fingerprints (423), with a modest abundance of samples during training (2,240). Additionally, its ability to tackle diverse training datasets, pooled from seven independent experiments, greatly alleviated the common challenge of data scarcity for training ML models in antifouling material studies. That flexibility in training dataset sources also unveiled the tantalizing potential of continual expansion and enrichment of the training dataset via extensive inter-laboratory collaborations, akin to the Human Genome Project.

We employed the approach of SVR assisted by autoencoders, as other methods attempted (e.g., neural nets) caused significant bias issues. Despite the challenges around model interpretability, this approach demonstrated distinct advantages in this study, such as a lower time complexity (e.g., compared to feature selection) in handling the size of the datasets in this study, which was a result of the SVR's capability to combine dimensionality reduction and implicit mapping. Furthermore, model interpretation can be achieved by identifying clusters with reduced feature vectors (computed by autoencoders) or other unsupervised mapping strategies to extract the key functional moieties that determine the antibiofilm performance. Compared with the state of the art, [6,22,23], the ensemble model reported here is uniquely able to address a broad polymer design space due to the unprecedentedly diverse training dataset employed in this work. Nevertheless, it could be outperformed in terms of prediction errors by models trained to operate within a confined domain of chemistry. Our future work will focus on improving the computational method for copolymer feature generation as the current approach of linear combination may have limited the SVR mapping ability. We will also expand the design space by incorporating ionizable functional groups, such that the model can predict the performance of a wider variety of AP (e.g., zwitterionic hydrophilic groups). We will improve the accuracy of model predictions by broadly sampling a variety of ML approaches, such as transfer learning while retraining the general model building framework reported here to address the fundamental distinctions among disparate antifouling mechanisms.

7. Experimental Section

Collection of bacterial microarray data

The synthesis procedure and biofilm formation measurement are followed the methods previously described.[6,22] Polymers on microarrays were derived from various monomers mixed pairwise at different compositions and incubated with PAO1 for 72 hours in RPMI-1640 media. Plasmids for constitutively expressing fluorescent proteins GFP (pBK-miniTn7-egfp) and mCherry (pMMR) were introduced into the relevant host strain by conjugation or electroporation to enable bacterial quantification.

Additional details of model development

Upon completing the computational dataset construction for the polymers, we performed data-preprocessing to clean the data and reduce its redundancy. Out of the 2,420 polymers, 180 polymers had a fluorescence intensity below the limit of detection (LOD) (e.g., these fluorescence intensities were labeled as “zero”, which did not reflect the actual anti-biofilm performances) and were thus removed from the dataset to prevent them from potentially misleading ML models, leaving 2,240 polymers in the final dataset. Since the original fluorescence intensity distribution was right-skewed (not showing), a logarithmic transformation was applied to all the remaining fluorescence intensity values (F_{PA}), which were labeled as $\mathrm{Log}(F_{\mathrm{PA}})$ as shown in **Figure 2**A. We further split the dataset into a training set and a holdout test set at a ratio of 80:20 with stratification due to the relatively small size of the dataset and applied a recallable minimum-maximum scaler to the training set.

The autoencoder, with its signature three-layer structure (**Figure 2**B),[44] was effective in distinguishing all polymers in the training space while reducing the noise and redundancy in the dataset. We constructed the autoencoder to reduce the training set’s dimension (labeled as the encoded training set) with proper reference validation assisted by TensorFlow and Keras. Despite the effectiveness of the autoencoder, there remains two sources for instabilities. First, an optimized autoencoder was unlikely to behave consistently once a completely new seed was used due to its nonconvexity (e.g., different local optimums would lead to different reduced feature contents). Additionally, the number of reduced features used could bring about variability in the final model predictions. To account for those instabilities, we developed an ensemble model by averaging the predictions of three autoencoders, which have similar mean squared error but a different number of reduced features. The ensemble method has been shown to simultaneously mitigate variance and bias issues, therefore improving the model's generalizability and performance.[45]

Hyperparameter optimization is also another key to balancing the bias and variance tradeoff. Building upon the strengths of SVR, e.g., its generalizability and low time-complexity by implicit mapping,[34] we developed a radial-basis-function-kernelized SVR model using the encoded training set.

To achieve a reasonable bias-variance trade-off we investigated the model's performance ranging from ten to one hundred reduced features with an increment of ten as shown in **Figure 2**C. Under each reduced dimension, an SVR model's hyperparameters were optimized by the random search algorithm in CV. We also assessed the averaged root mean squared error (RMSE), the averaged coefficient of determination (R^2), and their standard deviations as the evaluation metrics for every model's optimization process.

To the best of our knowledge, there is not a quantitative method widely accepted in the field of ML that allows non-arbitrary selection of the number of the reduced features. Optimization of such selection is of high theoretical and practical importance but is beyond the scope of this work. Therefore, we took the following qualitative approach to define an acceptable range for the number of reduced features. As shown in **Figure 2**C, when the number of reduced features was too small (e.g., 10), the models performed poorly as indicated by the high average RMSE during training (blue curve) and validation (red curve), despite a relatively narrow difference between these two sets (e.g., 0.013 RMSE difference in terms of $\mathrm{Log}(F_{\mathrm{PA}})$ when ten reduced features were applied), which is typical of a bias issue – i.e., not enough fitting dimensions to capture the trends contained in the training (or validation) set. In contrast, when the number of reduced features was too high (e.g., 70 and higher), the model performance started to diverge considerably between the two sets (e.g., 0.035 RMSE difference in terms of $\mathrm{Log}(F_{\mathrm{PA}})$ when 70 reduced features were applied). That gap between validation and training indicated a variance issue, suggesting that the models were not general enough to make accurate predictions for previously unseen polymer chemistries.

The ability to generalize would be essential for the purpose of virtual screening of novel AP for their antibiofilm performance. Therefore, to trade off the bias-variance issue, we identified an acceptable range for the number of reduced features to be between 45 to 49 in this study. To minimize potential variation in model prediction due to the selection of a single reduced feature number, we averaged the predictions generated by models developed using each of the five reduced feature numbers from 45 to 49. In total, this ensemble model consisted 15 individually train SVR models. It should also be noted that the reproducibility of the model can be guaranteed by using the same seeds for the initialization of autoencoder neural networks.

Vapor-deposition of AP

The amphiphilic copolymer thin films were synthesized using an all-dry deposition technique termed *initiated* chemical vapor deposition (iCVD). The iCVD mechanism involves free radical polymerization,[26] in which initiator and monomer(s) are introduced simultaneously into a reaction chamber, as depicted in **Figure 5**A. In this chamber, the heated filament wires are suspended above a continuously cooled substrate (i.e., silicon [Si] wafer and glass coverslips) and promote thermal decomposition of the initiator resulting in free radicals. Then, these radicals chemisorb to adsorbed monomers on the substrate via the Eley-Rideal mechanism[25,46] so that polymerization is initiated, and monomer mono/multilayers are formed.

Before iCVD deposition, the coverslips and Si wafer were cleaned in a plasma cleaner (PDC-001-HP, Harrick Plasma) for 2 minutes at a high radio frequency (RF) setting. No covalent grafting was required to achieve sufficient adhesion of the polymer thin films to the underlying substrates used in this study (Figure S6).[13,47,48] The substrate-independent nature of the iCVD technology and good adhesion of the iCVD coatings on most substrates have been well documented.[29,36,49] That adhesion was enabled by the strong molecular interactions between the iCVD polymer and the underlying substrate, due to the strong molecular adsorption of monomer precursors to the substrate during the iCVD synthesis process (which has been described using the Brunauer–Emmett–Teller theory). Subsequentially, pHEMA-*co*-HFBA copolymer thin films were synthesized using a customer-built iCVD reactor (335 mm diameter, 50 mm height), which was evacuated by an E2M40 rotary vane vacuum pump (Edwards Vacuum, UK). pHEMA-*co*-HFBA amphiphilic copolymer thin films with different compositions were achieved by varying HFBA and HEMA flow rates, respectively. The deposition conditions are summarized in Table S1. The pressure in the reactor chamber was set to 1.1 Torr by a throttling valve (MKS Instruments, USA), and it was maintained by a manometer (Baratron, MKS Instruments, USA). The stage temperature of the chamber was set to 20°C by an Accel 500 LC recirculating chiller (Thermo Fisher, USA), and the filament nickel-chromium wires (80% Ni/20% Cr) were heated to 220-230°C using a DC power source (B&K Precision, USA). A type K thermocouple (Omega Engineering, USA) measured both filament and stage temperatures. *In-situ* laser interferometry (He-Ne laser, JDSU, USA) was located above the reactor glass lid to enable the coating thickness control. During deposition, HFBA was slightly heated to 25ºC. To vaporize HEMA, the monomer jar was heated at 75ºC. All monomers were metered into the reactor using needle valves (Swagelok, USA). The initiator, TBPO was maintained at room temperature, and its flow rate was set using a mass flow controller (MKS Instruments, USA). Argon (Ar) was utilized as a carrier gas to maintain a constant total gas flow rate at four standard cubic centimeters per minute (sccm) and thus constant residence time for gas flow. Lastly, pHEMA homopolymer was deposited with a trace of ethylene glycol dimethacrylate (EGDMA) at a pressure of 0.2 Torr and a stage temperature of 30 ºC. EGDMA was used as a crosslinker to make the PHEMA film insoluble.[38] EGDMA was chosen due to its molecular resemblance to HEMA.

After the vapor deposition, AP coatings were left to equilibrate overnight (~16 hours) under vacuum at 20 °C. This equilibration time was chosen because it allowed ample time for pHEMA to equilibrate (typically over 60-100 min)[50,51] while avoiding the potential aging effect (typically over months-years).[52] This also allowed us to match training conditions, where the ML model was trained using material properties and biofilm data obtained on well-equilibrated but still fresh polymer microarrays, thus precluding potential effects of equilibration and aging on the learned properties-biofilm relationships.

Chemical Characterization.

Fourier-transform infrared spectroscopy (FTIR) (Bruker Vertex V80V Vacuum FTIR system with cooled MCT detector) was used to determine the final composition of copolymer thin films. The spectra were acquired over 400 - 4000 cm^{-1} with a resolution of 4 cm^{-1} and 256 total scans.[53] The spectra were analyzed, and the baseline was corrected using OPUS software (Bruker). The thickness of the thin films on a flat surface (Si wafers) was measured by variable angle spectroscopic ellipsometry (J.A. Woollam Alpha-SE ellipsometer) at three different angles (65º, 70º, and 75º) with a wavelength range from 315 to 718 nm.[54] The optical model with a Cauchy function was used to fit-in experimental data that consisted of 3 different layers: silicon wafer (Si) as a substrate, silicon oxide (IV) (SiO_2) as a calibration standard, and the copolymer thin film.

X-ray photoelectron spectroscopy (XPS) was performed using a Scienta Omicron ESCA 2SR (Uppsala, Sweden) with operating pressure of 1×10^{-9} Torr. X-rays were generated from monochromatic Al Kα at 300W (15 kV; 20 mA) with an analysis spot size of 2 mm in diameter. Survey scans were collected to determine the composition of the copolymer thin films.

Fineman-Ross

To determine the type of copolymerization for pHEMA-*co*-HFBA, Fineman-Ross copolymerization analysis was used to determine reactivities of HEMA and HFBA, respectively.[55]

The surface monomer composition, f_{HEMA}, was calculated as follows:

$$f_{HEMA} = \frac{\frac{P_{HEMA}}{P_{sat,HEMA}}}{\frac{P_{HEMA}}{P_{sat,HEMA}} + \frac{P_{HFBA}}{P_{sat,HFBA}}} \qquad (1)$$

where $\frac{P_{HEMA}}{P_{sat,HEMA}}$ or $\frac{P_{HFBA}}{P_{sat,HFBA}}$ is monomer partial pressure over monomer saturated pressure, which is also representative of the monomer surface concentration of each co-monomer.

The film composition (F_{HEMA}) was obtained from FTIR as mentioned above. To determine, the type of copolymerization (either random, block or alternating), Fineman-Ross equation was used as follows:

$$\frac{\mathrm{f_{HEMA}}(1-2\mathrm{F_{HEMA}})}{\mathrm{F_{HEMA}}(1-\mathrm{f_{HEMA}})} = \mathrm{r_{HFBA}} + \mathrm{r_{HEMA}} \frac{\mathrm{f_{HEMA}^2}(\mathrm{F_{HEMA}}-1)}{\mathrm{F_{HEMA}}(1-\mathrm{f_{HEMA}})^2} \qquad (2)$$

where r_{HEMA} and r_{HFBA} are reactivities of HEMA and HFBA, which can be determined from the slope and intercept, respectively.

Biofilm Assay at Solid-Liquid-Air and Solid-Liquid Interfaces

After confirming the hydrophilic content in each pHEMA-*co*-HFBA with FTIR, we performed a crystal violet biofilm assay to evaluate biofilm formation as a function of volume percentage of HEMA (vol.% HEMA). The reason for using vol.% is to keep consistent with the quantification method used in the training dataset, which approximates polymer compositions on the microarray slide based on the volume percentage of each monomer feed used for synthesis (assuming complete conversion of comonomers). To convert mol.% into vol.%, we adopted molar volumes of 121.63 mL mol^{-1} for HEMA[56] and 185.56 mL mol^{-1} for HFBA[57] (see SI for details). The procedures of the crystal violet biofilm assay are reported elsewhere.[12] Briefly, PAO1 were grown to stationary phase (37 °C, 16 h, 225 rpm) in Luria Bertani (LB, Lennox) broth and subcultured in fresh LB broth (1:100 dilution). For the biofilm assay at solid-liquid-air interfaces, the AP-coated glass substrates with hydrophilic content ranging from 0% to 100%, along with the uncoated controls, were each half-submerged vertically in the diluted LB culture contained the wells of 12-well culture plates. This was followed by static incubation for 8 hours at 37 °C based on P. *aeruginosa* biofilm lifecycle and previous work.[12,58,59] At this point in time (8 hours) when culture transitions from stage II to III, quantification is most reliable due to the presence of linear and vertical biofilm development; further growth would lead to biofilm maturation, giving rise to a three-dimensional community (stage IV, 14-24h) where biomass accumulation would reflect more of multiplication and quorum sensing,[60] rather than bacteria-material interactions. For the biofilm assay at solid-liquid interfaces, the hydrophobic pHFBA-coated substrates (i.e., “0% HEMA”), along with the uncoated controls, were placed horizontally with the coated side facing upward on the well bottom of 6-well culture plates. These substrates were fully submerged in 4 mL of the diluted LB culture and incubated with gentle shaking (60 rpm) for 72 h at 37 °C, with replacement of 3.5 mL of the spent medium in each well with fresh LB every 24 h. The shaking and extended incubation time were implemented to encourage biofilm formation by PAO1 at the solid-liquid interfaces. At the end of the incubation period specified for either type of interfaces, the biofilm-covered substrates were rinsed in Milli-Q water to dislodge loosely attached cells, followed by staining in crystal violet solution (0.5% w/v) for 15 min. After removing unbound dye molecules in Milli-Q water, crystal violet molecules bound to the biofilm were dissolved in 1 mL acetic acid (30% v/v). At least four biological replicates were included for each type of substrates. To correct for the artifact caused by crystal violet bound to polymer coatings, medium controls (LB without bacteria) were included for each substrate type alongside the inoculated samples. We measured the absorbance of the dissolved crystal violet from each sample at 550 nm and obtainedthe normalized absorbance based on the following equation.

$$A^{*}_{\mathrm{norm}} = \frac{A^{*}_{\mathrm{b,P}}}{A^{*}_{\mathrm{b,G}}} = \frac{(A_{\mathrm{t,P}} - A_{0,\mathrm{P}})/\mathrm{OD}_{600,\mathrm{P}}}{(A_{\mathrm{t,G}} - A_{0,\mathrm{G}})/\mathrm{OD}_{600,\mathrm{G}}} \tag{3}$$

Where $A_{\mathrm{t,P}}$ and $A_{\mathrm{t,G}}$ are total absorbance for polymer-coated (“P”) and glass coverslip (“G”), respectively. $A_{0,\mathrm{P}}$ and $A_{0,\mathrm{G}}$ are absorbance measured from bacteria-free controls. Thus, the difference between these two measured absorbances (e.g., $A_{\mathrm{t,P}} - A_{0,\mathrm{P}}$) is the corrected absorbance. This value was then divided by its corresponding OD_{600} at the end of the incubation, which normalizes the biofilm growth by the planktonic cell density to account for variations in growth conditions, resulting in $A^{*}_{\mathrm{b,P}}$ and $A^{*}_{\mathrm{b,G}}$. Lastly, the ratio of $A^{*}_{\mathrm{b,P}}$ over $A^{*}_{\mathrm{b,G}}$ is computed, denoted A^{*}_{norm}, which benchmarks biofilm formed on the coatings of interest against bare glass coverslip. A^{*}_{norm} less than unity suggests the coating outperformed glass in terms of its antibiofilm property, and vice versa.

In order to compare the ML predictions directly with the experimental observations, standardization of the biomass units was performed. The unit of ML-predicted biomass was $\mathrm{Log}F_{\mathrm{PA}}$, which was inherited from the fluorescent microarray experiment. Rescaling of the ML-predicted $\mathrm{Log}F_{\mathrm{PA}}$ values was performed by applying Eq.(4), such that the maxima and minima of the means predicted by ML were set to align with that of the experimentally observed maxima and minima, respectively.

$$\begin{aligned}\mathrm{Std.\,Log}(F_{\mathrm{PA}}) = {} & \frac{\mathrm{Log}(F_{\mathrm{PA}}) - \min[\mathrm{Log}(F_{\mathrm{PA}})]}{\max[\mathrm{Log}(F_{\mathrm{PA}})] - \min[\mathrm{Log}(F_{\mathrm{PA}})]} \\ & \cdot \{\max[\mathrm{Log}(A^{*}_{\mathrm{norm}})] - \min[\mathrm{Log}(A^{*}_{\mathrm{norm}})]\} + \min[\mathrm{Log}(A^{*}_{\mathrm{norm}})]\end{aligned} \tag{4}$$

Where $Biomass$ is the amount of biofilm quantified in either A^{*}_{norm} or F_{PA}; “min[]” and “max[]” are functions that return minima and maxima of the logarithmic mean biomass among all pHEMA-*co*-HFBA compositions.

Statistical Analysis

Analysis of variance (ANOVA) and post hoc TukeyHSD were performed using statistical programing language R (RStudio, Version 1.2.1335) to compare biofilm formation observed during experiments at solid-liquid-gas and solid-liquid interfaces and that predicted by the ML model, at a 95% confidence level.

Supporting Information

Supporting Information is available from Wiley Online Library or from the author.

Acknowledgments

The authors want to thank the Office of Naval Research (ONR) for support through Award MURI N00014-20-1-2418 to R.Y. and through Award USDA NIFA 2021-67034-35040 to Y.C. The authors also want to acknowledge the Samuel C. Fleming Family Graduate Fellowship for the support granted to A.K. Finally, the authors wish to thank Cornell NanoScale Facility (grant number NNCI-2025233), Cornell Center for Materials Research (grant number DMR-1719875), and Cornell Restricted Access Data Center for providing the facilities required for conducting this research. A.L.H kindly acknowledges the University of Nottingham for funding his Nottingham Research Fellowship. Funding from Wellcome Trust (grant numbers 103882 and 103884) is also kindly acknowledged.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] J. W. Costerton, P. S. Stewart, E. P. Greenberg, *Science (1979)* **1999**, *284*, 1318.

[2] A. SMITH, *Adv Drug Deliv Rev* **2005**, *57*, 1539.

[3] H.-C. Flemming, J. Wingender, U. Szewzyk, P. Steinberg, S. A. Rice, S. Kjelleberg, *Nat Rev Microbiol* **2016**, *14*, 563.

[4] M. Tyers, G. D. Wright, *Nat Rev Microbiol* **2019**, *17*, 141.

[5] J.-L. Mainardi, R. Villet, T. D. Bugg, C. Mayer, M. Arthur, *FEMS Microbiol Rev* **2008**, *32*, 386.

[6] A. L. Hook, C. Chang, J. Yang, S. Atkinson, R. Langer, D. G. Anderson, M. C. Davies, P. Williams, M. R. Alexander, *Advanced Materials* **2013**, *25*, 2542.

[7] E. Ostuni, R. G. Chapman, M. N. Liang, G. Meluleni, G. Pier, D. E. Ingber, G. M. Whitesides, *Langmuir* **2001**, *17*, 6336.

[8] R. Yang, H. Jang, R. Stocker, K. K. Gleason, *Advanced Materials* **2014**, *26*, 1711.

[9] T.-S. Wong, S. H. Kang, S. K. Y. Tang, E. J. Smythe, B. D. Hatton, A. Grinthal, J. Aizenberg, *Nature* **2011**, *477*, 443.

[10] C. S. Gudipati, J. A. Finlay, J. A. Callow, M. E. Callow, K. L. Wooley, *Langmuir* **2005**, *21*, 3044.

[11] J.-F. Dubern, A. L. Hook, A. M. Carabelli, C.-Y. Chang, C. A. Lewis-Lloyd, J. C. Luckett, L. Burroughs, A. A. Dundas, D. J. Humes, D. J. Irvine, M. R. Alexander, P. Williams, *Sci Adv* **2023**, *9*, DOI 10.1126/sciadv.add7474.

[12] T. B. Donadt, R. Yang, *Adv Mater Interfaces* **2021**, *8*, 2001791.

[13] A. Khlyustova, M. Kirsch, R. Yang, *ACS Sustain Chem Eng* **2022**, *10*, 15699.

[14] C. A. Amadei, R. Yang, M. Chiesa, K. K. Gleason, S. Santos, *ACS Appl Mater Interfaces* **2014**, *6*, 4705.

[15] R. Yang, E. Goktekin, M. Wang, K. K. Gleason, *J Biomater Sci Polym Ed* **2014**, *25*, 1687.

[16] S. Bhatt, J. Pulpytel, G. Ceccone, P. Lisboa, F. Rossi, V. Kumar, F. Arefi-Khonsari, *Langmuir* **2011**, *27*, 14570.

[17] C. Leng, K. A. Gibney, Y. Liu, G. N. Tew, Z. Chen, *ACS Macro Lett* **2013**, *2*, 1011.

[18] A. v. Dudchenko, P. Bengani-Lutz, A. Asatekin, M. S. Mauter, *ACS Appl Polym Mater* **2020**, *2*, 4709.

[19] S. H. Baxamusa, K. K. Gleason, *Adv Funct Mater* **2009**, *19*, 3489.

[20] Z. Zhao, H. Ni, Z. Han, T. Jiang, Y. Xu, X. Lu, P. Ye, *ACS Appl Mater Interfaces* **2013**, *5*, 7808.

[21] Y. Cheng, R. Yang, *Acc Mater Res* **2021**, *2*, 979.

[22] A. L. Hook, C.-Y. Chang, J. Yang, J. Luckett, A. Cockayne, S. Atkinson, Y. Mei, R. Bayston, D. J. Irvine, R. Langer, D. G. Anderson, P. Williams, M. C. Davies, M. R. Alexander, *Nat Biotechnol* **2012**, *30*, 868.

[23] P. Mikulskis, A. Hook, A. A. Dundas, D. Irvine, O. Sanni, D. Anderson, R. Langer, M. R. Alexander, P. Williams, D. A. Winkler, *ACS Appl Mater Interfaces* **2018**, *10*, 139.

[24] A. Khlyustova, Y. Cheng, R. Yang, *J Mater Chem B* **2020**, *8*, 6588.

[25] K. K. S. Lau, K. K. Gleason, *Macromolecules* **2006**, *39*, 3688.

[26] A. M. Coclite, K. K. Gleason, *Plasma Processes and Polymers* **2012**, *9*, 425.

[27] Y. Cheng, A. Khlyustova, P. Chen, R. Yang, *Macromolecules* **2020**, *53*, 10699.

[28] K. K. Gleason, *Nature Reviews Physics* **2020**, *2*, 347.

[29] T. Franklin, R. Yang, *ACS Biomater Sci Eng* **2020**, *6*, 182.

[30] A. A. Dundas, O. Sanni, J. Dubern, G. Dimitrakis, A. L. Hook, D. J. Irvine, P. Williams, M. R. Alexander, *Advanced Materials* **2019**, *31*, 1903513.

[31] O. Sanni, C. Chang, D. G. Anderson, R. Langer, M. C. Davies, P. M. Williams, P. Williams, M. R. Alexander, A. L. Hook, *Adv Healthc Mater* **2015**, *4*, 695.

[32] K. Adlington, N. T. Nguyen, E. Eaves, J. Yang, C.-Y. Chang, J. Li, A. L. Gower, A. Stimpson, D. G. Anderson, R. Langer, M. C. Davies, A. L. Hook, P. Williams, M. R. Alexander, D. J. Irvine, *Biomacromolecules* **2016**, *17*, 2830.

[33] D. Rogers, M. Hahn, *J Chem Inf Model* **2010**, *50*, 742.

[34] H. B. C. J. C. , K. L. S. A. V. V. Drucker, *Adv Neural Inf Process Syst* **1996**, *9*, 155.

[35] E. Ostuni, R. G. Chapman, R. E. Holmlin, S. Takayama, G. M. Whitesides, *Langmuir* **2001**, *17*, 5605.

[36] K. K. Gleason, *CVD Polymers: Fabrication of Organic Surfaces and Devices*, Wiley-VCH Verlag GmbH & Co., Weinheim, **2015**.

[37] A. Liu, E. Goktekin, K. K. Gleason, *Langmuir* **2014**, *30*, 14189.

[38] W. Reichstein, L. Sommer, S. Veziroglu, S. Sayin, S. Schröder, Y. K. Mishra, E. İ. Saygili, F. Karayürek, Y. Açil, J. Wiltfang, A. Gülses, F. Faupel, O. C. Aktas, *Polymers (Basel)* **2021**, *13*, 186.

[39] J. B. Gilbert, M. F. Rubner, R. E. Cohen, *Proceedings of the National Academy of Sciences* **2013**, *110*, 6651.

[40] A. K. Leonardi, C. K. Ober, *Annu Rev Chem Biomol Eng* **2019**, *10*, 241.

[41] V. Raghunathan, J. Luis Yagüe, J. Xu, J. Michel, K. KGleason, L. C. Kimerling, H. Ma, A. K. Y Jen, L. R. Dalton, J. M. Lee, D. J. Kim, H. Ahn, S. H. Park, G. Kim, G. H. Kim, O. K. Kwon, K. J. Kim, H. Zhang, X. Jian, X. Han, M. Zhao, G. Morthier, R. Baets, W. Groh, A. Zimmermann, *Fabrication and Design Considerations for Multilevel Active Polymeric Devices*, Optical Society Of America, **2002**.

[42] R. Yang, H. Jang, R. Stocker, K. K. Gleason, *Advanced Materials* **2014**, *26*, 1711.

[43] Y. Cheng, G. Feng, C. I. Moraru, *Front Microbiol* **2019**, *10*, DOI 10.3389/fmicb.2019.00191.

[44] Y. Wang, H. Yao, S. Zhao, *Neurocomputing* **2016**, *184*, 232.

[45] C. McGill, M. Forsuelo, Y. Guan, W. H. Green, *J Chem Inf Model* **2021**, *61*, 2594.

[46] A. Khlyustova, R. Yang, *Front Bioeng Biotechnol* **2021**, *9*, DOI 10.3389/fbioe.2021.670541.

[47] W. E. Tenhaeff, K. K. Gleason, *Adv Funct Mater* **2008**, *18*, 979.

[48] A. Khlyustova, M. Kirsch, X. Ma, Y. Cheng, R. Yang, *J Mater Chem B* **2022**, *10*, 2728.

[49] P. Chen, J. Lang, Y. Zhou, A. Khlyustova, Z. Zhang, X. Ma, S. Liu, Y. Cheng, R. Yang, *Sci Adv* **2022**, *8*, DOI 10.1126/sciadv.abl8812.

[50] C. Esen, R. H. Şenay, E. Feyzioğlu, S. Akgöl, *Journal of Nanoparticle Research* **2014**, *16*, 2255.

[51] N. L. Smith, Z. Hong, S. A. Asher, *Analyst* **2014**, *139*, 6379.

[52] T. V. Nguyen, X. H. Le, P. H. Dao, C. Decker, P. Nguyen-Tri, *Prog Org Coat* **2018**, *124*, 137.

[53] G. Ozaydin-Ince, J. M. Dubach, K. K. Gleason, H. A. Clark, *Proceedings of the National Academy of Sciences* **2011**, *108*, 2656.

[54] C. D. Petruczok, R. Yang, K. K. Gleason, *Macromolecules* **2013**, *46*, 1832.

[55] G. Odian, *Principles of Polymerization*, Wiley, **2004**.

[56] National Center for Biotechnology Information (2023), “PubChem Compound Summary for CID 13360, 2-Hydroxyethyl methacrylate,” can be found under https://pubchem.ncbi.nlm.nih.gov/compound/2-Hydroxyethyl-methacrylate, **2023**.

[57] Sigma-Aldrich, “Sigma-Aldrich Compound Summary for 2,2,3,4,4,4-Hexafluorobutyl methacrylate,” can be found under https://www.sigmaaldrich.com/US/en/product/aldrich/371971, **2023**.

[58] T. Rasamiravaka, Q. Labtani, P. Duez, M. el Jaziri, *Biomed Res Int* **2015**, *2015*, 1.

[59] C. S. L. Spake, E. M. Berns, L. Sahakian, A. Turcu, A. Clayton, J. Glasser, C. Barrett, D. Barber, V. Antoci, C. T. Born, D. R. Garcia, *Journal of Orthopaedic Research* **2022**, *40*, 2448.

[60] C. Solano, M. Echeverz, I. Lasa, *Curr Opin Microbiol* **2014**, *18*, 96.