ACS APPLIED MATERIALS & INTERFACES

www.acsami.org Research Article

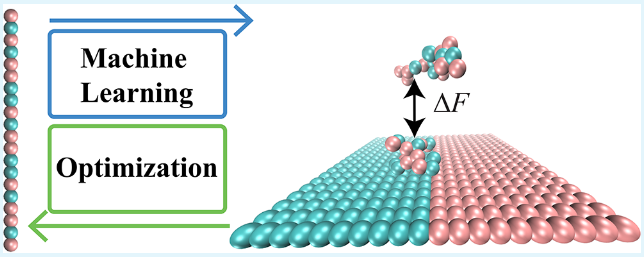
Predicting Adhesive Free Energies of Polymer−Surface Interactions with Machine Learning

Jiale Shi, Michael J. Quevillon, Pedro H. Amorim Valença, and Jonathan K. Whitmer*

Cite This: https://doi.org/10.1021/acsami.2c08891 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Polymer−surface interactions are crucial to many biological processes and industrial applications. Here we propose a machine learning method to connect a model polymer's sequence with its adhesion to decorated surfaces. We simulate the adhesive free energies of 20000 unique coarse-grained one-dimensional polymer sequences interacting with functionalized surfaces and build support vector regression models that demonstrate inexpensive and reliable prediction of the adhesive free energy as a function of sequence. Our work highlights the promising integration of coarse-grained simulation with data-driven machine learning methods for the design of functional polymers and represents an important step toward linking polymer compositions with polymer−surface interactions.



KEYWORDS: *polymer sequence, polymer−surface interaction, polymer adsorption, inverse design, machine learning, free energy calculation, molecular dynamics simulation, genetic algorithm*

INTRODUCTION

Polymer−surface interactions are critical to many industrial applications and biological processes.[1−4] Writing or painting on paper with macromolecular pigments and inks provides a ubiquitous example of polymer−surface interactions.[2] Many industrial products, such as coatings used on the surfaces of magnetic storage media, silicon capacitors, and computer hardware, exploit polymer−surface interactions. These interactions are also integral to novel processes. For example, Kim et al.[1] utilize the interactions of block copolymers with chemically patterned surfaces to induce epitaxial self-assembly of block polymer domains, allowing for molecular-level control in top-down fabrication techniques. Additionally, many biological processes also involve what in essence are polymer−surface interactions. The interactions between heterochromatin and nuclear lamina affect the cell nuclear reorganization, reflecting cellular senescence processes.[5] Vital biological processes such as intracellular signaling and incorporation of viruses into host cells[4,6] are initiated by a protein searching for and recognizing a specific receptor on a cell surface. A small change in the sequence of the polymers affects their interactions and adhesive properties. For instance, mutations in the virus, like D164G, N501Y, and 501.V2 of COVID-19,[7−9] change the spike protein structures and functionalities enabling the new spike protein to bind more easily with the ACE2 receptor, leading to a higher likelihood of infections. On a practical level, an understanding of the quantitative effects polymer sequences can have on surface adhesion is an essential ingredient for the design and synthesis of adhesive materials for tissues,[10] where the surfaces to be attached may have significant compositional heterogeneity.

Several early theoretical and computational studies have examined the effects of polymer composition on polymer−surface interactions and molecular recognition.[11−14] Chakraborty and Bratko[12] utilized Monte Carlo simulation to study the adsorption of random heteropolymers (RHPs) on disordered multifunctional surfaces, finding that a sharp adsorption transition occurs when statistical pattern matching exists between the RHP sequence and the surface site distribution. Muthukumar[13,14] performed studies utilizing both theoretical analysis and Monte Carlo simulations to study the interactions of a polyelectrolyte chain with a patterned surface of opposite charge, illustrating that the self-assembly of polymer molecules at patterned surfaces is largely affected by the charge density, the size of the pattern, and the Debye length. It is also known that polymer structural properties, such as the radius of gyration, can also influence polymer−surface interactions.[15] A coarse-grained statistical mechanical study of AB copolymers interacting with stripes of A and B beads on a surface was performed by Kriksin et al.,[16] who found that the adsorption behavior strongly depends on the copolymer sequence distribution and the arrangement of

Received: May 23, 2022
Accepted: July 18, 2022

selectively adsorbing regions on the substrate.[16] While some of these studies emphasize the importance of polymer sequence in surface adhesion,[12] both a qualitative understanding and a quantitative understanding of sequence design principles are lacking.[12] The formidable challenge here is that databases for polymer sequences and structures comparable to extant databases for comparable properties[17] do not exist and are expensive to create.

Machine learning (ML) and artificial intelligence (AI) have emerged as powerful tools for physical science and engineering,[18−21] highlighted by recent projects AlphaFold2[22,23] and RoseTTAFold.[24] Naturally, this has opened the door to many investigations exploring the predictions of polymer structure and therefore function from the sequence information, as well as "inverse design" research.[17,25−28] For instance, Statt et al.[29−31] have investigated the sequence-dependent aggregation behavior of sequence-defined macromolecules via the unsupervised learning method. Another interesting case is that Meenakshisundaram and co-workers[32] have designed sequence-specific copolymer compatibilizers using a genetic algorithm applied to a coarse-grained molecular dynamics model. In one important related example, Webb et al.[17] utilized a deep neural network (DNN) to predict the structural properties of sequence-controlled coarse-grained polymers just from the sequence information. These successful cases[17,29,30,32] inspire us to utilize ML and AI to investigate the quantitative relationships between the adhesive free energies and the polymer sequence information.

In this work, we utilize biased molecular dynamics simulations to generate a database of free energies that connect the polymer sequence and composition of a patterned surface to its adhesive properties. From this database, we build an inexpensive surrogate model using support vector regression (SVR), which demonstrates reliable prediction of the adhesive free energy of the polymer−surface interaction as a function of the provided polymer sequence information. Subsequently, we apply this model to design targeted sequences using a genetic algorithm. Finally, we illustrate how the polymer sequence can be manipulated to affect the adhesive free energy with the surface and how to do inverse engineering of the polymer sequence using the genetic algorithm.

METHODS

Molecular Dynamics (MD) Simulation. We utilize model polymer chains containing 20 backbone beads based on the classic model of Kremer and Grest,[33] which has been widely utilize to investigate polymer interfacial properties.[32,34−36] The pair interaction between beads is described via a 12-6 Lennard-Jones (LJ) potential:

$$E_{\mathrm{LJ}}^{ij} = 4\epsilon_{ij}\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{1}$$

where ϵ_{ij} sets the interaction energy between two types of beads (red beads are A, and green beads are B; $\epsilon_{AA} = \epsilon_{BB} = 1$, and $\epsilon_{AB} = 0.3$), σ sets the range of the interaction, and r is the distance between two beads in dimensionless LJ units. The AA and BB Lennard-Jones interactions are truncated at a distance of 2.5σ, while the AB Lennard-Jones interaction is truncated at a distance of $2^{1/6}\sigma$ so that it is purely repulsive. While simple, the construction of the model imposes an asymmetry in adhesion that can be optimized by searching over the sequence space of the polymers. Finally, bonds are handled via the finitely extensive nonlinear elastic (FENE) potential.[37−40] An N of 20 is chosen in this work as a chain length that balances the complexity of configurational space (2^{20} states are available) with the anticipated computational cost of the study; 20 polymers evolve sufficiently fast in molecular dynamics simulation to enable extensive (if not comprehensive) data gathering and exploration of the trade-off between the extent of simulation and accuracy in our surrogate model.

$$E_{\mathrm{bond}} = -\frac{1}{2}K{R_0}^{2}\ln\left[1 - \left(\frac{r}{R_0}\right)^{2}\right] \tag{2}$$

where K (=30ϵ/σ^2) is the spring constant and R_0 (=1.5σ) is the maximum extent.

The patterned surface (PS) is 20σ × 20σ, containing 400 beads of type A or B, which have the same 12-6 LJ potential setting as the respective polymer beads. The PS's position is fixed in the z = 0 plane, and the distance between each bead is σ, shown in Figure 1a. As

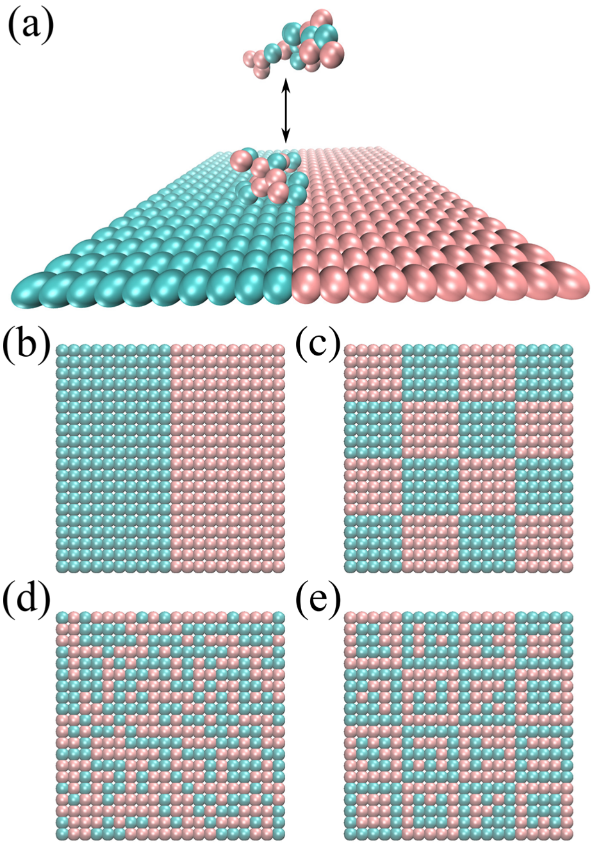


Figure 1. (a) Schematic representation of our simulations involving a polymer chain of defined sequence interacting with a patterned surface with a cubic simulation box whose side length a = 20σ. The polymer chain and surface are both composed of two types of beads, A (red) and B (green). The polymer is modeled as a flexible 20-bead linear chain. The surface is holonomically constrained with dimensions of 20σ × 20σ arranged in a simple square lattice with 400 beads. The four surfaces we examine (b−d) have different patterns with the composition divided approximately equally between beads A and B. (b) PS1, which is composed of half A beads and half B beads in two stripes. N_A = 200, and N_B = 200. (c) PS2, which is composed of 16 alternate small size squares (5σ × 5σ) of A and B beads. N_A = 200, and N_B = 200. (d) PS3 is randomly generated with a probability of $^1/_2$ for each site to be A or B. While even the composition is the most likely state, it is not guaranteed, and there is a significant probability for other configurations to be generated. The surface used as PS3 had an N_A of 184 and an N_B of 216. PS3 is one specific randomized surface pattern. (e) PS4, which is built upon PS2, but randomized within the interior of the 5σ × 5σ squares. The surface used as PS4 had an N_A of 206 and an N_B of 194. PS4 is also one specific randomized surface pattern in a slightly more restricted way.

shown in panels b−e of Figure 1, we investigate four different PSs to validate that our method is robust for surfaces with different patterns. Detailed composition information (N_A and N_B) for the four PSs is presented in Table 1.

Table 1. Compositions for Patterned Surfaces

surface	N_A	N_B
PS1	200	200
PS2	200	200
PS3	184	216
PS4	206	194

Enhanced Sampling. Elucidating the adhesive free energies of sequence-defined polymers with patterned surfaces requires efficient sampling of the rare events comprising removal and re-adhesion of a polymer to the interface, because significant energy barriers must be scaled to enable these rearrangements. Enhanced sampling calculations proceed by applying a bias to collective variables to accelerate the exploration of the simulation systems. Collective variables (CVs), closely related to the concept of reaction coordinates, are a low-dimensional projection of the high-dimensional space of MD simulations, which can clearly distinguish reactants from products and quantify dynamical progress along the pathway from reactants to products.[41] Generally, this defines a vector-valued function from the space of nuclear positions to the reduced CV space, ξ: $\mathbb{R}^{3N} \rightarrow \mathbb{R}^n$, where N is the number of atoms and n the desired reduced dimensionality. For studying the adsorption process on a surface, it is typically sufficient to define a single collective variable. Here, we opt for a single CV (d), the distance between the polymer chain's center of mass (z_{CM}), and the surface ($z_{\mathrm{surface}} = 0$)

$$d = z_{\mathrm{CM}} - z_{\mathrm{surface}} = z_{\mathrm{CM}} \tag{3}$$

Here $d \equiv z_{\mathrm{CM}}$ because the patterned surface is located at $z = 0$; we thus use z_{CM} as the CV for later descriptions. We obtain potentials of mean force (PMFs, here identical to the free energy landscape of a single polymer interacting with the surface) for this coordinate using the multiwalker adaptive bias force (ABF) algorithm[42] as implemented in SSAGES.[43] We choose to sample z_{CM} in the range of $[0.8\sigma, 9.8\sigma]$ with 90 bins. We use four walkers starting from different initial configurations, running each for a total of 2×10^7 MD time steps with a length Δt of 0.001τ ($\tau = \sqrt{m\sigma^2/\epsilon}$ defining the standard time scale incorporating the Lennard-Jones ϵ and the bead mass m) to obtain a converged result. Example PMFs (F vs z_{CM}) are plotted in Figure 2. Generally, as z_{CM} increases from 0.8σ, F first decreases to a minimum value $F_{\min}$ because of the repulsive force between the polymer and the surface. Subsequently, F then increases because of the attractive force between the polymer and the surface. Finally, when $z_{\mathrm{CM}} \gtrsim 7$, F becomes flat at the plateau free energy in the non-interacting state F_o as there is no interaction between the polymer and the surface. We define adhesive energy ΔF for each sequence-defined polymer chain interacting with the PS as the difference between the plateaued non-interacting state F_o and the minimum state $F_{\min}$ ($\Delta F = F_o - F_{\min}$)[15] and use this quantity to train the machine learning models explored in this paper. We note from Figure 2 that PMF landscapes have similar shapes but vary in magnitude for different polymer sequences. While each polymer−surface interaction can potentially have more subtle features, ΔF captures the essential adhesive property.

Machine Learning. We use support vector regression (SVR)[44−46] to build a model predicting polymer−surface interactions ΔF from limited polymer sequence information. The basic idea of a support vector machine (SVM)[45−48] is first to map the data into a high-dimensional space and then construct an optimal separating hyperplane in this space. The SVM thus constructed is then used to perform SVR. We utilize the SVR implementation in the open-source python package Scikit-learn[49] using radial basis functions for the regression. The settings of the optimized values of the regularization parameter (C) of the error term, the maximum error (ϵ) that specifies the penalty-free area, and the kernel coefficient (γ) are stored in the Github repository mentioned in the notes section. All of the parameters mentioned above are optimized using five-fold cross-validation.

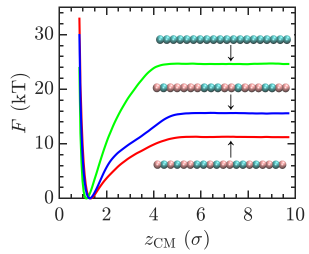


Figure 2. Example potentials of mean force (PMFs) F (in k_BT) plotted as a function of z_{CM} (in σ) (0.8σ, 9.8σ) for polymers with different sequences interacting with PS1. Generally, as z_{CM} increases from 0.8σ, F first decreases to a minimum value $F_{\min}$ because of the repulsive force between the polymer and the surface. Then F increases because of the attractive force between the polymer and the surface; finally when $z_{\mathrm{CM}} \gtrsim 7$, F becomes flat at the value of F_o as there is no interaction between the polymer and the surface. The adhesive energy for each sequence-defined polymer chain interacting with the surface is $\Delta F = F_o - F_{\min}$. The influence of sequence on these quantities is illustrated by the insets, which show the sequence of the polymer interacting with PS1 to obtain the given curve.

There are different types of polymer representations, like one-hot encoding,[17,26] molecular embedding,[26,50] molecular graph,[26,51] and BIGSMILES.[52] Because our coarse-grained model contains only two types of beads, it is both pragmatic and appropriate to use one-hot encoding[17] to preprocess the polymer sequence information. As illustrated in Figure 3a, we encode the one-dimensional 20-bead-length polymer chain's sequence information into a 20-dimensional vector, where type A beads are 1 and type B beads are 0. The resulting vector is the input for the SVR model and is trained to reproduce the corresponding ΔF of each polymer chain that is obtained from the aforementioned biased MD simulations. The polymer sequence is treated as headless; rather than inserting this symmetry into the model, we augment the data set with this symmetric property, adding each sequence's backward representation with the same ΔF output unless that sequence is a palindrome.[a]

For each PS in panels b−e of Figure 1, we collect 20000 polymer chains with unique sequences that have different compositions and different orders and the corresponding ΔF. A separate SVR model is trained for each patterned surface. To train the SVR model, we use five-fold cross-validation after shuffling the data, as shown in Figure 3b. We employ the coefficient of determination (R^2) and mean absolute error (MAE) to characterize the model's performance and optimize the SVR model's hyperparameters [C, γ, and ϵ (see the caption of Figure 3b)].

RESULTS AND DISCUSSION

We begin by exploring the performance of each surrogate model on the accumulated data sets for adhesive free energy. In Figures 4−7, we characterize the distribution of free energies within each data set in panel a, how diverse each free energy distribution is with respect to the average composition (quantified by the fraction x_A of A-type monomers) in panel b, and the errors in the training and test data in panels c and d. Examining the data for PS1 (Figure 4), we note that the distribution of adhesive free energies over the sequences space is quite broad, with the best binding occurring for nearly pure sequences ($x_A \approx 0$, or $x_A \approx 1$). Because the surface is symmetric in its placement of A- and B-type beads, the distribution of binding energies is symmetric over composi-

https://doi.org/10.1021/acsami.2c08891

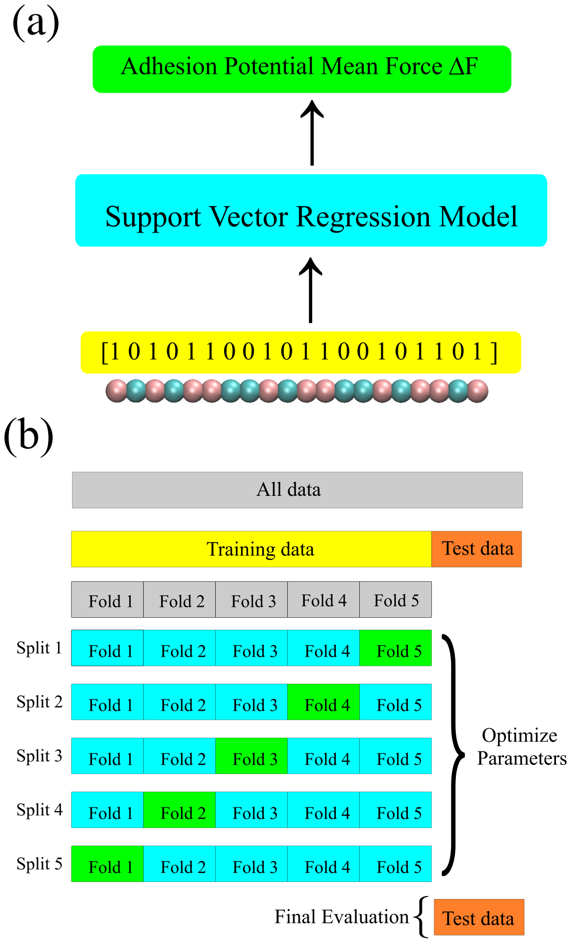


Figure 3. (a) Schematic of our machine learning framework for predicting polymer−surface interactions from sequence information. We use one-hot encoding to transfer the 20-bead-length one-dimensional polymer sequence into a 20-dimensional vector, where type A beads are 1 and type B beads are 0. The input is a 20-dimensional vector, while the output is the corresponding adhesive free energy ΔF. The support vector regression ML models used in this work are dependent on the surface patterns. We investigate four different patterned surfaces in this work and train each corresponding individual support vector regression machine learning model. (b) We separate the 20000 unique polymer sequences into train data (80%, 16000) and test data (20%, 4000). Inside the training data, to avoid overfitting, we use five-fold cross-validation to optimize the SVR model's parameters (regularization parameter C, maximum error ϵ, and kernel coefficient γ).[45,47,48] Next, we train on the whole training data with the optimized SVR model. Finally, we evaluate the model's performance on the remaining test data that have not been used in the five-fold cross-validation.

tions x_A. As shown in panels c and d of Figure 4, training data are clustered quite tightly around the SVR model, and all predicted free energies within the test set are within $\approx 1 k_BT$ of their actual values. The SVR model is thus seen to give good accuracy and predictive capability.

Similar results are obtained for the surrogate model developed for surface PS2. This surface has more fine-grained structure, which results in lower average binding energies and a more unimodal distribution in interaction energies relative to PS1 (Figure 5a). The distribution with respect to x_A is similarly more diffuse (Figure 5b). Interestingly, the predictions for this surface are better than PS1 when the MAE is taken into account, and the test data are more tightly distributed,

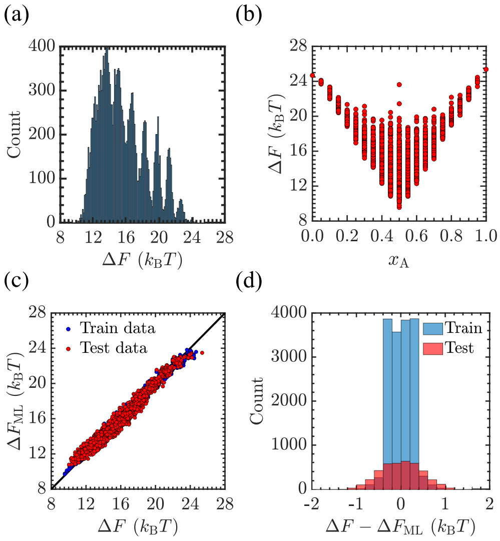


Figure 4. Adhesive free energy data for the interaction between sequence-specified polymers and surface PS1. (a) Histogram of ΔF illustrating the distribution of adhesive free energies of polymer chains. (b) Distribution of ΔF with respect to the overall composition fraction x_A of the polymer. (c) Training behavior and predictive performance of the SVR model, with the predicted value ΔF_{ML} (y-axis) plotted vs the simulated value ΔF (x-axis) for the training data (blue) and test data (red). For the testing set, the R^2 score is 0.973 and the MAE is $0.382 k_BT$. (d) Histogram of deviation $\Delta F - \Delta F_{ML}$ of the model from the true value, demonstrating the good predictive capability of our SVR model for this surface.

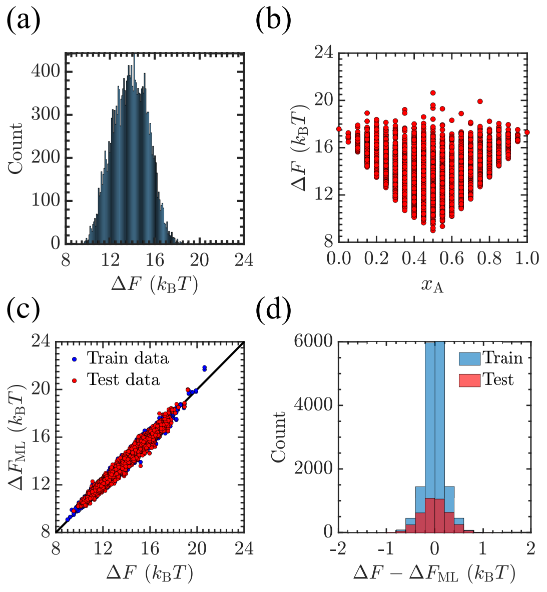


Figure 5. Adhesive free energy data for the interaction between sequence-specific polymers and surface PS2. (a) Histogram of ΔF illustrating the distribution of adhesive free energies of polymer chains. (b) Distribution of ΔF with respect to the overall composition fraction x_A of the polymer. (c) Training accuracy and predictive capability of the SVR model, with ΔF_{ML} (y-axis) plotted vs the simulated value ΔF (x-axis) for the training data (blue) and test data (red). For the test set, the R^2 score is 0.965 and the MAE is $0.230 k_BT$. (d) Histogram of deviation $\Delta F - \Delta F_{ML}$ of the model from the true value, demonstrating good accuracy for the behavior of SVR modeling with this surface.

https://doi.org/10.1021/acsami.2c08891

indicating the SVR again has good predictive capability (see Figure 5a,b and its caption).

PS3 is an exemplar surface containing a specific randomized surface pattern. This is used rather than an average over multiple surfaces in the context of this study so that errors in training and test performance resulting from the lack of an ordered structure can be decoupled from errors due to fluctuations in the surface used to generate input data. Modeling results for this surface are shown in Figure 6. The distribution is seen to have a narrower distribution and smaller mean in the energy distribution (Figure 6a) than PS1 and PS2, resulting from the randomized features. In addition, the randomization results in a skewed distribution of adhesive properties as a function of composition (Figure 6b). Despite this, the SVR model again performs extremely well, with $O(k_BT)$ accuracy in prediction, despite the relatively broad training set distribution (when compared to PS1 and PS2) (see Figure 6c,d). Similar effects are observed for surface PS4 (Figure 7), where the SVR model gives high-accuracy predictions in absolute terms and when the error relative to the mean is considered.

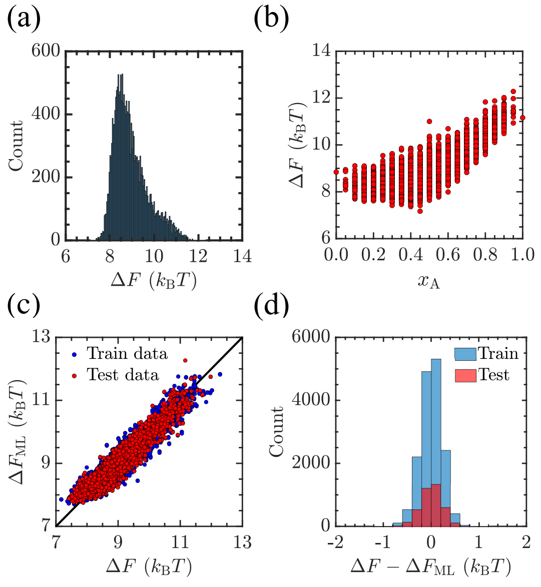


Figure 6. Adhesive free energy data for the interaction between sequence-specific polymers and surface PS3. (a) Histogram of ΔF illustrating the distribution of the adhesive free energies of polymer chains. (b) Distribution of ΔF with respect to the overall composition fraction x_A of the polymer. As this surface contains randomized elements, the distribution of adhesive energies is no longer symmetric. (c) Training and predictive performance of the SVR model, with the predicted value ΔF_{ML} (*y*-axis) plotted vs the value ΔF (*x*-axis) obtained from simulation for the training (blue) and test (red) sets. For the test set, the R^2 score is 0.909 and the MAE is $0.180 k_BT$. (d) Histogram of deviation $\Delta F - \Delta F_{ML}$ of the model from the true value, which demonstrates the good predictive capability of our SVR model for this surface, despite the broader distribution relative to the mean value of interaction energies with PS3 relative to PS2.

From Figures 4a–7a, we see that variations in surface patterns alter the distributions of ΔF significantly. These differences from the training data would be expected to affect the SVR models' prediction abilities. Though in some cases the R^2 is reduced, this primarily results from an overall broadening of the distribution relative to the value of the mean of the distribution. We see that even for the poorest R^2 performance (PS4), the held-out test data still exhibit a high value of 0.86922. Therefore, in general, we see that the SVR models have extremely good predictive capability.

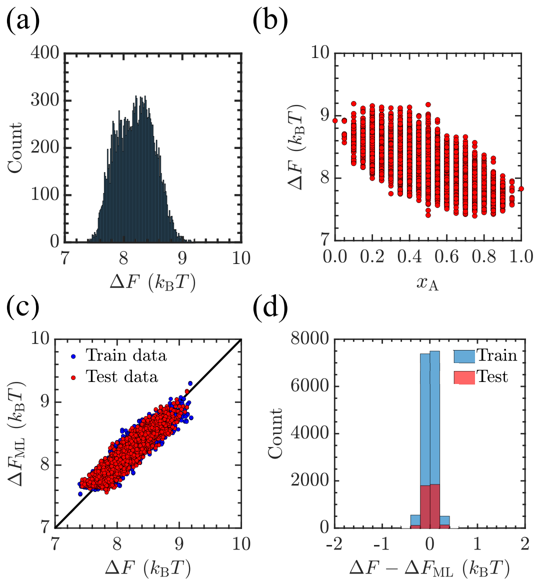


Figure 7. Adhesive free energy data for the interaction between sequence-specific polymers and surface PS4. (a) Histogram of ΔF illustrating the distribution of adhesive free energies of polymer sequences. (b) Distribution of ΔF with respect to the overall composition fraction x_A of the polymer. Because this surface contains randomized elements, the distribution of adhesive energies is not symmetric. (c) Training and prediction performance of SVR models, with the predicted value ΔF_{ML} (*y*-axis) plotted vs the simulated value ΔF (*x*-axis) for the training data (blue) and test data (red). For the test set, the R^2 score is 0.869 and the MAE is $0.090 k_BT$. (d) Histogram of deviation $\Delta F - \Delta F_{ML}$ of the model from the true value, which demonstrates the predictive capability of SVR models for this surface, with a narrower distribution of differences between predictions and data on the test set than other surfaces examined in this work.

In the investigations described above, we utilize 20000 unique data points for each case. Typically, larger data sets will lead to better predictive accuracy for the model. However, it is difficult to obtain large data sets in many materials sciene use-cases due to economic and time constraints involved in building a database from experimental and/or computational data. Therefore, it is important to quantify the amount of data necessary to train an effective SVR model. Again, we use a five-fold cross-validation strategy to choose the training set size by randomly selecting 5–100% of the data from the four remaining folds. Figure 8 illustrates the performance of SVR models, as quantified by the R^2 score (blue circles, left axis) and MAEs (red diamonds, right axis), for predicting ΔF for a set of 4000 held-out sequences as a function of training set size (800–16000). The quality of the SVR models (as measured by R^2) for all four surfaces improves monotonically as the training set size increases, with the sharpest increases coming prior to 8000 training data points; after this, the R^2 levels off considerably. The mean absolute error continues to decrease after this, though the decrease is more pronounced for PS1 and PS2 than for PS3 and PS4. This implies that our model could have performed approximately as well with half of the data utilized here. The fact that only a few thousand data points are necessary for predictive models of this quality is promising, as

https://doi.org/10.1021/acsami.2c08891

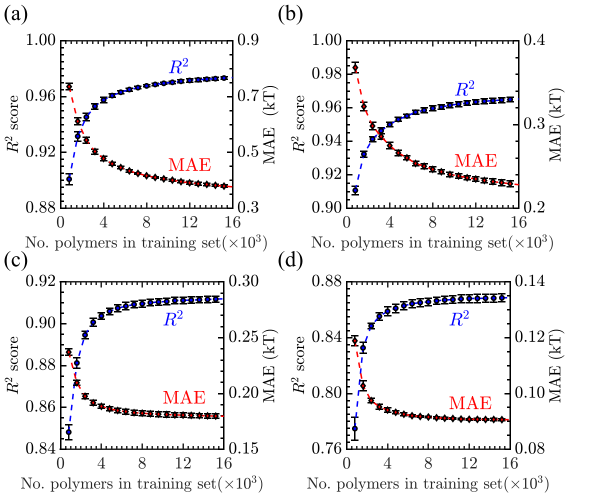


Figure 8. R^2 scores (blue circles, left axis) and MAEs (red diamonds, right axis) associated with ML regression models when predicting ΔF as a function of the number of polymers in a training set for (a) PS1, (b) PS2, (c) PS3, and (d) PS4. The error bars reflect the standard deviation (SD) of the R^2 scores and the SD of the MAEs from five-fold cross-validation, in which each fold is used as a test set for SVR models trained using 5−100% of the data from the remaining four folds.

it enables us to target the accuracy level and optimal size of the adhesion simulations used to populate our database. This benefits from the limited complexity of the input vector space and the simplicity of our optimization question (ΔF for a single surface). However, with appropriate foresight in database construction, such results should be generalizable, highlighting a potential benefit of using ML, because the SVR machine learning models might be constructed from more limited data if the tolerances seen here are acceptable. The need for relatively few data points to determine the essential character of polymer−interface interactions, along with the physical relations of each of the patterned surfaces to each other, indicates that transfer learning[25] could be a viable strategy for improving the model's accuracy in small data sets.

With models in hand, we demonstrate how their efficiency may be exploited to perform inverse design of polymer sequences with a desired ΔF. Here, we apply a genetic algorithm to perform the optimization. We use an initial population size of polymer sequences of 1000, a mating probability of 0.8, and a mutation probability of 0.003. The algorithm proceeds for 1000 steps. To search for the polymer sequences with the largest ΔF_{ML}, we set the fitness function so that a larger ΔF_{ML} has a larger fitness:

$$\mathrm{fitness}_i = \Delta F^i_{\mathrm{ML}} - \min(\Delta F_{\mathrm{ML}}) + 0.001 \tag{4}$$

where i is the index of the sequence, ΔF^i_{ML} is the corresponding ML-predicted adhesive free energy, fitness$_i$ is the corresponding fitness value, and $\min(\Delta F_{\mathrm{ML}})$ is the smallest ML-predicted adhesive free energy among 1000 sequences in the current generation. The small numerical offset is applied so that the fitness function is always strictly greater than zero. Though the algorithm is run for 1000 generations, in practice it converges much more quickly (≈50 generations), so that after this point, it is probing repeats within this model. We have verified by testing with different initial populations that changing the initial population does not affect the convergence rate significantly. The results are summarized in Table 2. The target sequences for PS1−PS3 are already in the existing database of 20000 sequences, while the target sequence for PS4 lies outside the analogous set. We find that the genetic algorithm provides excellent performance in determining optimized sequences, and use of the SVR model enables efficient searching of the compositional space. Though the search is much more broad, our results have clearly selected for more block-like polymers that balance local enthalpy with global entropy. Using the output string, we calculate the corresponding ΔF using MD simulations. Relative to the database, these are plotted using the red diamond in Figure 9.

Table 2. Target Sequences Obtained by Maximizing ΔF_{ML} for the SVR Models and the Corresponding ΔF for Each Surface Examined

surface	target sequence	ΔF_{ML} ($k_{\mathrm{B}}T$)	ΔF ($k_{\mathrm{B}}T$)
PS1	[00000000000000000000]	23.74	24.67
PS2	[00000000001111111111]	21.68	20.64
PS3	[11111111111111111111]	12.27	11.16
PS4	[00000000000011111111]*	9.36	9.43

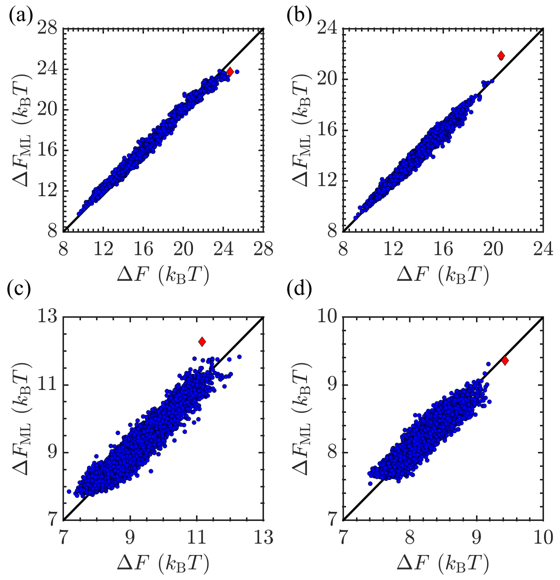


Figure 9. Largest ΔF_{ML} from application of a genetic algorithm (red diamonds; see the text for the description) vs the corresponding simulated ΔF depicted against all sequences (blue circles) in the existing database for (a) PS1, (b) PS2, (c) PS3, and (d) PS4.

Though some small differences exist between predicted and actual free energies, these sequences are uniformly at the top end of the distribution in terms of maximal adhesion. One may improve these results, if desired, by first using a genetic model to obtain the top 100 sequences and then running MD simulations to obtain the true ΔF for these top 100 sequences. From these outputs, the best ΔF may be chosen. Furthermore, if searching for other properties, like the sequence corresponding to smallest ΔF_{ML} or one specific ΔF_{ML} value, one needs to modify only the fitness function where the desired property has the largest fitness value.

Overall, we have shown our algortihm can develop a surrogate model for the generation of sequences that optimally bind to the surfaces of interest. Because the surfaces and the

https://doi.org/10.1021/acsami.2c08891

polymers in question here are relatively simple, the optimal solutions are, as well. However, a surrogate model trained as we train the models in this study will also work on more complex polymer−surface interactions, and in those cases, it will not necessarily select for simple homopolymer and diblock structures.

CONCLUSION

We utilize a support vector regression model to predict adhesion free energy ΔF of polymer−surface interaction with its sequence information as input. In our work, we test four decorated surfaces with different patterns. The free energy ranges and energy distributions observed on the surfaces explored exhibit significant differences. The model inexpensively and reliably predicts adhesive free energies of polymer−surface interactions from sequential information. Though the free energies for four different surfaces are very different, each model exhibits very good accuracy in prediction. We identify how similar accuracy may be obtained with slightly fewer data and use the output of these models to design adhesive polymer sequences using a genetic algorithm, demonstrating good success in terms of this inverse design problem.

Our work highlights the promising integration of coarse-grained simulation with data-driven machine learning methods for obtaining quantitative relationships between polymer sequences and adhesive free energies and to use this information for the inverse design of polymer sequences. Our work thus represents a step forward from predicting the structural and functional properties of sequence-defined polymer chains themselves to predicting their interactions with surfaces, enabling the design of polymer sequences for desired polymer−surface interactions.[20] While our molecular simulation model in this work is a toy coarse-grained model that contains only two types of backbone beads, the techniques of the data-driven machine learning workflow are readily generalized to more complex and realistic polymer chain models that can help mimic biological processes[2−4] and practical applications.[1] Extensions of these studies to incorporate more specificity into the models and more general predictions of surface−polymer interactions represent targets for future work. As highlighted by the results presented here, there is reason to believe such refined strategies will be extremely successful in the creation of new adhesive materials.

AUTHOR INFORMATION

Corresponding Author

Jonathan K. Whitmer − *Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana 46556, United States; Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556, United States;* orcid.org/0000-0003-0370-0873; Email: jwhitme1@nd.edu

Authors

Jiale Shi − *Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana 46556, United States*

Michael J. Quevillon − *Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana 46556, United States*

Pedro H. Amorim Valença − *Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana 46556, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsami.2c08891

Author Contributions

J.S. and J.K.W. conceived the study and designed the research plan. J.S. conducted molecular dynamics simulation, machine learning model training, and inverse design using the genetic algorithm. M.J.Q. and P.H.A.V. aided J.S. with simulation design, scripting, and analysis. J.S., M.J.Q., P.H.A.V., and J.K.W. interpreted the results and wrote the paper.

Notes

The authors declare no competing financial interest.

Example scripts and information necessary to run the examples contained in this article are available at https://github.com/shijiale0609/ML_PSI.

ACKNOWLEDGMENTS

J.S., M.J.Q., P.H.A.V., and J.K.W. acknowledge support through the Midwest Integrated Center for Computation of Materials (MICCoM), a U.S. Department of Energy (DOE) Computational Materials Science Center funded by the DOE, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, for the development and maintenance of the SSAGES code used to calculate free energies in this work. Additional support is acknowledged from startup funding provided by the University of Notre Dame and computational resources at the Notre Dame Center for Research Computing (CRC). J.S. and J.K.W. acknowledge fruitful discussions with Ruimin Ma (University of Notre Dame) and Yamil Colón (University of Notre Dame).

ADDITIONAL NOTE

[a]To illustrate, if one polymer sequence is [00110011001100110011] and the corresponding adhesive energy is ΔF_A, we can add its reversed sequence [11001100110011001100] and ΔF_A without running MD simulations, but if one polymer sequence is a palindrome like [00000111111111100000] with ΔF_B, whose forward order and backward order are the same, only one sequence is used so as not to impart undue influence from that sequence to the SVR model.

REFERENCES

(1) Ouk Kim, S.; Solak, H. H.; Stoykovich, M. P.; Ferrier, N. J.; De Pablo, J. J.; Nealey, P. F. Epitaxial Self-Assembly of Block Copolymers on Lithographically Defined Nanopatterned Substrates. *Nature* **2003**, *424*, 411−414.

(2) Chakraborty, A. K.; Golumbfskie, A. J. Polymer Adsorption−Driven Self-Assembly of Nanostructures. *Annu. Rev. Phys. Chem.* **2001**, *52*, 537−573.

(3) Chakraborty, A. K. Disordered Heteropolymers: Models for Biomimetic Polymers and Polymers with Frustrating Quenched Disorder. *Phys. Rep.* **2001**, *342*, 1−61.

(4) Xiu, S.; Dick, A.; Ju, H.; Mirzaie, S.; Abdi, F.; Cocklin, S.; Zhan, P.; Liu, X. Inhibitors of SARS-CoV-2 Entry: Current and Future Opportunities. *J. Med. Chem.* **2020**, *63*, 12256−12274.

(5) Chiang, M.; Michieletto, D.; Brackley, C. A.; Rattanavirotkul, N.; Mohammed, H.; Marenduzzo, D.; Chandra, T. Polymer Modeling Predicts Chromosome Reorganization in Senescence. *Cell Rep* **2019**, *28*, 3212−3223.

(6) Wong, J. P.; Damania, B. SARS-CoV-2 Dependence on Host Pathways. *Science* **2021**, *371*, 884−885.

(7) Callaway, E. Making Sense of Coronavirus Mutations. *Nature* **2020**, *585*, 174−177.

(8) Plante, J. A.; Liu, Y.; Liu, J.; Xia, H.; Johnson, B. A.; Lokugamage, K. G.; Zhang, X.; Muruato, A. E.; Zou, J.; Fontes-Garfias, C. R.; Mirchandani, D.; Scharton, D.; Bilello, J. P.; Ku, Z.; An, Z.; Kalveram, B.; Freiberg, A. N.; Menachery, V. D.; Xie, X.; Plante, K. S.; Weaver, S. C.; Shi, P.-Y. Spike Mutation D614G Alters SARS-CoV-2 Fitness. *Nature* **2021**, *592*, 116−121.

(9) Hie, B.; Zhong, E. D.; Berger, B.; Bryson, B. Learning the Language of Viral Evolution and Escape. *Science* **2021**, *371*, 284−288.

(10) Annabi, N.; Tamayol, A.; Shin, S. R.; Ghaemmaghami, A. M.; Peppas, N. A.; Khademhosseini, A. Surgical Materials: Current Challenges and Nano-Enabled Solutions. *Nano Today* **2014**, *9*, 574−589.

(11) Ozboyaci, M.; Kokh, D. B.; Corni, S.; Wade, R. C. Modeling and Simulation of Protein—Surface Interactions: Achievements and Challenges. *Q. Rev. Biophys.* **2016**, *49*, No. e4.

(12) Chakraborty, A. K.; Bratko, D. A Simple Theory and Monte Carlo Simulations for Recognition Between Random Heteropolymers and Disordered Surfaces. *J. Chem. Phys.* **1998**, *108*, 1676−1682.

(13) Muthukumar, M. Pattern Recognition by Polyelectrolytes. *J. Chem. Phys.* **1995**, *103*, 4723−4731.

(14) Muthukumar, M. Pattern Recognition in Self-Assembly. *Curr. Opin. Colloid Interface Sci.* **1998**, *3*, 48−54.

(15) Chauhan, G.; Simpson, M. L.; Abel, S. M. Crowding-Induced Interactions of Ring Polymers. *Soft Matter* **2021**, *17*, 16−23.

(16) Kriksin, Y. A.; Khalatur, P. G.; Khokhlov, A. R. Adsorption of Multiblock Copolymers onto a Chemically Heterogeneous Surface: A Model of Pattern Recognition. *J. Chem. Phys.* **2005**, *122*, 114703.

(17) Webb, M. A.; Jackson, N. E.; Gil, P. S.; de Pablo, J. J. Targeted Sequence Design Within the Coarse-Grained Polymer Genome. *Sci. Adv.* **2020**, *6*, abc6216.

(18) Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best Practices in Machine Learning for Chemistry. *Nat. Chem.* **2021**, *13*, 505−508.

(19) de Pablo, J. J.; Jackson, N. E.; Webb, M. A.; Chen, L.-Q.; Moore, J. E.; Morgan, D.; Jacobs, R.; Pollock, T.; Schlom, D. G.; Toberer, E. S.; Analytis, J.; Dabo, I.; DeLongchamp, D. M.; Fiete, G. A.; Grason, G. M.; Hautier, G.; Mo, Y.; Rajan, K.; Reed, E. J.; Rodriguez, E.; Stevanovic, V.; Suntivich, J.; Thornton, K.; Zhao, J.-C. New Frontiers for the Materials Genome Initiative. *npj Comput. Mater.* **2019**, *5*, 41.

(20) Gormley, A. J.; Webb, M. A. Machine Learning in Combinatorial Polymer Chemistry. *Nat. Rev. Mater.* **2021**, *6*, 642−644.

(21) Huang, K.; Fu, T.; Glass, L. M.; Zitnik, M.; Xiao, C.; Sun, J. DeepPurpose: A Deep Learning Library for Drug−target Interaction Prediction. *Bioinformatics* **2021**, *36*, 5545−5547.

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

(23) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; Velankar, S.; Kleywegt, G. J.; Bateman, A.; Evans, R.; Pritzel, A.; Figurnov, M.; Ronneberger, O.; Bates, R.; Kohl, S. A. A.; Potapenko, A.; Ballard, A. J.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Clancy, E.; Reiman, D.; Petersen, S.; Senior, A. W.; Kavukcuoglu, K.; Birney, E.; Kohli, P.; Jumper, J.; Hassabis, D. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596*, 590−596.

(24) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373*, 871−876.

(25) Ma, R.; Colón, Y. J.; Luo, T. Transfer Learning Study of Gas Adsorption in Metal−Organic Frameworks. *ACS Appl. Mater. Interfaces* **2020**, *12*, 34041−34048.

(26) Ma, R.; Huang, D.; Zhang, T.; Luo, T. Determining Influential Descriptors for Polymer Chain Conformation based on Empirical Force-Fields and Molecular Dynamics Simulations. *Chem. Phys. Lett.* **2018**, *704*, 49−54.

(27) Perry, S. L.; Sing, C. E. 100th Anniversary of Macromolecular Science Viewpoint: Opportunities in the Physics of Sequence-Defined Polymers. *ACS Macro Lett.* **2020**, *9*, 216−225.

(28) Ferguson, A. L.; Ranganathan, R. 100th Anniversary of Macromolecular Science Viewpoint: Data-Driven Protein Design. *ACS Macro Lett.* **2021**, *10*, 327−340.

(29) Statt, A.; Casademunt, H.; Brangwynne, C. P.; Panagiotopoulos, A. Z. Model for Disordered Proteins with Strongly Sequence-Dependent Liquid Phase B ehavior. *J. Chem. Phys.* **2020**, *152*, 075101.

(30) Statt, A.; Kleeblatt, D. C.; Reinhart, W. F. Unsupervised Learning of Sequence-Specific Aggregation Behavior for a Model Copolymer. *Soft Matter* **2021**, *17*, 7697−7707.

(31) Reinhart, W. F.; Statt, A. Opportunities and Challenges for Inverse Design of Nanostructures with Sequence Defined Macromolecules. *Acc. Mater. Res.* **2021**, *2*, 697−700.

(32) Meenakshisundaram, V.; Hung, J.-H.; Patra, T. K.; Simmons, D. S. Designing Sequence-Specific Copolymer Compatibilizers Using a Molecular-Dynamics-Simulation-Based Genetic Algorithm. *Macromolecules* **2017**, *50*, 1155−1166.

(33) Kremer, K.; Grest, G. S. Dynamics of Entangled Linear Polymer Melts: A Molecular-Dynamics Simulation. *J. Chem. Phys.* **1990**, *92*, 5057−5086.

(34) Estridge, C. E.; Jayaraman, A. Diblock Copolymer Grafted Particles as Compatibilizers for Immiscible Binary Homopolymer Blends. *ACS Macro Lett.* **2015**, *4*, 155−159.

(35) Lang, R. J.; Merling, W. L.; Simmons, D. S. Combined Dependence of Nanoconfined T_g on Interfacial Energy and Softness of Confinement. *ACS Macro Lett.* **2014**, *3*, 758−762.

(36) Zhan, B.; Shi, K.; Dong, Z.; Lv, W.; Zhao, S.; Han, X.; Wang, H.; Liu, H. Coarse-Grained Simulation of Polycation/DNA-Like Complexes: Role of Neutral Block. *Mol. Pharmaceutics* **2015**, *12*, 2834−2844.

(37) Grest, G. S.; Kremer, K. Molecular Dynamics Simulation for Polymers in the Presence of a Heat Bath. *Phys. Rev. A* **1986**, *33*, 3628−3631.

(38) Stevens, M. J.; Kremer, K. Structure of Salt-Free Linear Polyelectrolytes. *Phys. Rev. Lett.* **1993**, *71*, 2228−2231.

(39) Kremer, K.; Grest, G. S. Molecular Dynamics (MD) Simulations for Polymers. *J. Phys.: Condens. Matter* **1990**, *2*, SA295−SA298.

(40) Zhou, Q.; Akhavan, R. Cost-effective Multi-mode FENE Bead-spring Models for Dilute Polymer Solutions. *J. Non-Newton. Fluid Mech* **2004**, *116*, 269−300.

(41) Peters, B. *Reaction Rate Theory and Rare Events*; Elsevier, 2017.

(42) Darve, E.; Rodríguez-Gómez, D.; Pohorille, A. Adaptive Biasing Force Method for Scalar and Vector Free Energy Calculations. *J. Chem. Phys.* **2008**, *128*, 144120.

(43) Sidky, H.; Colón, Y. J.; Helfferich, J.; Sikora, B. J.; Bezik, C.; Chu, W.; Giberti, F.; Guo, A. Z.; Jiang, X.; Lequieu, J.; Li, J.; Moller, J.; Quevillon, M. J.; Rahimi, M.; Ramezani-Dakhel, H.; Rathee, V. S.; Reid, D. R.; Sevgen, E.; Thapar, V.; Webb, M. A.; Whitmer, J. K.; de Pablo, J. J. SSAGES: Software Suite for Advanced General Ensemble Simulations. *J. Chem. Phys.* **2018**, *148*, 044104.

(44) Drucker, H.; Burges, C. J.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. *Advances in Neural Information Processing Systems 9 (NIPS 1996)* **1997**, *9*, 155−161.

(45) Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*; Springer-Verlag: Berlin, 2006.

(46) Murphy, K. P. *Machine Learning: A Probabilistic Perspective*; MIT Press, 2012.

(47) Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273−297.

(48) Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1−27.

(49) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825−2830.

(50) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27−35.

(51) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757−1772.

(52) Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; Jensen, K. F.; Olsen, B. D. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Cent. Sci.* **2019**, *5*, 1523−1531.