Data-driven prediction of polymer surface adhesion using high-throughput MD and hybrid network models

Sibasankar Panigrahy
sibasankar panigrahy@iitd.ac.in alenj@iitd.ac.in divyanayar@iitd.ac.in

Department of Materials Science and Engineering, Indian Institute of Technology Delhi, New Delhi, India

Abstract

Designing of functional polymers with tailored applications requires an understanding of their interactions with surfaces that find applications in chemical, biological and industrial processes. This work involves computing and predicting the adhesive free energies of generic, coarse-grained polymers to generic functionalized surfaces having variable sequences, composition, spatial patterns and polymer-surface interaction strengths. Enhanced molecular dynamics simulations (approximately 0.85 million simulations) are employed to generate a unique synthetic dataset of 8464 polymer-surface binding free energies (BFEs). The computed BFEs account for the effects of polymer sequences, polymer/surface composition and variable polymer-surface interaction strengths. This dataset is used for training the CNN-GRU attention-based hybrid model that accurately predicts the BFEs as well as complete free energy profiles of the polymers binding to surfaces with unknown sequences and composition. Our work provides a unified approach using MD simulations with ML to accelerate the design of functional polymers for various interfacial applications.

1 Introduction

Polymer-surface interactions are central to applications ranging from drug delivery and tissue engineering to biosensing and nanofabrication.[15, 4] In biological systems, these interactions govern adhesion, uptake, and molecular recognition on biological membranes or protein surfaces.[7] Such interactions are central to the design of next-generation drug delivery and therapeutic systems, with critical roles in controlling drug release, achieving tissue-specific targeting, and enabling immune evasion.[12, 8, 9] The heterogeneous hydrophilic and hydrophobic patches on these surfaces affect the adsorption, binding and folding of biomolecules. Small modifications in biomolecule (or biopolymer) and surface composition can drastically alter biological interactions and therapeutic efficiency. Therefore, an optimized design of the functional polymers and surfaces would require accounting for these factors. Conventional approaches often involve extensive experimental screening of these materials, which is ineffcient in terms of cost and time.

Molecular simulations and machine learning (ML) approaches have emerged as promising tools for accelerating the design of such materials.[1, 2] Prior molecular simulation studies have examined polymer conformational transitions and adsorption interfacial thermodynamics.[6, 5] A Monte Carlo simulation study involved adsorption of random heteropolymers on disordered multifunctional surfaces, as determined by the interplay of polymer sequence and the surface site distribution.[3] Another study utilized both theoretical analysis and Monte Carlo simulations to show that the self-assembly of polyelectrolyte chains on oppositely charged patterned surfaces is largely affected by the

charge density, the size of the pattern, and the Debye length. [10, 11] As it is known that for training ML models accurately, large datasets are required. Generating large datasets for different patterned real surfaces and different sequences of real polymers would require performing thousands of experiments and simulations, that would be intractable. Although MD simulations explicitly capture sequence-dependent behavior, they are computationally intensive and are feasible only for relatively short polymer chains. This limits large-scale exploration of polymer-surface design spaces, even with enhanced sampling methods that are constrained by choice of collective variables. Therefore, there is a need for complementary data-driven models that can generalize across sequence and surface design spaces. Prior studies have involved ML-guided strategy to optimize polymer properties,[13] and integration of simulations and experiments to improve biomaterial design. [16] A recent investigation involved extensive training of ML models on 20,000 coarse-grained polymer sequences interacting with structured surfaces, revealing that even small changes in monomer arrangement can drastically alter the binding free energy landscape.[14] These studies mainly highlighted the role of polymer sequence on surface adhesion. However, extensive investigation of design principles based on the combined effects of polymer sequence, polymer/surface composition, spatial patterning, and polymer-surface interaction strengths is lacking.

In this work, we present a robust polymer-surface binding model that minimizes the number of simulations by selecting generic coarse-grained models for the polymers and surfaces. Although coarse-grained (CG) MD simulations may not reproduce experimental free energies quantitatively, they have been shown to capture reliable qualitative trends in adhesion, sequence dependence, and conformational transitions. We employ umbrella sampling simulations to generate high-resolution potential of mean force (PMF) profiles across a large dataset of surfaces and polymers, where the underlying patterns were systematically designed via an Ising Monte Carlo framework to capture diverse spatial and compositional arrangements. ML models such as neural network models are accurate predictors of polymer-surface binding free energies (BFEs), although they are limited in yielding direct physical principles or simple design rules. Therefore, we build a hybrid neural architecture that integrates convolutional, recurrent, and attention-based layers that yield highly accurate predictions of BFEs with the ability to extract underlying design principles.

2 Methods

2.1 Dataset generation

The adsorption of coarse-grained model of random co-block polymer chains was examined onto solid surfaces. The surfaces were modelled to be homogeneous but chemically heterogeneous. The polymers and surfaces were modelled using strongly (A' or A) and weakly (B' or B) interacting beads (the scheme is summarized in Fig. S1). Polymer sequences and surface patterns of different compositions were designed using the Ising model. Monte Carlo simulations of Ising models (1D for polymer sequences and 2D for surface patterns) were utilized to generate sequences of specified fractions (0.1 to 0.9 in steps of 0.1) of a bead based on the Metropolis-Hastings algorithm (Fig. 1). The detailed methodology is described in the Supplementary material. In this framework, at each fraction we sampled 10 independent configurations to capture statistical variability. In total, we obtained 92 unique surface patterns $(20\times20 \text{ lattices})$ and 92 unique polymer sequences (40-bead chains). The surfaces were fixed and polymers were free to move in three dimensions. Free energy profiles were obtained via umbrella sampling molecular dynamics simulations using the polymer center-of-mass distance from the surface (z-axis) as the collective variable (CV). More details are provided in the Supplementary material.

2.2 Model architecture and training

Convolutional neural networks (CNNs) are particularly effective for learning representations from high-dimensional spatial data, where local correlations and translational invariances are important. Conversely, recurrent neural networks (RNNs) have a dynamic memory of past inputs so they may find dependencies in sequential data. In the context of polymer–surface interactions, the surface can be naturally represented as a two-dimensional matrix encoding spatial information, while the polymer can be described as a one-dimensional sequence of beads whose binding affinity depends on local motifs and overall ordering. By merging both architectures, it is possible to create a forward model that utilizes CNNs to encode surface characteristics and RNNs to obtain polymer sequence data,

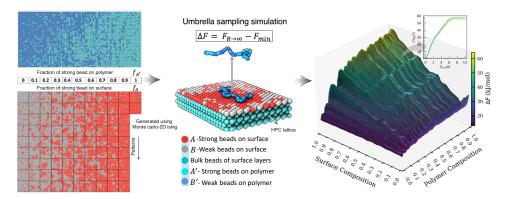


Figure 1: Workflow for generating the polymer-surface binding free energy dataset showing the surface patterns (20×20) and polymer sequences (40-bead) (total 92 surfaces × 92 polymers) with beads of different interaction strength. Umbrella sampling simulations (491104) provided ΔF as a function of polymer and surface composition to map the entire space.

facilitating predictive modeling of their combined potential mean force (PMF) landscape. Building

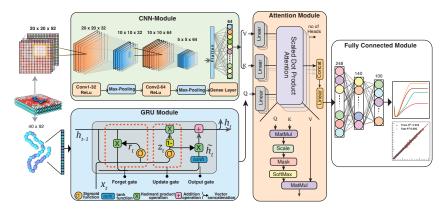


Figure 2: Hybrid deep learning architecture for forward prediction of polymer–surface interactions. Surfaces are processed through a CNN module, polymer sequences are encoded with a GRU module, and the learned representations are combined via a multi-head attention block. The fused features are passed through fully connected layers to predict the PMF profile.

on this rationale, a hybrid deep learning model is developed to predict the potential of mean force (PMF) from a given surface-polymer configuration (Fig. 2). The model splits the input into two components:1) a surface segment reshaped to a 20x20x1 tensor and passed through 2D convolutional layers,2) a polymer segment itself reshaped to (40,1) and then run through a single-layer gated recurrent unit (GRU). The CNN encoder includes three 2D-convolutional layers with increasing filter depths, each followed by a max-pooling. A dense layer flattens and projects onto a 64-dimensional feature vector, allowing sequence-processing layers to jointly learn interaction patterns between local polymer beads and corresponding surface motifs. The model incorporates a cross-attention mechanism based on a transformer block, where the polymer encoded from the GRU serves as the query, while the surface encoding provides the keys and values. The mathematics and details are discussed in Supplementary material. The architecture allows each polymer bead to attach selectively to spatial features of the surface, enabling the model to learn position-specific interaction preferences. This is essential for reconstructing characteristic features of the PMF, such as the depth of the free energy minimum and the steepness of the adsorption barrier, which are sensitive to how and where these strong monomer-surface associations occur along the sequence. A full description of the network architecture, along with the corresponding parameters and hyperparameters, is given in Table S1.

3 Results and Discussion

3.1 Binding free energies from the dataset

Umbrella sampling simulations of polymer–surface binding provided a comprehensive dataset of potential of mean force (PMF) profiles and the binding free energy, ΔF . Fig. 3a shows that the distribution of ΔF as a function of polymer composition is not sensitive to change in polymer composition and varies smoothly over a wide range of ΔF . Whereas, Fig. 3b indicates that the distribution is sensitive to the change in surface composition and exhibits a steep gradient. This implies that surface heterogeneity strongly affects the binding as compared to polymer composition, highlighting the significant role of surface composition in determining the binding strength. Representative PMF profiles in Fig. 3c show that the binding free energy increases sharply with the change in surface composition, as indicated by steeper slopes and deeper free energy minima. For better visualization, the PMF profiles have been scaled such that the plateau region is set to zero, providing a clearer comparison of free-energy differences (see Fig. S2). In contrast, varying polymer composition at fixed surface composition (Fig. 3d) produces more gradual changes, with narrower differences in adsorption strength.

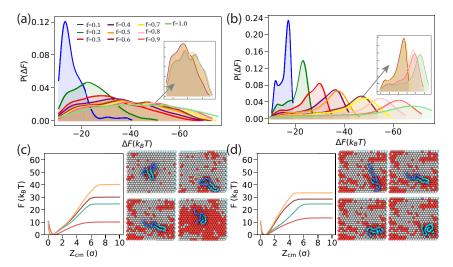


Figure 3: Distribution of ΔF values across different (a) polymer compositions. (b) surface compositions with an inset showing for f= 0.6 and 1.0. (c,d) PMF profiles for selected surface–polymer systems and simulation snapshots corresponding to the polymer conformations adsorbed on the surfaces.

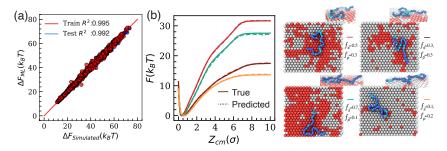


Figure 4: (a)Correlation plot between predicted and simulated ΔF values. (b) True and predicted PMF profiles for the surface and polymer composition as shown in the snapshots.

3.2 Predicting binding free energy profiles

Building on this dataset, we developed a hybrid deep learning model to predict PMFs directly from sequence and pattern information. The hybrid CNN–GRU–Transformer forward model was trained

on this dataset and achieved a mean absolute error (MAE) of 0.45 and an R² of 0.992 in predicting complete PMF profiles, demonstrating excellent agreement with simulation data (Fig. 4a and Fig. S2). Its performance benefits from complementary encoders i.e. CNNs that encode surface spatial features, GRUs that represent polymer sequences, and the transformer block that connects them by attending to critical surface-polymer relationships. Representative predictions (Fig. 4b) captured both strongly and weakly adsorbing systems, with only minor deviations (1-3 k_BT) in plateau regions. To assess whether the model's attention captures physically meaningful information, we projected the attention outputs into two dimensions using t-SNE (Fig. S4). The resulting embedding exhibits a gradient in ΔF values, with points on the right representing systems with high ΔF and points on the left representing systems with low ΔF . To evaluate the predictive performance of our proposed model, we benchmarked it against a baseline deep neural network (DNN) trained with the same input representations. We then performed an ablation study on the polymer encoder while keeping the 2D CNN surface encoder fixed, testing alternatives including 1D CNN, GRU, BiGRU, LSTM, and BiLSTM. As shown in Table S2, recurrent polymer encoders are particularly effective for capturing sequence-dependent interactions. These results confirm that the hybrid model effectively integrates spatial features of surfaces and sequential information of polymers to generalize across diverse polymer adsorption regimes.

4 Conclusions and Future work

An integrated approach is presented for predicting polymer adhesion to chemically heterogeneous surfaces using MD simulations and ML approach. A unique extensive synthetic dataset of binding free energies is developed using enhanced MD simulations of polymers on surfaces with varying, sequences, compositions, spatial patterns and interaction strengths. This dataset is used to train the attention-based hybrid CNN model that performs accurately in predicting the binding free energies as well as complete binding free energy profile of polymers on surfaces. With simplified coarse-grained polymer and surface models, we have developed a machine learning workflow that can be generalized to more complex and realistic polymer chains. This opens up the possibility of studying biologically relevant interactions and designing functional materials. In the next steps, the aim is to inverse design coarse-grained polymer and surface models based on the desired adhesion free energy using extended tandem neural network architecture. The ML workflow developed in this work can be extended using atomistic simulations to apply sequence-design principles to more realistic self-assembled monolayers (SAMs) and biomolecular systems, and for engineering interfacial interactions between biomolecules and surfaces.

References

- [1] Nongnuch Artrith, Keith T. Butler, François Xavier Coudert, Seungwu Han, Olexandr Isayev, Anubhav Jain, and Aron Walsh. Best practices in machine learning for chemistry. *Nature Chemistry*, 13(6):505–508, 2021.
- [2] Simon Axelrod. Learning matter: Materials design with machine learning and atomistic simulations. *Accounts of Materials Research*, 3(3):343–357, 2022.
- [3] D. Bratko, Arup K. Chakraborty, and Eugene I. Shakhnovich. Recognition between random heteropolymers and multifunctional disordered surfaces. *Chemical Physics Letters*, 280(1-2): 46–52, 1997. ISSN 00092614. doi: 10.1016/S0009-2614(97)01075-0.
- [4] Junghyun Cho, Jinwoo Oh, Joona Bang, Jai Hyun Koh, Hoon Yeub Jeong, Seungjun Chung, and Jeong Gon Son. Roll-to-plate 0.1-second shear-rolling process at elevated temperature for highly aligned nanopatterns. *Nature Communications*, 14(1), 2023.
- [5] Alexandros Chremos, Emmanouil Glynos, Vasileios Koutsos, and Philip J. Camp. Adsorption and self-assembly of linear polymers on surfaces: A computer simulation study. *Soft Matter*, 5 (3):637–645, 2009.
- [6] Arthi Jayaraman, Carol K. Hall, and Jan Genzer. Designing pattern-recognition surfaces for selective adsorption of copolymer sequences using lattice Monte Carlo simulation. *Physical Review Letters*, 94(7):4–7, 2005.

- [7] Wen Dong Liu and Bai Yang. Patterned surfaces for biological applications: A new platform using two dimensional structures as biomaterials. *Chinese Chemical Letters*, 28(4):675–690, 2017.
- [8] Yutaka Miura, Tomoya Takenaka, Kazuko Toh, Shourong Wu, Hiroshi Nishihara, Mitsunobu R. Kano, Yasushi Ino, Takahiro Nomoto, Yu Matsumoto, Hiroyuki Koyama, Horacio Cabral, Nobuhiro Nishiyama, and Kazunori Kataoka. Cyclic rgd-linked polymeric micelles for targeted delivery of platinum anticancer drugs to glioblastoma through the blood-brain tumor barrier. ACS Nano, 7(10):8583–8592, 2013.
- [9] Jingxin Mo, Xianjue Chen, Meiying Li, Wenxu Liu, Wei Zhao, Lee Yong Lim, Richard D. Tilley, J. Justin Gooding, and Qinghua Li. Upconversion nanoparticle-based cell membrane-coated crgd peptide bioorthogonally labeled nanoplatform for glioblastoma treatment. ACS Applied Materials Interfaces, 14(44):49454–49470, 2022.
- [10] Muthukumar. Pattern recognition by polyelectrolytes. *The Journal of Chemical Physics*, 103 (11):4723–4731, 09 1995.
- [11] Murugappan Muthukumar. Pattern recognition in self-assembly. *Current Opinion in Colloid Interface Science*, 3(1):48–54, 1998. ISSN 1359-0294.
- [12] Roberto Oliva, Serena Maria Torcasio, Olivier Coulembier, Anna Piperno, Antonino Mazzaglia, Silvia Scalese, Arianna Rossi, Giada Bassi, Silvia Panseri, Monica Montesi, and Angela Scala. Rgd-tagging of star-shaped pla-peg micellar nanoassemblies enhances doxorubicin efficacy against osteosarcoma. *International Journal of Pharmaceutics*, 657:124183, 2024.
- [13] Roshan A. Patel and Michael A. Webb. Data-Driven Design of Polymer-Based Biomaterials: High-throughput Simulation, Experimentation, and Machine Learning. *ACS Applied Bio Materials*, 7(2):510–527, 2024. ISSN 25766422. doi: 10.1021/acsabm.2c00962.
- [14] Jiale Shi, Michael J. Quevillon, Pedro H. Amorim Valença, and Jonathan K. Whitmer. Predicting Adhesive Free Energies of Polymer-Surface Interactions with Machine Learning. ACS Applied Materials and Interfaces, 14(32):37161–37169, 2022. ISSN 19448252. doi: 10.1021/acsami. 2c08891.
- [15] Nguyen Hoc Thang, Truong Bach Chien, and Dang Xuan Cuong. Polymer-based hydrogels applied in drug delivery: An overview. *Gels*, 9(7), 2023. ISSN 2310-2861.
- [16] Michael A. Webb, Nicholas E. Jackson, Phwey S. Gil, and Juan J. de Pablo. Targeted sequence design within the coarse-grained polymer genome. *Science Advances*, 6(43), 2020. ISSN 23752548. doi: 10.1126/sciadv.abc6216.

A Supplementary Material

A.1 Methods

System setup details The initial configurations for the simulations were randomly generated sequences or orderly patched sequences with varied fractions of particles. We start by choosing a random site to check the feasibility of flipping the spin associated with the site. The energy required to flip the spin associated with a site i is calculated using $\Delta E_i = 2J.\sigma_i. \sum_j \sigma_j$. Here J is the interaction strength, σ_i is the current spin state (+1, -1) of site i, and σ_j is the current spin state (+1, -1) of neighbouring sites j. The interaction neighborhood for spins can be defined using either the von Neumann or Moore topology, with periodic boundary conditions applied. The algorithm differs from the usual Metropolis-Hastings algorithm by the introduction of a biasing term (f_{bias}) for fraction control. By controlling the strength (k) of this term, we ensure that the system configuration is aptly biased towards the desired fraction of particles. If $\Delta E_i + k.(f_A - f_B) < 0$, the spin flip is implemented, else we execute a spin flip with an acceptance probability of $e^{-\beta.\Delta E}$, where β is a noise term. he composition of strong bead sites is systematically varied from 0 to 1, and ten distinct surface patterns are generated for each composition. The surfaces are designed with a hexagonal closed packing lattice system, with an energy of adsorption ranging from 10 to 70 k $_B$ T. The surface has four layers and is positioned 1 nm above the bottom of the simulation box.

MD Simulation details The interaction energy between polymer beads is tuned to control flexibility, allowing the polymer to better investigate the complex energy landscape of surfaces. A flat-bottom potential is applied in the x and y directions to prevent the polymer from crossing into periodic images and ensure lateral confinement. The polymer was allowed to move under the influence of inter- and intra-molecular interactions and polymer surface interactions. The bonded interactions in the polymer were represented using the FENE potential. The simulations were run for 4,000,000 steps, equivalent to 20 ns of simulation time. The surface beads are fixed during the simulations, while the polymer chains move freely in three dimensions. Molecular Dynamics (MD) simulations were conducted using the GROMACS 2021.4 package. A Langevin thermostat maintained the system temperature at 300 K.

Umbrella sampling simulations were performed by applying a harmonic bias potential to the collective variable (CVs), which was defined as the distance between the surface and the center-of-mass of the polymer along the z-axis. A force constant of 500 kJ mol⁻¹ nm⁻². 58 independent simulations (windows) were performed for CV ranging between 0 and 10.4 nm. The potential of mean force (PMF) profiles were computed as a function of this CV. All molecular dynamics simulations were performed using GROMACS [version 2021.4] with PLUMED [version 2.8] plugin. Biased trajectories were subsequently reweighted using the weighted histogram analysis method (WHAM) to reconstruct the PMFs. The dataset is accessible in the following link: https://github.com/siba-p/ML_surfacepoly.

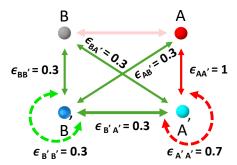


Figure S1: Schematic of the modelled interactions between polymer and surface beads. Polymer was modelled with A' and B' beads. Surfaces were modelled with A and B beads.

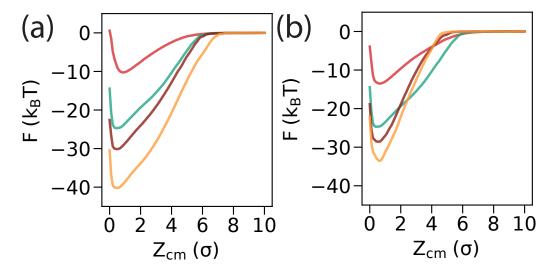


Figure S2: PMF profiles for selected surface–polymer systems and simulation snapshots corresponding to the polymer conformations adsorbed on the surfaces. The PMF curves have been scaled such that the plateau region is set to zero

Data preprocessing To reduce data redundancy and improve model generalization, we implemented a symmetry-based data augmentation for both surfaces and polymers. An important part of this stage was realizing that various combinations of surface polymers are physically identical because of underlying symmetries (like flipping and rotation). For example, rotating a surface by 90°, 180°, or 270° does not change how it interacts physically. Here each surface is represented as an OHE of binary matrix $S \in \{0,1\}^{20 \times 20}$, and each polymer is a binary sequence $P \in \{0,1\}^{40}$.

Canonical Form and Symmetry Invariance

We define a canonical form for each surface and polymer by selecting the lexicographically minimal transformation, making all symmetry-equivalent variants into a single, representative form. For a surface S, its canonical form is given by:

$$Canonical(S) = \min \{ Flatten(T(S)) \mid T \in \mathcal{T} \}$$

where \mathcal{T} includes all valid 90° rotations and horizontal/vertical flips:

$$\mathcal{T} = \{I, R_{90}, R_{180}, R_{270}, \text{flip}_{h}, \text{flip}_{v}\}$$

All unique invariants for both surfaces and polymers are computed and stored. This guaranteed that structurally equivalent sequences are treated uniformly throughout training. Many current methods use random augmentation or interpret symmetry-transformed configurations as different/distinct. Our method, on the other hand, deterministically finds a canonical representative for each surface-polymer. This not only cuts down on duplication, but also makes sure that the training data is physically consistent, which is important in pairwise interaction systems where positional equivalence is quite important.

We defined the model $f: \mathbb{R}^{440} \to \mathbb{R}^{100}$, which maps $\mathbf{x} = [\mathbf{x}_{\text{surf}}, \mathbf{x}_{\text{poly}}])$ to a PMF profile $\mathbf{y} \in \mathbb{R}^{100}$. The surface input $\mathbf{x}_{\text{surf}} \in \{0,1\}^{400}$ represents a binary 20×20 matrix $\mathbf{S} \in \{0,1\}^{20 \times 20}$. The polymer input $\mathbf{x}_{\text{poly}} \in \{0,1\}^{40}$, the targeted output is a PMF curve $\mathbf{y} \in \mathbb{R}^{100}$.

The final polymer feature is: $\mathbf{H_{poly}} \in \mathbb{R}^{T \times d}$

We applied a Transformer attention block, where:

Query:
$$Q = LN(\mathbf{H_{poly}})$$

Key, Value:
$$K, V = LN(\mathbf{S_{surf}})$$

For Multi-head attention with h heads and each head of size $d_h = \frac{d}{h}$, then i^{th} head can be calculated as:

$$head_i = Softmax \left(\frac{\mathbf{QW}_i^{Q}.(\mathbf{KW}_i^{K})}{\sqrt{d_h}} \right).(\mathbf{VW}_i^{V})$$

After concatenating all heads and projecting::

$$A = Concat(head_1, head_2, \cdots head_i).\mathbf{W}^{\mathbf{o}}$$

This output is then then added to the original Query and residual connection: $\mathbf{H_1} = \mathbf{H_{poly}} + A$ Then a position wise FFN is applied:

$$H_2 = H_1 + \Phi(H_1W_1 + b_1)W_2 + b_2$$

After flattening: $\mathbf{z} = vec(\mathbf{H_2}) \in \mathbb{R}^{T \times d}$

This output is passed through FFN, where $\hat{y}(PMF)$ is predicted.

Table S1: Summary of the parameters used in hybrid CNN-GRU-Transformer forward model

Model Element	Description / Value	
Input dimension	440	
Surface encoder	2D CNN (2 conv blocks: 32, 64 filters), with MaxPooling	
Polymer encoder	GRU with 64 hidden units	
Transformer block	2-head attention with key dim 64	
Attention inputs	Query: polymer; Key/Value: surface sequence	
Feedforward network	Two dense layers after attention (64 and 64 units)	
Activation	ReLU (Conv, Dense), LeakyReLU (final MLP)	
Final dense layers	[250, 140, 100] with dropout = 0.2 after transformer	
Optimizer	Nadam ($\alpha = 10^{-4}, \beta_1 = 0.97, \beta_2 = 0.97$)	
Loss function	Mean Absolute Error (MAE)	

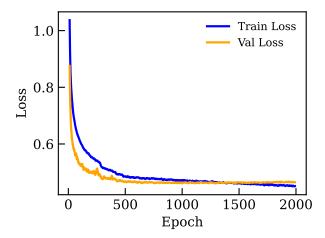


Figure S3: Training and validation loss as a function of epochs for the CNN–GRU attention model. Both losses decrease converge with increasing epochs, shows effective model learning and good generalization without significant overfitting

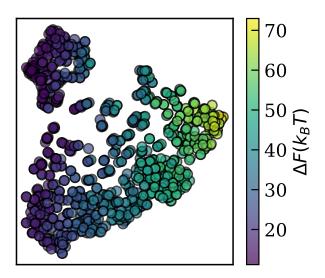


Figure S4: Two-dimensional t-SNE projection of the model's attention outputs.

Table S2: Comparison of hybrid CNN–polymer encoder models with transformer attention on PMF prediction.

Model	MAE	GPU time (Sec)
DNN (without attention)	0.631	862
CNN+GRU	0.452	1042
CNN+BiGRU	0.447	1175
CNN+LSTM	0.487	1098
CNN+BiLSTM	0.496	1231
2D CNN + CNN1D	0.539	556