Modeling Functional Random Heteropolymers with k-mer Representations

Anonymous Author(s)
Affiliation
Address
email

Abstract

Designing synthetic macromolecules with targeted functions is a long-standing challenge in materials science. Random heteropolymers (RHPs) and their blends provide a vast combinatorial design space whose physicochemical behavior emerges from short-range monomer interactions rather than global sequence order. However, the absence of explicit sequence information and the difficulty of simulating disordered polymer ensembles make structure–property prediction challenging. Here, we introduce a *k*-mer representation learning framework for modeling and optimizing functional random copolymers, trained on data acquired from an autonomous robotic blending platform. The platform executes high-throughput synthesis, testing, and iterative optimization of RHP blends for protein stabilization, generating $>10^3$ labelled experiments in closed-loop optimization campaigns. Each polymer or blend is encoded as a concatenated *k*-mer fingerprint that captures segment-level statistics of monomer connectivity derived from stochastic polymerization models. We demonstrate that the resulting *k*-mer features outperform one-hot composition encodings in predictive accuracy, revealing non-additive, physically interpretable correlations such as charge-pattern complementarity. The *k*-mer-based model generalizes across different experimental setups. This work shows how physics-grounded statistical representations of polymer structure can bridge experimental data and machine learning, providing a scalable framework for physically faithful modeling of disordered soft-matter systems.

1 Introduction

Blending polymers provides a practical and efficient route to create new materials with tailored properties using existing components Khan et al. [2019], Utracki et al. [2014]. Such blends have broad applications across plastics recycling, energy storage, and biomedical materials Lin et al. [2024], Blatt and Hallinan Jr [2021], Leyden et al. [2024]. However, discovering functional polymer blends remains difficult because their properties emerge from complex, non-additive interactions among components. Traditional computational methods, such as Flory–Huggins theory or molecular dynamics, can provide qualitative insights into compatibility but are often too limited or expensive to capture the nonlinear relationships that determine function Ethier et al. [2024], Liang et al. [2022].

Random heteropolymers (RHPs) are synthetic polymers composed of multiple monomer types that are statistically linked along the chain, producing heterogeneous sequences with complex local structures and emergent macroscopic behaviors. We recently developed an autonomous experimental workflow for exploring the design space of random heteropolymer blends (RHPBs), the mixture of RHPs (Figure 1). Using a robotic platform that performs high-throughput synthesis, testing, and optimization, we generated a dataset of over 10^3 labeled blend compositions, each evaluated by its ability to stabilize proteins under thermal stress. This dataset captures the effects of varying polymer composition, molecular weight, and mixing ratio on emergent functionality, providing a

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

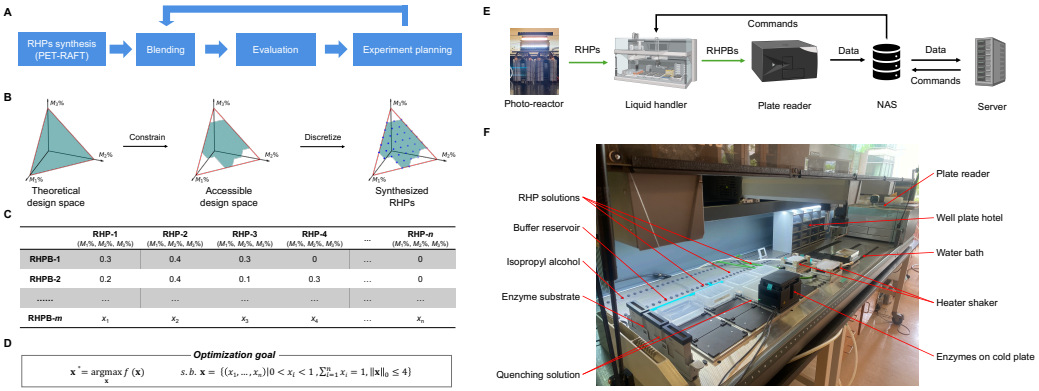


Figure 1: Overview of the autonomous closed-loop discovery process for RHPBs. (A) Workflow of the optimization process. (B) Illustration the RHP design space discretization. The theoretical design space is firstly constrained by physical factors such as the solubility of the resulting polymers, and then it is discretized according to a specified step length. (C) Exemplary representation of RHPBs. Each RHPB is a mixture of different RHPs and described by its composition (D) Mathematical formulation of the optimization goal with constraints. (E) Material and information flow during the optimization process. The green arrows indicate material flow and the black arrows indicate information flow. NAS: Network-Attached Storage. (F) Detailed view of the physical experimental platform and key components.

rare large-scale, physically grounded experimental corpus for modeling disordered macromolecular systems.

In this work, we applied the structure-informed k-mer representation for modeling this dataset. Each polymer or blend is encoded as a concatenated k-mer fingerprint that reflects segment-level connectivity and short-range monomer correlations derived from stochastic polymerization theory Compeau et al. [2011], Smith et al. [2018]. We demonstrate that these k-mer features outperform composition-based one-hot encodings in predictive accuracy and physical interpretability, capturing non-additive effects such as charge-pattern complementarity. This framework establishes a scalable and physically faithful approach for linking polymer composition to macroscopic function, offering a foundation for data-driven design of disordered soft-matter systems.

2 Methods

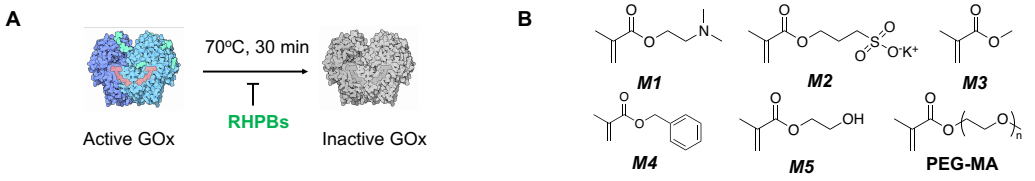


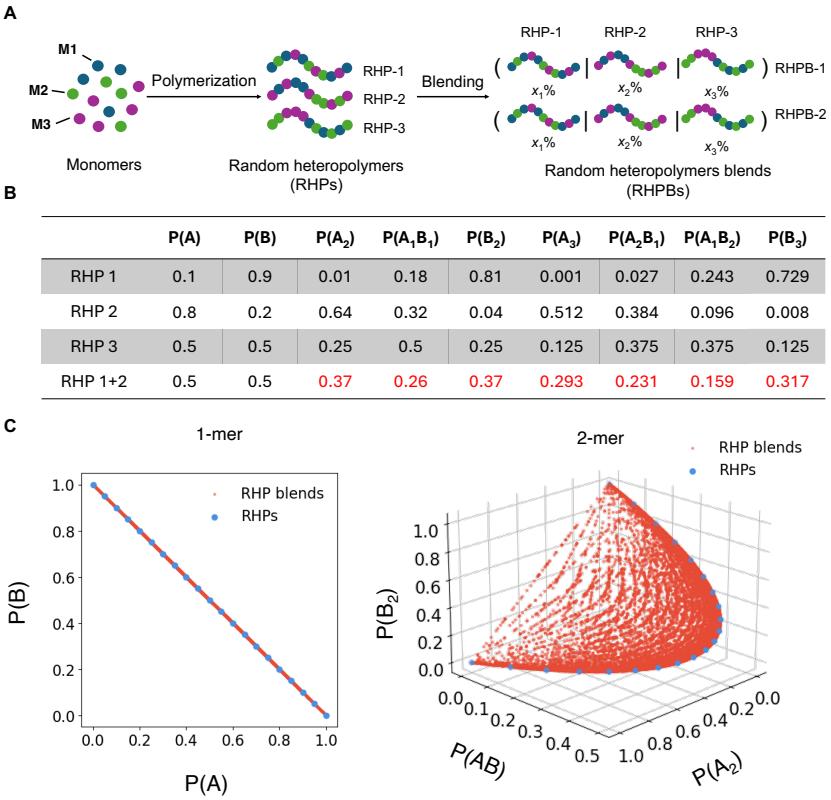
Figure 2: Dataset overview. (A) Illustration of the enzyme assay. (B) Structures of the monomers used in this research.

2.1 Dataset

The dataset used in this study was generated using an autonomous experimental platform. Briefly, the platform integrates automated liquid handling, plate-based high-throughput experimentation, and real-time data analysis to perform closed-loop optimization of RHPBs. Each RHP was synthesized from a set of pre-selected monomers via one-step copolymerization using PET-RAFT (Figure 2), resulting in polymers with statistically distributed sequences characterized by distinct monomer compositions and molecular weights. During the optimization, the robotic system blended up to four

RHPs in programmable ratios, dispensed the resulting mixtures into multi-well plates, and evaluated their ability to stabilize glucose oxidase (GOx) under thermal stress. The residual enzyme activity (REA) after heat treatment served as the functional metric (Figure 2). The final dataset comprises over 10^3 labeled blend compositions, each annotated with quantitative polymer descriptors (composition, molecular weight, blending ratio) and the corresponding measured REA.

2.2 k-mer representaion



	P(A)	P(B)	$P(A_2)$	$P(A_1B_1)$	$P(B_2)$	$P(A_3)$	$P(A_2B_1)$	$P(A_1B_2)$	$P(B_3)$
RHP 1	0.1	0.9	0.01	0.18	0.81	0.001	0.027	0.243	0.729
RHP 2	0.8	0.2	0.64	0.32	0.04	0.512	0.384	0.096	0.008
RHP 3	0.5	0.5	0.25	0.5	0.25	0.125	0.375	0.375	0.125
RHP 1+2	0.5	0.5	0.37	0.26	0.37	0.293	0.231	0.159	0.317

Figure 3: Blending RHPs achieves a broader design space than individual RHPs alone. (A) Schematic illustration of the synthesis of random heteropolymers (RHPs) with three monomers (M1, M2, and M3) through polymerization, and the blending of these RHPs with various compositions to form random heteropolymer blends (RHPBs). $x_i\%$ is the blending ratio of individual RHP-i. (B) Demonstration of how blending expand the design space as reflected by the change in k-mer distributions under ideal polymerization conditions where monomer reactivities are identical. The rows show the probabilities of k-mer ($k < 3$) for individual RHPs (RHP-1, RHP-2, RHP-3) and for a 1:1 blend of RHP-1 and RHP-2. Single RHP-3 can recover the 1-mer distribution of the blend, but is not able to match the $k \geq 2$ probabilities shown in red. (C) Visual comparison of 1-mer and 2-mer distributions for all potential RHPs versus all potential RHP blends in a two-monomer system. 21 RHPs (blue dots) were uniformly sampled from the design space, and 10,000 RHPBs (red dots) were generated by randomly selecting two RHPs and mixing them with a random blending ratio. While blending does not shift the 1-mer distribution, it accesses new regions of composition space for k-mers ($k \geq 2$), demonstrating the expanded diversity achievable through blending.

Blending RHPs expands the accessible design space beyond what can be achieved by tuning a single polymer's monomer composition (Figure 3A). Each RHP is characterized by a 1-mer representation—the normalized abundance of individual monomers along the chain, while its full k-mer representation describes the normalized abundance of all contiguous monomer segments of length k, analogous to k-gram statistics in genomics Compeau et al. [2011]. The k-mer distribution arises from stochastic polymerization governed by monomer reactivity ratios Smith et al. [2018], Yu et al. [2024]. Previous studies have typically represented RHPs using one-hot monomer composition vec-

tors Tamasi et al. [2022], Wu et al. [2023], implicitly assuming additive behavior between monomers. In contrast, blending multiple RHPs effectively decouples the relationship between 1-mer composition and higher-order k-mer statistics: mixing different RHPs alters the distribution of segmental motifs without changing the overall monomer composition (Figure 3B) Jiang et al. [2020], Ruan et al. [2023]. In this k-mer feature space, a blend corresponds to a convex combination of its constituent RHP representations (Figure 3C), introducing an additional continuous degree of freedom for engineering emergent, non-additive properties and enabling data-driven exploration of disordered polymer ensembles where each blend can be viewed as a weighted mixture of learned one-hot embeddings that encode both composition and local sequence correlations.

The k-mer distribution of the RHP with different k were generate with quantecon and scipy package.

```python
import numpy as np
import quantecon as qe
from scipy.stats import multinomial
def calculate_kmer(p:np.array, k:int):

    distribution = []
    index = np.argmax(p)
    p_rest = np.delete(p,index)
    rest_sum = np.sum(p_rest)
    p[index] = 1 - rest_sum
    for i in range(k):
        kmer = qe.simplex_grid(p.shape[0],i+1)
        kmer_distribution = multinomial.pmf(x=kmer,n=i+1,p=p)
        distribution = np.append(distribution,kmer_distribution)

    return distribution
```

2.3 Predictive Modeling and SHAP Analysis

Predictive modeling and feature attribution analyses were performed in Python using the XGBoost library for regression and the SHAP package for model interpretation. An XGBoost model was trained on the dataset, with a 9:1 train–test split. Training was repeated three times with different random seeds, and model hyperparameters were optimized for each run. Model performance was evaluated using the coefficient of determination (R^2) and mean squared error (MSE).

For interpretability, SHAP values were computed on the model trained with the 4-mer representation to quantify the contribution of each segment feature to the predicted REA. SHAP summary plots and interaction values were analyzed to identify the most influential k-mer motifs and synergistic relationships among polymer segments driving functional performance.

3 Results

We performed analysis of the experimental outcomes to identify sequence-level features contributing to enzyme stabilization. Each RHPB was represented by a concatenated k-mer fingerprint that encodes segment-level monomer connectivity at different lengths (k) (Figure 4A). Using XGBoost regression, predictive accuracy improved with increasing k. When k = 1, the model reproduced the same monomer-level trends (M1, M2, M3) previously observed by Tamasi et al. Tamasi et al. [2022], while for $k > 2$ the k-mer model outperformed one-hot encoding based solely on blending ratios (Figure 4B). The best performance was achieved at $k = 4$, which was used for subsequent analyses. A model trained on the data accurately predicted the outcomes of an independent Bayesian-optimization campaign from selected RHPs (Figure 4C), suggesting that the learned k-mer features captured the key structural factors governing thermal stabilization.

Feature attribution using SHAP (SHapley Additive exPlanations)Lundberg and Lee [2017] revealed that specific k-mers strongly correlate with model performance (Figure 4D). The motif $(M1)_3(M4)_1$ exhibited the highest positive contribution, while SHAP interaction analysisLundberg et al. [2020] showed that the presence of $(M1)_2(M4)_1$ synergizes with $(M2)_1$ but not with $(M1)_3$ (Figure 4E). These

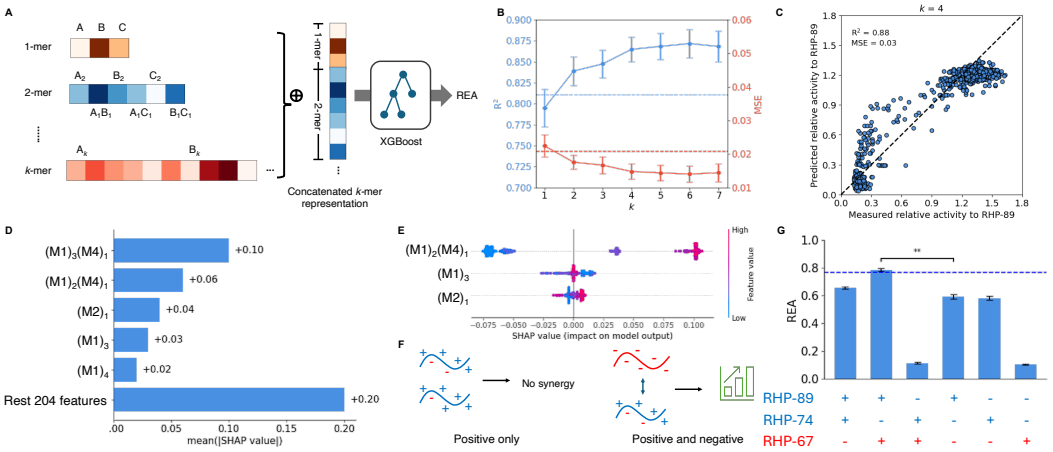


Figure 4: Modeling the experimental results of RHPBs with k-mer representation. (A) Schematic illustration of the generation of k-mer representation. RHPs and RHPBs are encoded as concatenated k-mer segments and used as input features for XGBoost to predict the REA. (B) Model performance in terms of R^2 (blue) and mean squared error (MSE; red) as a function of k, demonstrating improved predictive accuracy with higher-order k-mer representations. Each data point is represented as mean $\pm$ standard deviation of 3 different train-test splits. The dashed blue and red lines indicate the R^2 and MSE values obtained using one-hot encoding based on the blending ratios of RHPs. (C) Parity plot of relative REA values to RHP-89 of RHPBs from another optimization campaign with RHPs selected from the library, predicted by the model trained on data from original dataset with concatenated 4-mer representation. The results demonstrate the model's ability to capture key features driving performance improvements. (D) SHAP analysis ranking the top-5 segment features from 4-mer representations contributing to the model performance. (E) SHAP interaction values for the interactions between selected segment features and $(M1)_2(M4)_1$ from model trained with 4-mer representation. The x-axis is the combined effect of these features on the model's output. The dot colors reflect the magnitude of one of the interacting features. The results indicate that $(M1)_2(M4)_1$ itself has a strong contribution to the improvement. Meanwhile, its presence together with specific occurrences of $(M2)_1$ also contributes to improved performance. In contrast, combinations of $(M1)_2(M4)_1$ with $(M1)_3$ exhibit an negligible effect on the prediction. (F) Illustration of the synergistic effect. RHPBs composed solely of positively charged components did not show synergy (left), whereas incorporating both positive and negative components (right) leads to synergistic improvement. (G) Ablation experiment of a selected RHPB demonstrating that both the specific identity of the positively charged polymer (RHP-89 and RHP-74) and the inclusion of oppositely charged components (RHP-67) synergistically enhance performance.

results indicate that separating positively and negatively charged segments into distinct polymer chains enhances GOx stabilization, and that the balance of charged RHP components must complement the protein surface (Figure 4F).

Ablation experiments on a top-performing blend (RHP-89:RHP-74:RHP-67:RHP-78 = 0.4:0.15:0.3:0.15) confirmed these findings (Figure 4G). Removing individual components revealed strong synergy between negatively charged RHP-67 and positively charged RHP-89, whereas replacing RHP-89 with another positive polymer (RHP-74) diminished activity. These observations demonstrate that optimizing segment-level charge distributions through blending yields benefits beyond simple monomer-composition tuning.

This phenomenon mirrors mechanisms of molecular recognition by intrinsically disordered regions (IDRs) in proteins, which use short linear motifs (SLiMs) of 5–12 residues to mediate specific yet flexible interactions Holehouse and Kragelund [2024]. The redundancy and tolerance of SLiMs to sequence variation parallels the statistical diversity of RHP sequences. Exploring RHPB space thus not only enables materials discovery but also illuminates how local segment motifs contribute to macroscopic function. Our analysis identified the co-occurrence of M1 and M4 within short segments as critical for GOx stabilization, likely reflecting complementarity to charged patches on the protein surface Panganiban et al. [2018]. All-atom explicit-solvent molecular-dynamics

simulations of GOx–RHPB complexes (data not shown) confirmed that multiple polymer chains interact simultaneously with a single enzyme, consistent with previous adsorption models Jin et al. [2023, 2025]. The observed synergy between oppositely charged segments likely arises from complementary electrostatic interactions at the polymer–protein interface that collectively enhance structural stabilization.

4 Conclusion

In summary, we developed a physics-grounded k-mer representation learning framework for modeling the functional performance of RHPBs. By encoding segment-level connectivity statistics derived from stochastic polymerization theory, the model captures non-additive sequence effects that govern emergent functionality and enables accurate prediction of enzyme stabilization outcomes across experimental campaigns. SHAP analysis revealed combinations of segment motifs responsible for synergistic performance, highlighting the role of charge complementarity and short-range monomer correlations in protein–polymer interactions.

Despite its predictive performance, the current representation does not explicitly capture molecular-weight distributions or chain-length heterogeneity, both of which can substantially influence polymer conformations and blending behavior. Incorporating these structural descriptors, together with coarse-grained physical parameters, would further enhance the model's quantitative accuracy and transferability. However, direct simulation of such large, disordered systems using conventional molecular dynamics (MD) remains computationally expensive. Looking ahead, we anticipate that accelerated MD techniques will enable mechanistic insight into how local sequence statistics give rise to emergent macroscopic functions. When integrated with data-driven modeling, these approaches will contribute to a unified and physically interpretable framework for designing and understanding complex polymer ensembles.

References

Ibrahim Khan, Muhammad Mansha, and Mohammad Abu Jafar Mazumder. *Polymer Blends*, pages 513–549. (Springer), 2019. doi: 10.1007/978-3-319-95987-0_16.

Leszek A. Utracki, P. Mukhopadhyay, and R. K. Gupta. *Polymer Blends: Introduction*, pages 3–170. (Springer Dordrecht), 2014.

Ting-Wei Lin, Omar Padilla-Vélez, Parin Kaewdeewong, Anne M LaPointe, Geoffrey W Coates, and James M Eagan. Advances in nonreactive polymer compatibilizers for commodity polyolefin blends. *Chemical Reviews*, 124(16):9609–9632, 2024.

Michael Patrick Blatt and Daniel T Hallinan Jr. Polymer blend electrolytes for batteries and beyond. *Industrial & Engineering Chemistry Research*, 60(48):17303–17327, 2021.

Michael C Leyden, Felipe Oviedo, Sonashree Saxena, Ramya Kumar, Ngoc Le, and Theresa M Reineke. Synergistic polymer blending informs efficient terpolymer design and machine learning discerns performance trends for pdna delivery. *Bioconjugate Chemistry*, 35(7):897–911, 2024.

Jeffrey Ethier, Evan R. Antoniuk, and Blair Brettmann. Predicting polymer solubility from phase diagrams to compatibility: a perspective on challenges and opportunities. *Soft Matter*, 20:5652–5669, 2024. doi: 10.1039/D4SM00590B.

Zhilong Liang, Zhiwei Li, Shuo Zhou, Yiwen Sun, Jinying Yuan, and Changshui Zhang. Machine-learning exploration of polymer compatibility. *Cell Reports Physical Science*, 3:100931, 2022. doi: 10.1016/j.xcrp.2022.100931.

Phillip EC Compeau, Pavel A Pevzner, and Glenn Tesler. How to apply de bruijn graphs to genome assembly. *Nature biotechnology*, 29(11):987–991, 2011.

Anton AA Smith, Aaron Hall, Vincent Wu, and Ting Xu. Practical prediction of heteropolymer composition and drift. *ACS Macro Letters*, 8(1):36–40, 2018.

Hao Yu, Luofu Liu, Ruilin Yin, Ivan Jayapurna, Rui Wang, and Ting Xu. Mapping composition evolution through synthesis, purification, and depolymerization of random heteropolymers. *Journal of the American Chemical Society*, 146(9):6178–6188, 2024.

Matthew J Tamasi, Roshan A Patel, Carlos H Borca, Shashank Kosuri, Heloise Mugnier, Rahul Upadhya, N Sanjeeva Murthy, Michael A Webb, and Adam J Gormley. Machine learning on a robotic platform for the design of polymer–protein hybrids. *Advanced Materials*, 34(30):2201809, 2022. doi: 10.1002/adma.202201809.

Guangqi Wu, Haisen Zhou, Jun Zhang, Zi-You Tian, Xingyi Liu, Shuo Wang, Connor W Coley, and Hua Lu. A high-throughput platform for efficient exploration of functional polypeptide chemical space. *Nature Synthesis*, 2(6):515–526, 2023.

Tao Jiang, Aaron Hall, Marco Eres, Zahra Hemmatian, Baofu Qiao, Yun Zhou, Zhiyuan Ruan, Andrew D Couse, William T Heller, Haiyan Huang, et al. Single-chain heteropolymers transport protons selectively and rapidly. *Nature*, 577(7789):216–220, 2020.

Zhiyuan Ruan, Shuni Li, Alexandra Grigoropoulos, Hossein Amiri, Shayna L Hilburg, Haotian Chen, Ivan Jayapurna, Tao Jiang, Zhaoyi Gu, Alfredo Alexander-Katz, et al. Population-based heteropolymer design to mimic protein mixtures. *Nature*, 615(7951):251–258, 2023.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

Alex S Holehouse and Birthe B Kragelund. The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology*, 25(3):187–211, 2024.

Brian Panganiban, Baofu Qiao, Tao Jiang, Christopher DelRe, Mona M Obadia, Trung Dac Nguyen, Anton AA Smith, Aaron Hall, Izaac Sit, Marquise G Crosby, et al. Random heteropolymers preserve protein function in foreign environments. *Science*, 359(6381):1239–1243, 2018.

Tianyi Jin, Connor W Coley, and Alfredo Alexander-Katz. Adsorption of biomimetic amphiphilic heteropolymers onto graphene and its derivatives. *Macromolecules*, 56(5):1798–1809, 2023.

Tianyi Jin, Connor W Coley, and Alfredo Alexander-Katz. Sequence-sensitivity in functional synthetic polymer properties. *Angewandte Chemie International Edition*, 64(2):e202415047, 2025. doi: 10.1002/anie.202415047.