Machine Learning on a Robotic Platform for the Design of Polymer-Protein Hybrids

Matthew J. Tamasi[1†], Roshan A. Patel[2†], Carlos H. Borca[2†], Shashank Kosuri[1†], Heloise Mugnier[1], Rahul Upadhya[1], N. Sanjeeva Murthy[1], Michael A. Webb[2*], and Adam J. Gormley[1*]

[1]Department of Biomedical Engineering, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA
[2]Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey 08544, USA

Polymer-protein hybrids are intriguing materials that can bolster protein stability in non-native environments, thereby enhancing their utility in diverse medicinal, commercial, and industrial applications. One stabilization strategy involves designing synthetic random copolymers with compositions attuned to the protein surface, but rational design is complicated by a vast chemical and composition space. Here, we report a strategy to design protein-stabilizing copolymers based on active machine learning, facilitated by automated material synthesis and characterization platforms. The versatility and robustness of the approach is demonstrated by the successful identification of copolymers that preserve, or even enhance, the activity of three chemically distinct enzymes following exposure to thermal denaturing conditions. Although systematic screening results in mixed success, active learning appropriately identifies unique chemistries for each enzyme. Overall, this work broadens our capabilities to design fit-for-purpose synthetic copolymers that promote or otherwise manipulate protein activity, with extensions towards the design of robust polymer-protein hybrid materials.

Polymer-protein hybrids (PPHs) have emerged as attractive materials that leverage polymers to improve protein solubility and stability in often denaturing and abiological environments.[2–6] One strategy, which has resulted in remarkable hours-long enzyme activity in toluene,[7] tailors the composition of random copolymers based on protein surface chemistry. In principle, copolymers might be precisely designed to stabilize any given protein without compromising activity. However, identifying such copolymers, whether via rational design or screening, is challenging due to a large combinatorial design space (e.g., monomer chemistry, chain length, architecture).[8] Thus, fit-for-purpose PPHs could facilitate myriad applications– biofuel production,[9] plastics degradation,[10,11] pharmaceutical synthesis[12]– but a robust strategy for their design remains elusive.

Over the last decade, machine learning (ML) has dramatically accelerated materials discovery across disciplines,[13–15] enabling more efficient identification of materials with target properties.[13,16–21] Nonetheless, ML-guided copolymer design is limited by several factors, including the availability of quality data necessary to train the underlying models.[8,22–25] Most polymer databases predominantly feature homopolymers,[26] and the laborious nature of polymer synthesis and characterization severely limits the number of systems that can be examined "in-house".[27] Several copolymer design efforts have thus relied on data generated *in silico*.[21,28,29] Meanwhile, recent experimental work has used flow reactors or parallel batch synthesizers to provide modest data (< 500 samples).[18,30,31] More scalable approaches would substantially extend capabilities to design copolymers for PPHs and other materials applications.

Here, we use active ML to rapidly design copolymers to form thermostable PPHs with glucose oxidase (GOx), lipase (Lip), and horseradish peroxidase (HRP) (Fig. 1). To efficiently acquire data, we use automated oxygen-tolerant radical polymerization for copolymer synthesis[32,33] and develop a facile, thermal-stability assay to characterize PPHs. With this platform and five iterations of active learning for each enzyme, we successfully identify PPHs with significant enzyme activity; these PPHs generally outperform those derived from a systematic screen with over 500 unique copolymers. Notably, we demonstrate that our strategy appropriately adapts data acquisition to yield chemically distinct sets of top-performing copolymers for each enzyme. *Post hoc* analysis of our data and ML models reveals important relationships between specific copolymer chemistries and PPH stability, while biophysical characterization of our most efficacious PPHs provide mechanistic insight into how copolymers may preserve enzyme function under thermal stress. Overall, this framework will automate and accelerate the design of copolymers for stable PPHs across applications.

Overview of design space and strategy

To test our ML-based design paradigm, we consider three chemically distinct enzymes–HRP, GOx, and Lip–with the design goal to maximize retained enzyme activity (REA) following thermal stressing. For reference, a PPH exhibiting 100% REA provides the same level of activity

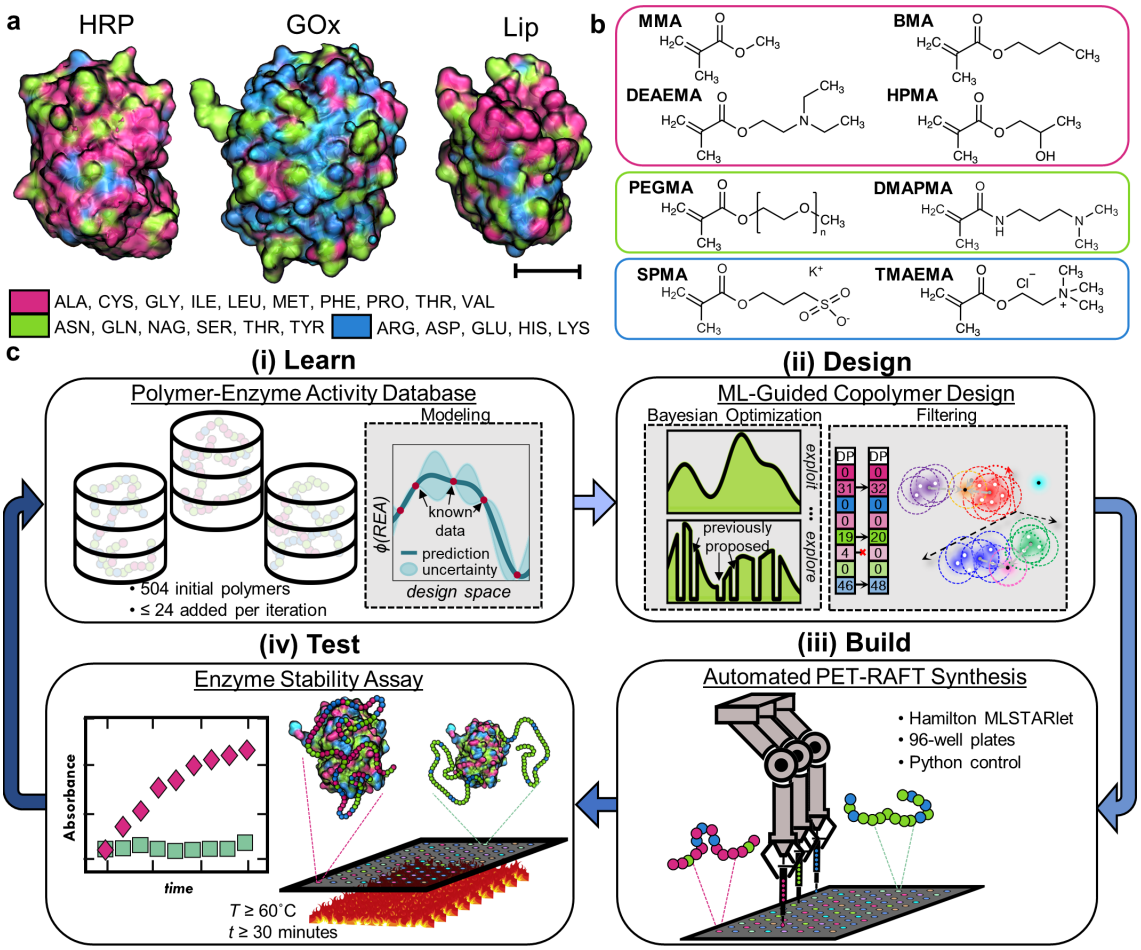

Fig. 1| Overview of study. a, Schematic illustration of the surface chemistry for horseradish peroxidase (HRP), glucose oxidase (GOx), and lipase (Lip). Amino acids are colored based on classification as hydrophobic (magenta), hydrophilic (green), or ionic (blue). Scale bar = 2 nm. Images for the protein are rendered using Visual Molecular Dynamics.[1] **b,** Monomers utilized for copolymer design. The colored boxes delineate rough classifications as hydrophobic (magenta), hydrophilic (green), and ionic (blue). **c,** Schematic representation of closed-loop learn-design-build-test discovery process used in this work. After initialization with a seed dataset, the process consists of (i) training an enzyme-specific Gaussian process regression (GPR) surrogate model to predict the REA of a PPH based on copolymer charachteristics (learn), (ii) Bayesian optimization of copolymers to satisfy an expected improvement acquisition function and subsequent filtering to propose new copolymers (design), (iii) automated synthesis of proposed copolymers via photoinduced electron/energy transfer reversible addition–fragmentation chain transfer (PET-RAFT) polymerization (build), and (iv) mixing of synthesized copolymers with enzyme to form PPHs that are thermally stressed and assessed for REA (test). Newly acquired data can then be used to restart the closed-loop discovery process.

as the enzyme prior to thermal stressing. Because these enzymes possess distinct surface chemistries and molecular weights (Fig. 1a), we consider a copolymer design space with eight possible monomers (Fig. 1b) copolymerized with target degree of polymerization (DP) between 50 and 200 in increments of 25. The chosen monomers are classified as hydrophobic (DEAMA, HPMA, BMA, MMA), hydrophilic (DMAPMA, PEGMA), or ionic (SPMA, TMAEMA); this set enables various interactions (e.g., van der Waals, hydrogen-bonding, electrostatic) with the enzyme, while balancing aqueous solubility. To encourage reproducible synthesis and minimize latency, up to four distinct monomers are selected for copolymerization.

Fig. 1c schematically presents the Learn-Design-Build-Test cycle employed here. After constructing an initial seed dataset featuring 504 copolymers and corresponding REA measurements, we performed five iterations for each enzyme. Within each iteration, we *(i)* developed ML models to predict REA from copolymer characteristics, *(ii)* identified batches of 24 candidate copolymers for PPHs using active and unsupervised ML, *(iii)* synthesized candidate copolymers, and *(iv)* performed thermal activity assays to determine REA for candidate PPHs; these results augmented the dataset to begin the next iteration.

Inefficiency of screening

To gain perspective on the viability of brute-force search, our seed dataset consisted of a systematic screen over 504 copolymers with distinct monomer combinations and DPs. The vast majority of copolymers in this dataset did not result in substantial REA, with the median values of 3.2% (HRP), 10.0% (GOx), and 0.118% (Lip). These poor results are partly explained by the limited design space surveyed during systematic screening (Fig. S1, S2). Additionally, the REA for PPHs with Lip, HRP, and GOx vary significantly for copolymers in the seed dataset, suggesting that copolymers should be tuned to specific enzymes and that systematic screening is likely to have mixed success across different enzymes.

Active learning in a combinatorial design space

To guide data acquisition beyond the seed database, we devised an active learning (AL) paradigm based on Bayesian optimization (BO)[34] of a ML surrogate model (see Methods). Preliminary comparisons using the seed datasets indidcated that GPR modeling with simple, machine-readable copolymer representations as input provided the best predictive performance and was thus selected as our surrogate modeling approach over other ML algorithms and copolymer featurization strategies[35] (Fig. S3). At early stages of the design process, our objective was to iteratively identify batches of copolymers that are likely to exhibit improvements in REA according to our ML models and/or explore unknown regions of design space based on model uncertainty. To achieve this balance between exploitation and exploration, we optimized copolymer compositions and DP according to a series of modified expected improvement acquisition functions (see Methods, Candidate copolymer generation, Candidate copolymer down-selection); similar acquisition functions have been used in previous work related to polymer design.[36,37] Following four iterations of this data acquisition approach, we transitioned to a policy of pure exploitation in the fifth iteration; we refer to the fifth iteration as the "exploit round."

Fig. 2a-c shows that the AL-BO paradigm facilitated identification of numerous, diverse copolymers that enhanced retained activity for each of the three enzymes. The median REA of PPHs found in the intermediate and final iterations of AL show progressive and significant increase over those in the seed database. In particular, there is a difference of 46.2%, 31.5 %, and 87.6% between the median REA of seed PPHs and those found in the exploit round for HRP, GOx, and Lip, respectively. Even within the intermediate iterations (1-4), we typically find improvements in median REA iteration-over-iteration (Fig. S4), despite data acquisition sometimes foregoing potentially promising designs in favor of diversity or uncertainty. For Lip and GOx, the best PPHs are found within the exploit round and exhibit remarkable REA values of 107.9% and 67.4%, which significantly improve upon both the average and maximum values observed in the seed datasets. For HRP, the top-performing PPH is found during the initial screen, but many of the top hybrids are still identified by AL, including one with an REA of 81.0%. More generally, we find that a large number of diverse copolymers offer reasonable stabilization of HRP, and AL identifies some promising regions that are not exposed by our systematic search. Quantitatively, AL-guided copolymers are disproportionately represented as top performers, comprising 70.2%, 40.5%, and 42.5% of the top twentieth percentile of REA for Lip, GOx, and HRP, respectively. Interestingly, the exploit round also produces three PPHs for Lip that not only preserve but enhance its activity relative to the unstressed enzyme.

Fig. 2d-i examine both the progression of AL and PPH performance as a function of the chemical constitution of copolymers. Based on the totality of measured REA values, we find that best-performing PPHs for each enzyme utilize entirely different copolymer chemistries, which justifies a tailored design strategy. In particular, optimal copolymers for HRP stabilization predominantly feature hydrophobic and ionic monomers and smaller DP ($<$100) (Fig. 2a,d). While AL-generated candidates primarily focus on uncovering this region of the chemical space, there are also many effective PPHs that limit ionic content as identified by the seed dataset (Figs. 2g and S2c). In this case, a wide range of diverse, high-performing PPHs are identified by AL, despite outlier points in the HRP dataset (Table S1). For GOx, optimal copolymers are either predominantly hydrophobic or hydrophilic with very little ionicity and have DP typically in the range of 100-150 (Fig. 2b,e). Accordingly, AL for GOx stabilization predominantly probed these regions of the chemical space and remained globally stagnant in its search (Fig. 2e,h), fine-tuning relatively promising regions identified in the seed dataset (Fig. S2a). Conversely, optimal copolymers for Lip stabilization possess sizable incorporations of monomers from all three chemical groupings with generally larger DP (Fig. 2c,f). AL-proposed candidates progress towards this promising region of the chemical space with each subsequent iteration (Fig. 2f,i); notably, this region of the chemical space is completely avoided in the seed dataset (Fig. S2b). This suggests that the Lip design campaign benefited from both exploration and exploitation data acquisition polices. Therefore, the AL/BO paradigm appropriately adapted optimization to identify high-performing PPHs for each enzyme across chemical space, with less than 20% additional data beyond the initial systematic screen.

Understanding chemical features driving PPH performance

Given the identification of highly stable PPHs for each enzyme, we sought to understand the important chemical

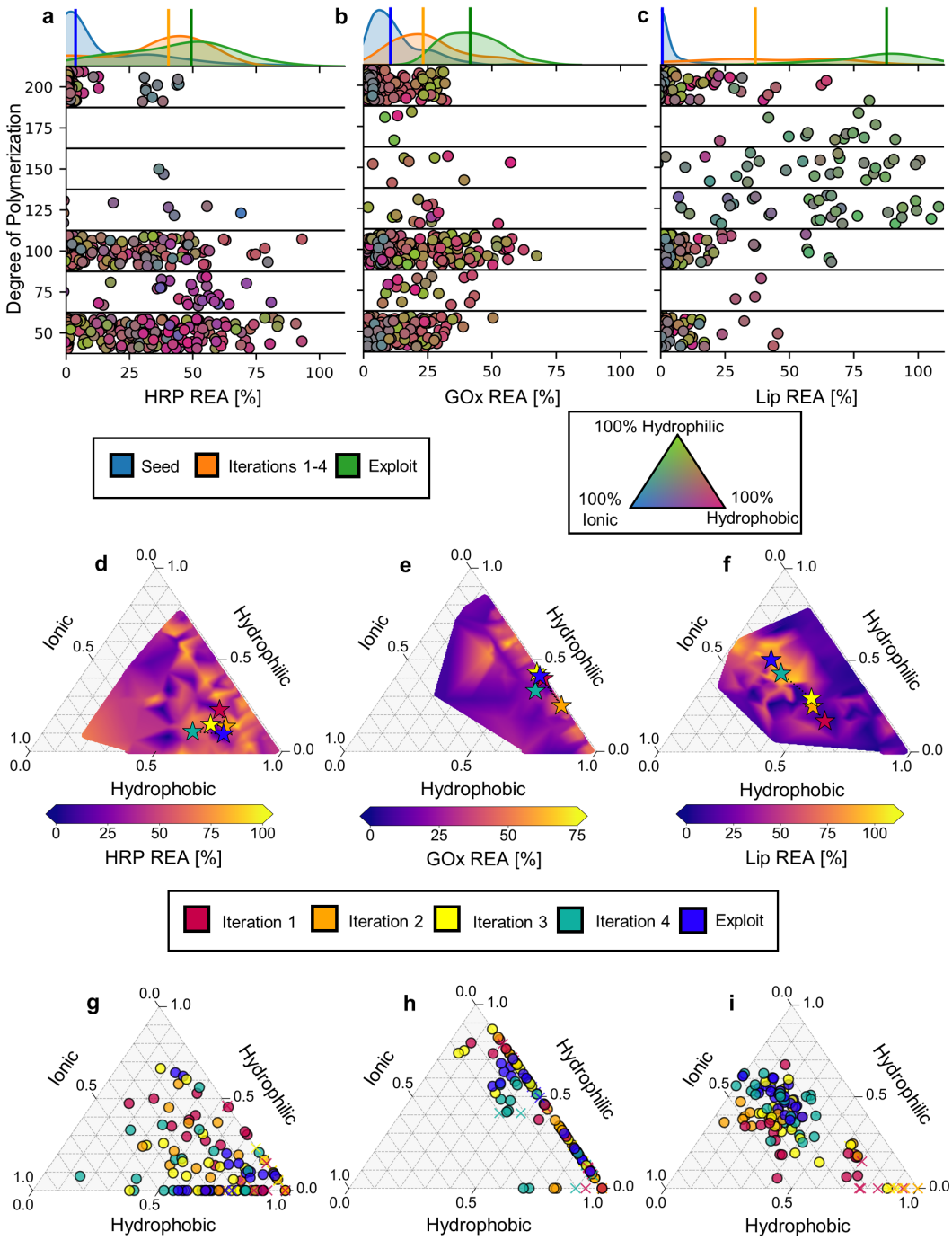

Fig. 2| Machine learning guides design of highly stable polymer-protein hybrids. a-c, Copolymer designs and their measured REAs for HRP, GOx, and Lip. Marginal axes at the top contain Gaussian kernel density estimate distributions of REA in the seed database (blue), active learning iterations 1-4 (orange), and the final exploitation round (green). Medians of distributions are indicated by vertical lines. Main axes show the experimentally measured REA for all tested PPHs; individual markers are vertically located in bins according their degree of polymerization with random fluctuations added within bins to improve visual clarity. The marker color reflects the composition of the copolymer according to the ternary diagram (bottom right). **d-f,** Representation of active learning path traversed through copolymer chemical space for each enzymes. The chemical space is represented as a ternary diagram with coordinates providing the fraction of incorporation of hydrophobic, hydrophilic, and ionic monomers in copolymers. Colored stars indicate the mean composition of copolymers proposed during a given active learning iteration. The ternary diagrams are additionally colored by maximum REA observed for a PPH in a given region of the chemical space spanned by the ternary axes. **g-i,** Individual chemical compositions of copolymers proposed during each stage of active learning. The centroid of all points at a given iteration yields the position of the stars **d-f**. The crosses denote copolymers that showed undesirable gelation during synthesis (see Methods, Handling polymer gelation).

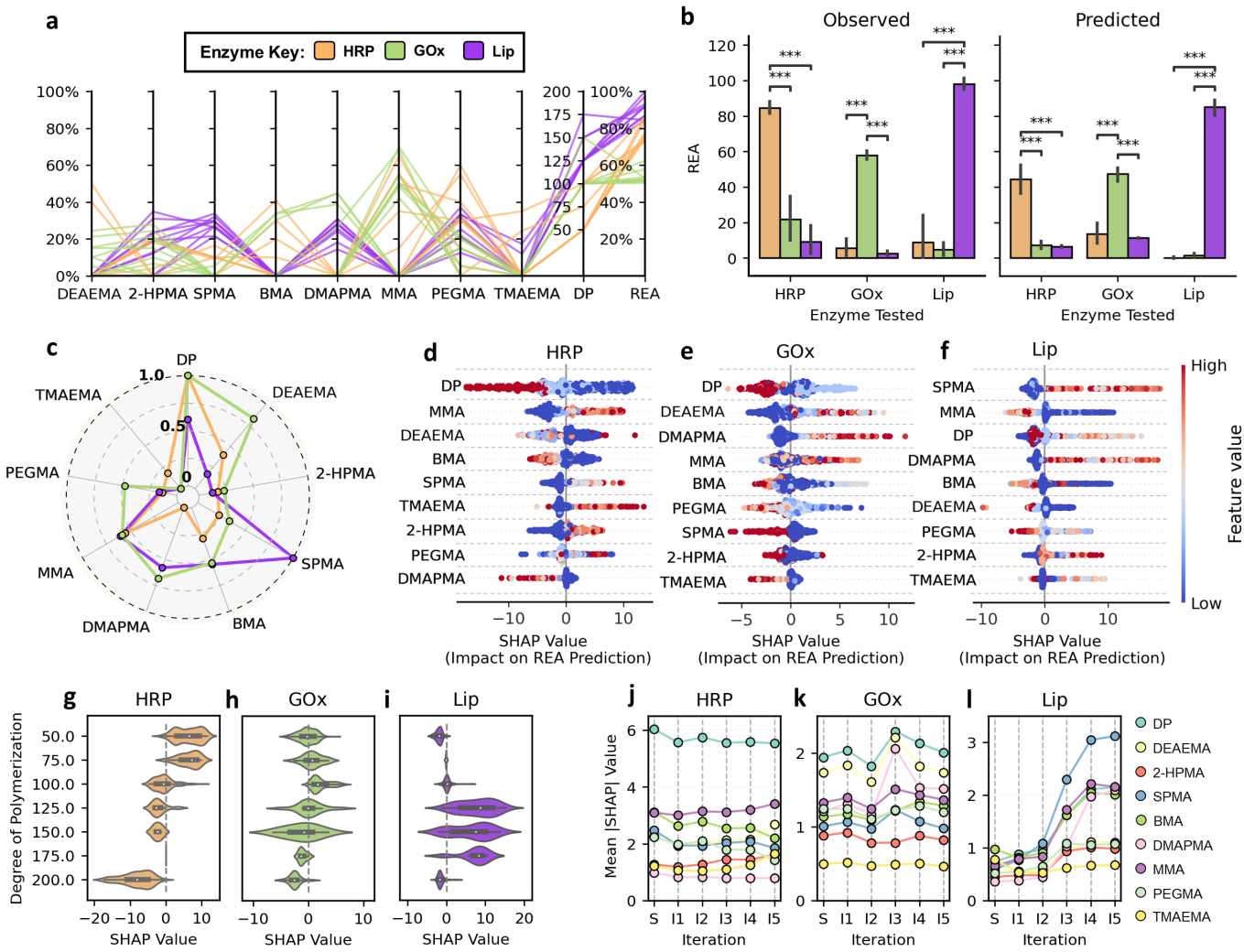

Fig. 3| Analysis reveals distinct priorities in copolymer features for each protein. a, Copolymer compositions and degree of polymerization (DP) for the top ten performing PPHs for HRP (orange), GOx (green), and Lip (purple). **b,** Cross-evaluation of top-performing copolymers across enzymes showing mean observed and predicted REA for each copolymer-enzyme pairing. Statistical significance was determined by Mann-Whitney U test. * ($p<0.05$), ** ($p<0.005$), *** ($p<0.0005$), unlabeled pairs are not significantly different. Top 10 performers for each enzyme demonstrate high specificity in agreement with predicted activity. **c,** Normalized mean |SHAP| values calculated for HRP, GOx, and Lip for each model to quantify relative feature importance. **d-f,** SHAP summary values for GPR models calculated from available data after all five active learning iterations. Each point corresponds to a unique evaluated PPH, and the point's position along the X-axis shows the impact of a feature on predicted REA. **g-i,** SHAP value distributions demonstrating the effect of degree of polymerization on REA predictions. Polymer chain lengths with maximum calculated SHAP values are distinct between enzymes. Black candlesticks range from second to third quartiles of SHAP values and white dots represent the distribution mean. **j-l,** Mean |SHAP| values calculated for all model features after model training on the seed dataset and after each iteration of active learning.

features of copolymers that gave rise to their performance. Fig. 3a compares the features of copolymers underlying PPHs with the top ten highest REA for each enzyme. Although top-performing PPHs for a given enzyme tend to have some chemical similarity across effective copolymers, there is substantial chemical diversity between PPHs for different enzymes. To demonstrate that copolymer pairing with enzymes is highly specific, we cross-examined the efficacy of the copolymers in Fig. 3a to stabilize the other enzymes; the results are provided in Fig. 3b. Experimentally, we empirically confirmed that the REA of PPHs designed for a specific enzyme are significantly higher than that of PPHs formed by the same copolymers but other enzymes. Virtual cross-evaluation using enzyme-specific GPR models trained on all iterations of data similarly suggest that REA is significantly diminished when top-performing copolymers for one enzyme are paired with another. Together, these results not only suggest an intricate connection between copolymer chemistry and size and the stability of PPHs but that such correlations can be effec-

tively learned from data.

To further explore the relationship between copolymer features and PPH activity, we computed Shapley additive explanations (SHAP) values[38,39] to quantify how chemical features of the copolymers (fractions of incorporation and DP) contributes to REA predictions by our GPR models. Here, positive SHAP values indicate positive contributions REA (negative SHAP values suggest negative contributions), and we use the mean absolute SHAP value of a feature as a proxy for its overall importance to model prediction. Fig. 3c shows that different copolymer features have distinct impact on REA predictions. To elucidate these differences, we compare SHAP values for the fractions of incorporation for each monomer (Fig. 3d-f) and DP (Fig. 3g-i) for each enzyme. Although we previously associated hydrophobic chemistry with high-performing PPHs for HRP (Fig. 2f,i), Fig. 3d reveals that the *exclusion* of BMA is favorable (higher REA), while the *inclusion* of MMA, a similar hydrophobic monomer, is associated with higher REA. Similar observations can be readily identified for Lip (Fig. 3f), for which SPMA and TMAEMA monomers (both highly ionic) represent the most and least important features based on their mean absolute SHAP values. Such differences in SHAP values between monomers with the same chemical classifications underscores the intricacy of designing effective polymer-enzyme pairing.

Fig. 3c-i also indicate that the relative importances of copolymer features varies across enzyme models. For example, we find that different chain length regimes favor high predictions on REA, depending on the enzyme-specific GPR model. (Fig. 3g-i). For HRP, smaller polymers (DP = 50, 75) display the highest SHAP values, while the highest SHAP values for Lip are observed for DP = 125 or 150. DP = 200 is generally associated with lower REA, perhaps suggesting that shorter copolymer sequences enable more facile pairing with enzyme chemical domains to promote stabilization.

To understand the evolution of feature importances during AL, we compared mean absolute SHAP values for all non-gelling copolymers derived from GPR models trained after each stage of data acquisition. Fig. 3j-l shows that the importance of features can shift significantly, even with the addition of small amounts of data (typically 20 data points added per iteration or less than 4% increase in prior data available). This is most evident following for Lip, wherein mean absolute SHAP values for SPMA, MMA, DMAPMA, and DP all substantially increase after the third and fourth iteration. This behavior might be related to data acquisition over previously unexplored regions of chemical space, which is partly shown in Fig. 2e. The effects for HRP and GOx are overall less dramatic; most rankings are unchanged between iterations, with occasional shifts of one or two ranks upon exposure to new data. Nonetheless, even if the rank-ordering of features is unchanged, mean improvement in measured REA for PPHs across iterations suggests that GPR models had sufficient fidelity to effectively optimize REA, at least within a local chemical space.

Revealing mechanisms with biophysical characterization

Although mechanisms of stabilization for PPHs based on random copolymers have been hypothesized and studied in limited fashion using molecular dynamics simulation,[7] experimental examination of these biophysical interactions is nascent. Therefore, we characterized (Fig. S5) and investigated a particular PPH for HRP identified in the exploit round–dubbed HRP-Exploit Polymer 1 (HRP-EP1)– using circular dichroism (CD) spectroscopy, small-angle X-ray scattering (SAXS), dynamic light scattering (DLS), and quartz crystal microbalance with dissipation (QCM-D). HRP was selected due to its amenability to these characterization techniques, while detailed characterization of other enzyme systems proved challenging due to weak CD spectroscopy signal-to-noise and solubility limitations. We first investigated the impact of heating and cooling on the secondary structure of HRP by CD spectroscopy (Fig. 4a). The corresponding measured α-helix, β-sheet, and random coil content is provided in Table S2. We initially hypothesized that the addition of copolymer EP1 would reduce thermally induced unfolding of HRP; however, the CD data suggests only a slight retardation of unfolding. Upon heating, the α-helix content for HRP degrades from ca. 34.8% to 17.4%, while the α-helix content for the HRP-EP1 system is 20.3% after heating. However, following cooling, HRP-EP1 exhibited 31.6% α-helix content compared to just 24.6% for HRP alone. This suggests that EP1 facilitates significant refolding of HRP in a chaperone-like manner.

To further understand the nature of the HRP-EP1 interactions, we used SAXS to compare the physical dimensions of HRP and its complexes in pre- and post-stress states. Guinier analysis of the data (Table S3, Fig. S6) showed that both HRP and HRP-EP1 have the same radius of gyration (R_g, 24.6 - 25.0 Å) in the pre-stressed state. Similarly, in the pre-stressed state, the pair-distance distribution function $P(r)$ remains highly similar upon complexation of HRP with EP1 (Fig. 4b). Post-stress, the differences are dramatic in the pair-distance distribution function. While the maximum particle diameter ($Dmax$) of native HRP increases from 80 to 200 Å, that of HRP-EP1 increased only to 94 Å (Table S3). Additionally, while the R_g of HRP-EP1 increases only slightly to 26.9 Å, a larger 51.9 Å component appears in the Guinier plots of HRP (Fig. 4c, blue line), likely indicative of a denatured or aggregated sub-species of HRP created through thermal stress. Additionally, Kratky plots (Fig. S7) show peaks at q = 0.065 and 0.075 Å^{-1} in HRP and HRP-EP1, respectively, which indicates a compact structure similar to that of the native protein. This clearly suggests that the com-

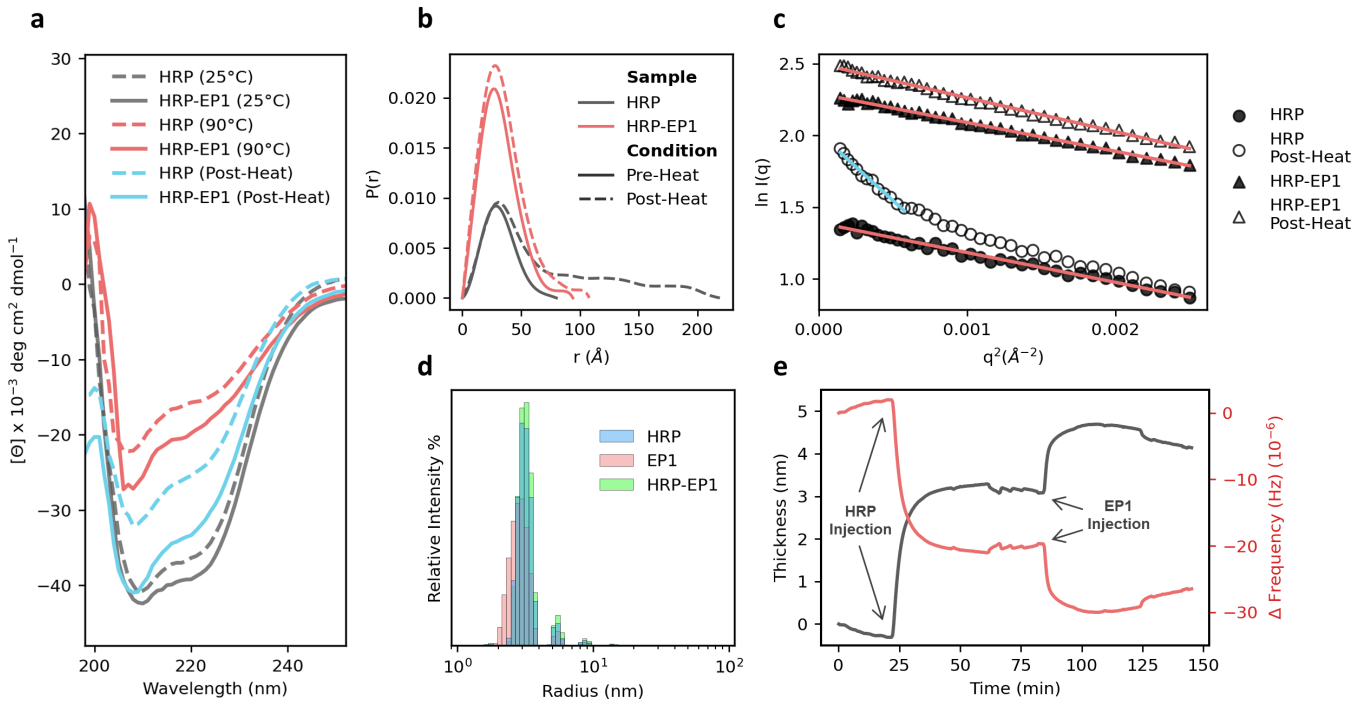

Fig. 4| Biophysical characterization indicates copolymer-assisted refolding. **a,** Circular dichroism wavelength scans of HRP (dashed lines) and HRP-EP1 (solid lines) at room temperature (black), upon heating (red), and after cooling for 24hrs (blue), demonstrating that HRP-EP1 promotes retention of secondary structure in HRP during thermal stress and promotes significant protein refolding in comparison to HRP control. **b,** Pair-distance distribution function of HRP and HRP-EP1 by small-angle X-ray scattering demonstrating retained HRP-PPH morphology and size after exposure to thermal stress in comparison to native enzyme. **c,** Guinier analysis of HRP and HRP-EP1 before and after heating suggesting the development of a denatured or aggregated sub-population of HRP (blue line) in comparison to a single species observed in HRP, HRP-EP1, and HRP-EP1 after thermal stress (red lines). **d,** Dynamic light scattering size distributions of HRP with and without polymer EP1, demonstrating that no larger structures were observed after mixing. **e,** Surface thickness measured by Quartz crystal microbalance with dissipation after direct adsorption of HRP ($t = 22$ min) followed by injection of polymer EP1 ($t = 82$ min).

plex promotes a certain level of conformational integrity in HRP even if secondary structure is impacted.

Finally, DLS was performed to complement the SAXS results by providing the distribution of hydrodynamic radii (R_h) in the samples (Fig. 4d). All samples show peak intensities between 3.0 - 3.3 nm with minimal signal intensity for $R_h > 10$ nm. Additionally, measured polydispersity index remained under 0.2 for all samples, suggesting relatively monodisperse solutions (Fig. S8, Table S4). These results indicate that stabilization of HRP in PPH-EP1 is indeed driven by the formation of a complex rather than via larger macromolecular assembly. Further support of complex formation by QCM-D showed significant differences in the Sauerbrey mass thickness following injection of EP1 onto surface immobilized HRP (Figs. 4e and S9). While native HRP exhibited a thickness of 3.6 nm, HRP-EP1 increased to 5.1 nm post injection at 80 minutes.

Outlook

Polymer-protein hybrids offer a powerful approach to stabilize sensitive proteins in a range of environments. Here, we developed a robust design framework integrating automated polymer chemistry and machine learning to efficiently discover polymer-protein hybrids with enhanced thermostability for three chemically distinct enzymes. Notably, the machine learning-guided acquisition of data was effectively tailored to each enzyme. In addition, by analysis of developed surrogate machine learning models, we determined particular chemical features of copolymers that drive increased retained activity for each enzyme. Furthermore, the biophysical characterization of a successful polymer-protein hybrid design reveals chaperone-like assistance in structural refolding as a possible mechanism of stabilization. Taken together, these results highlight the existence of a complex structure-function relationship underlying protein-polymer hybrid activity that can be learned and exploited for materials optimization.

This discovery platform for polymer-protein hybrids can be extended in numerous directions. First, it provides an exemplary approach that can be extended to other proteins, other copolymer chemistries, and/or alternative design objectives, such as other environmental stresses. One

intriguing possibility is also to generalize the surrogate models to incorporate chemical features of both proteins and their encapsulating polymers. Additionally, the assay data collected in this study can be used in conjunction with simulation-based models to further elucidate and validate molecular-level mechanisms for stability. Such simulations might also aid in identifying and selecting key features for surrogate models or even provide *in silico* figures of merit that correlate with stability. Furthermore, the copolymer chemical space is large and flexible to accommodate the simultaneous pursuit of multiple design objectives, which could accelerate their adoption as functional, commercial materials.

References

1. Humphrey, W., Dalke, A. & Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **14**, 33–38 (1996).

2. Chapman, R. & Stenzel, M. H. All wrapped up: Stabilization of enzymes within single enzyme nanoparticles. *Journal of the American Chemical Society* **141**, 2754–2769 (2019).

3. Lancaster, L., Abdallah, W., Banta, S. & Wheeldon, I. Engineering enzyme microenvironments for enhanced biocatalysis. *Chemical Society Reviews* **47**, 5177–5186 (2018).

4. Pelegri-O'Day, E. M., Lin, E.-W. & Maynard, H. D. Therapeutic protein–polymer conjugates: advancing beyond pegylation. *Journal of the American Chemical Society* **136**, 14323–14332 (2014).

5. Ko, J. H. & Maynard, H. D. A guide to maximizing the therapeutic potential of protein–polymer conjugates by rational design. *Chemical Society Reviews* **47**, 8998–9014 (2018).

6. Kosuri, S. *et al.* Machine-assisted discovery of chondroitinase abc complexes towards sustained neural regeneration. *Advanced Healthcare Materials* **In Press**, 10.1002/adhm.202102101.

7. Panganiban, B. *et al.* Random heteropolymers preserve protein function in foreign environments. *Science* **359**, 1239–1243 (2018).

8. Gormley, A. J. & Webb, M. A. Machine learning in combinatorial polymer chemistry. *Nature Reviews Materials* **6**, 642–644 (2021).

9. Satari, B., Karimi, K. & Kumar, R. Cellulose solvent-based pretreatment for enhanced second-generation biofuel production: A review. *Sustainable Energy & Fuels* **3**, 11–62 (2019).

10. DelRe, C. *et al.* Synergistic enzyme mixtures to realize near-complete depolymerization in biodegradable polymer/additive blends. *Advanced Materials* **33**, 2105707 (2021).

11. DelRe, C. *et al.* Near-complete depolymerization of polyesters with nano-dispersed enzymes. *Nature* **592**, 558–563 (2021).

12. Wu, S., Snajdrova, R., Moore, J. C., Baldenius, K. & Bornscheuer, U. T. Biocatalysis: Enzymatic synthesis for industrial applications. *Angewandte Chemie International Edition* **60**, 88–119 (2021).

13. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: Recent applications and prospects. *npj Computational Materials* **3**, 54 (2017).

14. de Pablo, J. J. *et al.* New frontiers for the materials genome initiative. *npj Computational Materials* **5**, 1–23 (2019).

15. MacLeod, B. P. *et al.* Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances* **6**, eaaz8867 (2020).

16. Gómez-Bombarelli, R. *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials* **15**, 1120–1127 (2016).

17. Kumar, J. N. *et al.* Machine learning enables polymer cloud-point engineering via inverse design. *npj Computational Materials* **5**, 1–6 (2019).

18. Kumar, R. *et al.* Efficient polymer-mediated delivery of gene-editing ribonucleoprotein payloads through combinatorial design, parallelized experimentation, and machine learning. *ACS Nano* **14**, 17626–17639 (2020).

19. Wu, Y., Guo, J., Sun, R. & Min, J. Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells. *npj Computational Materials* **6**, 120 (2020).

20. Barnett, J. W. *et al.* Designing exceptional gas-separation polymer membranes using machine learning. *Science Advances* **6**, eaaz4301 (2020).

21. Wang, Y. *et al.* Toward designing highly conductive polymer electrolytes by machine learning assisted coarse-grained molecular dynamics. *Chemistry of Materials* **32**, 4144–4151 (2020).

22. Audus, D. J. & de Pablo, J. J. Polymer informatics: Opportunities and challenges. *ACS Macro Letters* **6**, 1078–1082 (2017).

23. Upadhya, R. *et al.* Automation and data-driven design of polymer therapeutics. *Advanced Drug Delivery Reviews* **171**, 1–28 (2021).

24. Chen, L. *et al.* Polymer informatics: Current status and critical next steps. *Materials Science and Engineering: R: Reports* **144**, 100595 (2021).

25. Lin, T.-S. *et al.* BigSMILES: A structurally-based line notation for describing macromolecules. *ACS Central Science* **5**, 1523–1531 (2019).

26. Ma, R. & Luo, T. PI1m: A benchmark database for polymer informatics. *Journal of Chemical Information and Modeling* **60**, 4684–4690 (2020).

27. Knox, S. T. & Warren, N. J. Enabling technologies in polymer synthesis: Accessing a new design space for advanced polymer materials. *Reaction Chemistry & Engineering* **5**, 405–423 (2020).

28. Webb, M. A., Jackson, N. E., Gil, P. S. & de Pablo, J. J. Targeted sequence design within the coarse-grained polymer genome. *Science Advances* **6**, eabc6216 (2020).

29. Jablonka, K. M., Jothiappan, G. M., Wang, S., Smit, B. & Yoo, B. Bias free multiobjective active learning for materials design and discovery. *Nature Communications* **12**, 1–10 (2021).

30. Reis, M. *et al.* Machine-learning-guided discovery of 19f mri agents enabled by automated copolymer synthesis. *Journal of the American Chemical Society* **143**, 17677–17689 (2021).

31. Rubens, M., Vrijsen, J. H., Laun, J. & Junkers, T. Precise polymer synthesis by autonomous self-optimizing flow reactors. *Angewandte Chemie International Edition* **58**, 3183–3187 (2019).

32. Tamasi, M., Kosuri, S., DiStefano, J., Chapman, R. & Gormley, A. J. Automation of controlled/living radical polymerization. *Advanced Intelligent Systems* **2**, 1900126 (2020).

33. Gormley, A. J. *et al.* An oxygen-tolerant PET-RAFT polymerization for screening structure-activity relationships. *Angewandte Chemie International Edition* **57**, 1557–1562 (2018).

34. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & de Freitas, N. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE* **104**, 148–175 (2016).

35. Patel, R. A., Borca, C. H. & Webb, M. A. Featurization strategies for polymer sequence or composition design by machine learning. *ChemRxiv* 10.33774/chemrxiv–2021–m74c8.

36. Kim, C., Chandrasekaran, A., Jha, A. & Ramprasad, R. Active-learning and materials design: The example of high glass transition temperature polymers. *MRS Communications* **9**, 860–866 (2019).

37. Shmilovich, K. *et al.* Discovery of self-assembling pi-conjugated peptides by active learning-directed coarse-grained molecular simulation. *The Journal of Physical Chemistry B* **124**, 3873–3891 (2020).

38. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777 (2017).

39. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence* **2**, 56–67 (2020).

40. Xu, J., Jung, K. & Boyer, C. Oxygen tolerance study of photoinduced electron transfer–reversible addition–fragmentation chain transfer (pet-raft) polymerization mediated by ru(bpy)3cl2. *Macromolecules* **47**, 4217–4229 (2014).

41. Ng, G. *et al.* Pushing the limits of high throughput pet-raft polymerization. *Macromolecules* **51**, 7600–7607 (2018).

42. Hopkins, J. B., Gillilan, R. E. & Skou, S. Bioxtas raw: improvements to a free open-source program for small-angle x-ray scattering data reduction and analysis. *Journal of Applied Crystallography* **50**, 1545–1553 (2017).

43. Franke, D. *et al.* Atsas 2.8: A comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *Journal of Applied Crystallography* **50**, 1212–1225 (2017).

44. Petoukhov, M. V. *et al.* New developments in the atsas program package for small-angle scattering data analysis. *Journal of Applied Crystallography* **45**, 342–350 (2012).

45. Huang, X., Bai, Q., Hu, J. & Hou, D. A practical model of quartz crystal microbalance in actual applications. *Sensors* **17**, 1785 (2017).

46. Su, X., Zong, Y., Richter, R. & Knoll, W. Enzyme immobilization on poly (ethylene-co-acrylic acid) films studied by quartz crystal microbalance with dissipation monitoring. *Journal of Colloid and Interface Science* **287**, 35–42 (2005).

47. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for hyper-parameter optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, 2546–2554 (Curran Associates Inc., Red Hook, NY, USA, 2011).

Methods

Materials. Hydroxypropyl methacrylate (HPMA), 2-diethylamino ethyl methacrylate (DEAEMA), [2-(methacryloyloxy)ethyl] trimethylammonium chloride solution (TMAEMC), and *N*-[3-(dimethylamino)propyl] methacrylamide (DMAPMA) were purchased from Sigma-Aldrich; methyl methacrylate (MMA) and 3-sulfopropyl methacrylate potassium salt (SPMA) from VWR; butyl methacrylate (BMA) from Alfa Aesar; and poly(ethyleneglycol) (*n*) monomethyl ether monomethacrylate (PEGMA, $M_n \approx 400$ g/mol) from Polysciences. PEGMA was deinhibited prior to use by passing over mono-methyl ether hydroxyquinone inhibitor removal resin. Ethyl 2-(phenylcarbonothioylthio)-2-phenylacetate, 4-nitrophenyl butyrate (PNB), hydrogen peroxide (H_2O_2), D-(+)-glucose, sodium acetate, lithium bromide were purchased from Sigma-Aldrich; zinc tetraphenyl porphyrin (ZnTPP), dimethyl sulfoxide (DMSO), 3,3',5,5'-tetramethylbenzidine (TMB) from Fisher Scientific; and potassium phosphate (mono and dibasic) and sodium acetate anhydrous from VWR.

Automated PET-RAFT synthesis. Copolymers were prepared by automated photoinduced electron/energy transfer reversible addition–fragmentation chain transfer (PET-RAFT) polymerization in 96 well plates as previously described.[32,33,40,41] Briefly, the sequences and processes to be conducted by the Hamilton MLSTARlet liquid-handling robot were programmed in Python, indicating information on sample concentration, reagent volumes, and well position. Files containing reaction information were transferred to the Hamilton MLSTARlet to prime the robotic transfers. Stock solutions of monomer (2 M), ethyl 2-(phenylcarbonothioylthio)-2-phenylacetate (RAFT chain-transfer agent (CTA), 100 or 50 mM) and ZnTPP (4 or 2 mM) were prepared in DMSO as 1 mL aliquots. Aliquots were loaded into the Hamilton MLSTARlet liquid-handling robot and automatically pipetted into 96-wells clear flat-bottom well plates (Greiner bio-one). Monomer/CTA ratio was varied from 100 – 400 while ZnTPP/CTA remained at 0.01. Polymer mixtures were dispensed to a total volume of 200 µL and final monomer concentration of 1 M. The mixtures were then covered with well-plate sealing tape and radiated under 560 nm LED light (5 mW/cm2, TCP 12 Watt Yellow LED BR30 bulb) for 16 hours.

HRP thermal stability assay. The activities of PPHs for HRP were evaluated by its ability to oxidize TMB in the presence of H_2O_2. Copolymers were synthesized and diluted in DMSO before further dilution into assay buffer (50 mM sodium acetate, pH 5.0) to a final concentration of 22.7 µM (<1% DMSO). From the 22.7 µM polymer samples, 50 µL were mixed with 50 µL of 10 µg/mL HRP (0.11 µM) in polystyrene 96 well plates. The solutions were thermally sealed with plate-sealing film and then thermally challenged in a water bath at 60°C for 30 minutes. Substrate solution was prepared by diluting 40 mM of TMB in DMSO to a final concentration of 0.4 mM in 1% H_2O_2 assay buffer. 5 µL of polymer-enzyme mixtures were added to 245 µL of substrate solution. Absorbance was measured in kinetic mode for 5 minutes in 20 second intervals; measurements were made at 653 nm, which is the maximum of the absorption peak. The initial rate of change of absorbance was used to calculate the activity of HRP. Native HRP activity at time $t = 0$ served as a positive control, while HRP heated at 60°C for 30 minutes served as the negative control.

GOx thermal stability assay. The activities of PPHs for GOx were evaluated using an assay buffer containing glucose, TMB, and HRP. Copolymers were diluted in DMSO and then in assay buffer (50 mM sodium acetate, pH 5.0) to a final concentration of 12 µM. Resulting solutions were mixed with equal volumes of stock GOx solution (5 µg/mL 30 nM) in polystyrene 96 well plates. The solutions were thermally sealed with plate-sealing film and then thermally challenged in a water bath at 65°C for 30 minutes. After heating, 20 µL of the PPH samples were added to 100 µL of substrate solution (5% glucose, 0.4 mM TMB, 0.11 µM HRP in assay buffer). Absorbance was measured in kinetic mode for 5 minutes in 20 second intervals; measurements were made at 653 nm, which is the maximum of the absorption peak. The initial rate of change of absorbance was used to calculate the enzyme activity. Native GOx activity at time $t = 0$ served as a positive control, while GOx heated at 65°C for 30 minutes served as the negative control.

Lip thermal stability assay. Activities of PPHs for Lip were evaluated using PNB as the substrate. Copolymers were diluted in DMSO and then in assay buffer (50 mM K_2HPO_4, 16.66 mM K_2HPO_4, pH 7.4) to a final concentration of 120 µM. From the 120 µM copolymer solutions, 50 µL were mixed with 50 µL of stock lipase solution (0.8 mg/mL 24 µM) in polystyrene 96 well plates. The solutions were thermally sealed with plate-sealing film and heated in a water bath at 70°C for one hour. Substrate solution was prepared by diluting stock PNB solution (5.4 M) first to 10 mM in DMSO, followed by a final dilution to 0.5 mM in assay buffer. Absorbance was measured in kinetic mode for 10 minutes in 20 second intervals; measurements were made at 410 nm to monitor the production of p-nitrophenol. The initial rate of change of absorbance was used to calculate the enzyme activity. Native Lip activity at time $t = 0$ served as a positive control, while Lip heated at 70°C for one hour served as the negative control.

Circular dichroism spectroscopy. CD wavelength and temperature scans of samples were collected using an AVIV Model 400 CD spectrometer (AVIV Biomedical Inc.). Wavelength scans consisted of measurements from 260 nm to 190 nm, collecting points every 0.5 nm with a 1-nm bandwidth for 5 seconds, at all required temper-

atures. Temperature scans were consisted of measuring mean residue ellipticity at 222 nm from 30 to 90°C with a 5-second averaging time and 1.5-nm bandwidth. The ramp rate was 2°C/minute, and samples were equilibrated for 5 minutes at each temperature before measurement. The fraction of protein unfolding at different temperatures were calculated by assuming fully folded state at 30°C and fully unfolded state at 90°C. The melting temperature T_{m} was determined by fitting the temperature scans to a Boltzmann sigmoidal equation. The fractions of α-helices and β-sheets in the protein samples were calculated using CD deconvolution algorithms for wavelength scans (Table S2).

Dynamic light scattering. DLS of copolymers and polymer-enzyme mixtures were performed on a DynaPro DLS Plate Reader III, Wyatt Technologies. Concentration of HRP for DLS experiments was maintained at 0.2 mg/mL while polymer concentration was at 1 mg/mL. The data was collected using a wavelength of 830 nm and a scattering angle of 173°. Fifteen acquisitions were collected for each sample with an acquisition time of 5 seconds per acquisition using auto attenuation. Regularization analysis was performed using Rayleigh spheres model for hydrodynamic size measurement.

Small-angle X-ray scattering. All scattering experiments were carried out at the Life Science X-ray Scattering (LiX) beamline 16-ID of the National Synchrotron Light Source II (NSLS-II) at Brookhaven National Laboratory (Upton, NY). HRP was prepared at a final concentration of 1 mg/mL in 50 mM sodium acetate (pH 5.15) while lyophilized polymers were reconstituted in sodium acetate buffer and mixed with HRP at a final concentration of 2.61 mg/mL (10:1 molar concentration of polymer:HRP). Samples were denatured by heating in a water bath at 65 °C for 1 hour. All solutions were loaded into 96-well PCR plates and mailed in for data collection. An X-ray energy of 15.14 keV was utilized for solution SAXS. Three Pilatus detectors were employed to provide a q range of 0.005 - 3.13 Å, while the range 0.005 - 0.25 Å was taken as the small-angle region. For background subtraction, sodium acetate buffer blanks were run for every three samples. The subtracted data were analyzed in BioXTAS RAW 2.1 with ATSAS 3.0.4-6. Guinier analysis was performed to quantify the radius of gyration R_g, whereas pair-distance distribution analysis by an indirect Fourier transform method was conducted to quantitatively assess R_g, maximum dimension, and macromolecular structure.[42–44]

Quartz crystal microbalance with dissipation. All quartz crystal microbalance experiments were carried out on the Q-Sense Omega Auto (Biolin Scientific) with 5 MHz sensitivity, less than 1 nm surface roughness, and theoretical mass sensitivity of 17.7 ng cm^{-2} Hz^{-1}. HRP was dissolved in 50 mM sodium acetate buffer (pH 5.15) at 0.2 mg/mL whereas the final concentration of lyophilized polymers was set to 0.52 mg/mL (10:1 molar concentration of polymer:HRP). Sodium acetate buffer was flowed as an initial equilibration step at 20 µL/min for 25 min. HRP, polymer, and mixtures of HRP with polymer were flowed at 40 µL/min for 10 min. Sodium acetate was flowed after each step at 20 µL/min for 25 min to remove any loosely associated enzyme or polymer. Transformations using the Sauerbery equation[45,46] were completed on the fifth harmonic frequency and dissipation responses to obtain surface thickness.

Polymer characterization. The molecular weights (M_w and M_n) and dispersity (Đ) were measured by gel-permeation chromatography using an Agilent 1260 Infinity II. Polymer samples were eluted through a Phenomenex 5.0 µm guard column (50 x 7.5 mm) preceded by superose Phenogel 12 10/300 GL column (Cytiva 17-5173-01, column L × I.D. 30 cm × 10 mm, 11 m avg. part. size) in 0.5x PBS (0.2% NaN3) using a flow rate of 0.5 ml/min. GPC calibration was completed with Agilent PEG standards. Polymers were prepared at 50:1 eluent/polymer ratio in 0.5x PBS (0.2% NaN3) and filtered with a 0.45 µm nylon filter. Polymer conversion was calculated by obtaining ^{1}H NMR spectra using a Varian VNMRS 500 MHz spectrometer with mesitylene as an internal standard and processed using Mestrenova 11.0.4.

Machine learning surrogate models. All copolymers were featurized as DP-explicit composition vectors with one-hot encoded fingerprints of the monomer units.[35] With eight possible monomers, the resulting feature vector possesses nine dimensions, with the first containing the DP of the copolymer divided by 200 and the remaining eight containing the fractions of incorporation for each monomer; the division in the first dimension represents DP on a similar scale as the remaining features. Gaussian process regression (GPR) models, trained to predict the Yeo-Johnson transformation of the REA for a PPH, were preferred due to their superior predictive performance compared to other ML algorithms (Fig. S3). In addition, preliminary comparisons amongst GPR models trained over the seed datasets revealed no evident advantage to using more advanced fingerprinting strategies over simple one-hot encoding (Fig. S3). Using available experimental data of various PPHs, we constructed enzyme-specific datasets wherein each datum is described by this feature vector and labeled by REA.

We modelled the relationship between our copolymer features and REA using GPR to both capture the nontrivial, nonlinear mapping and to facilitate AL as GPR naturally provides uncertainty estimates on predicted labels. Covariances of points that are modeled by the Gaussian Process are calculated using the squared exponential kernel basis function:

$$k(\vec{x},\vec{x}') = \sigma^2 \exp(-\frac{1}{2}\frac{(\vec{x}-\vec{x}')^2}{l^2}) + \sigma_n^2,$$

where $\vec{x}$ is the feature vector of the copolymer, and textitl, σ, σ_n are kernel hyperparameters. Anisotropic kernels were explored but did not improve model performance. Hyper-

parameters were tuned using the Tree-structured Parzen Estimator Approach (TPE), implemented by the Hyperopt Python package.[47]

GPR models for each enzyme are constructed as follows: the dataset is first split into five folds. Four of five of the folds are then used to tune the GPR model hyperparameters, which are identified with 20-fold cross-validation and optimization by TPE to minimize the mean squared error of labels. The optimal hyperparameters, along with data from four of five folds, are used to train a GPR model that makes predictions on the remaining fold of data. This process is repeated four more times, such that all five of the original folds have served as test sets. The five sets of optimized hyperparameters are then averaged and used to define a final GPR model with the full set of data available for an enzyme at a given iteration. The five sets of held-out test performance metrics are also averaged to quantify and validate the predictive capabilities of the model.

Candidate copolymer generation. We use Bayesian optimization (BO) in tandem with a GPR model to propose promising candidate copolymers. For the first four rounds of active learning, we select candidates that maximize the expected improvement (EI) acquisition function given by

$$f(\vec{x}) = Z\sigma(\vec{x})\Phi(Z) + \sigma(\vec{x})\phi(Z)$$

$$Z = \begin{cases} \frac{(\mu(\vec{x}) - f' - \xi)}{\sigma(\vec{x})} & \sigma(\vec{x}) > 0 \\ 0 & \sigma(\vec{x}) = 0 \end{cases}$$

where $f(\vec{x})$ is the predicted mean REA from the GPR, f' is the current largest mean REA observed by the model, $\sigma(\vec{x})$ is the standard deviation from the GPR, Φ and ϕ are the cumulative and probability density functions of the normal distribution, respectively, and ξ is a hyperparameter that controls the balance between exploring unobserved regions of the chemical space and exploiting known regions of it to obtain high performing polymers.

To effectively sample copolymer designs that live on the exploit-explore spectrum, we sequentially generate 200 copolymer candidates for distinct ξ values that logarithmically vary from 0.001 to 30. To avoid proposing previously synthesized polymers or those within the margin of synthetic experimental error previously synthesized or already proposed polymers, an additional penalty function is added to the acquisition function based on $\vec{x}$ (see also Supporting Information). In the final iteration or exploit round, copolymers that simply maximize REA predictions from the GPR model are proposed as candidates, although the penalty function is retained to avoid redundant proposals.

Candidate copolymer down-selection. Unsupervised clustering methods were used to select 24 candidates for synthesis from a larger set of 200 candidates generated by the BO procedure. In particular, the following protocol was used for candidate selection in the first four AL iterations. First, a filter was applied to ensure that no copolymer featured fractions of incorporation of any given monomer that was less than 5%. This filter was imposed to establish reasonable margins of experimental control over the process of dispensing the monomer reagents with the robotic arm used to automatically synthesize the copolymers. Second, candidates were subsequently clustered using Density-based spatial clustering of applications with noise (DBSCAN) using a distance threshold of $0.05\sqrt{2}$ and a minimum of three points per cluster. Following the formation of clusters, the copolymer with the shortest Euclidean distance to the centroid poistion of the cluster in the copolymer feature vector space was selected as a representative candidate for futher consideration. All non-clustered candidates, or noise-points, were also considered In this fashion, the procedure produced a set of relatively diverse and representative copolymer candidates that fairly considers "outliers." Third, in cases where DBSCAN produced more than 24 candidates (this always occurred), we ensured that precisely 24 candidates were proposed by application of k-Means clustering. Here, again, representative candidates are chosen based on proximity to the cluster centroid. If a cluster consisted of only two points, then the candidate with the higher REA was used. A different down-smapling procedure was used in the exploit round, since diversity was no longer a priority for selection. Specifically, after producing the 200 polymer designs with BO, candidates were ranked by their REA in descending order and iteratively chosen for the final set of 24 candidates, provided they had compositions that were unique (within synthetic precision) from any polymers that constituted the growing list at that point.

Handling polymer gelation. Upon construction of the seed database and throughout the AL, a handful of copolymers were found to phase separate into a liquid and gel phase. While gelling polymers recorded nonzero REA values, they were excluded from the dataset used to train the GPR models from iteration 1 onward due to the potential uncontrolled differences in copolymer - enzyme interaction environments that could obfuscate model training. However, the penalty function was used during the active learning procedure to avoid suggesting polymer candidates proximate to gelling polymers across discovery campaigns across all three enzymes up to that iteration. While this strategy limited the number of gelled polymers per iteration per enzyme to an average of six copolymers in the first two rounds of AL, it ultimately proved ineffective for GOx as hydrophobic monomers were found to be effective for GOx stabilization but increased polymer gelation (Fig. S11). To combat this issue, a classifier that leveraged knowledge of prior polymer gelation across all enzymes and iterations up to that point was designed and integrated in the AL scheme. The use of the classifier was limited to and ultimately facilitated the discovery of primarily soluble polymers for iterations 4 and 5 of AL for GOx. Further discussion on the development and integration of the

classifier into the active learning scheme is supplied in the supporting information (Table S5, Fig. S11, Fig. S12).

Data and Code availability

All experimental data used to develop machine learning models are available in supporting information. In addition, all datasets will be published and available for download in .csv format from DataSpace, Zenodo, and Materials Data Facility. The code used in the development of the Gaussian process regression model development and training will be available on GitHub (https://github.com/webbtheosim/PPH_public) with trained machine learning models available in .pkl format as described on the Github repository. Python scripts used to perform SHAP analysis will also be available. Prior to publication, these materials are available by reasonable request from the corresponding authors.

Acknowledgements

A.J.G. acknowledges support from the National Institutes of Health under NIGMS MIRA Award R35GM138296, and the National Science Foundation under DMREF Award NSF-DMR-2118860 and CBET Award Number NSF-ENG-2009942. R.A.P., C.H.B., and M.A.W. acknowledge support from the National Science Foundation under DMREF Award Number NSF-DMR-2118861 as well as start-up funds from Princeton University. M.J.T. acknowledges additional support from the National Institute of Health (GM135141). The training of and optimization with machine learning models was performed with resources from Princeton Research Computing at Princeton University, which is a consortium led by the Princeton Institute for Computational Science and Engineering (PICSciE) and Office of Information Technology's Research Computing. A.J.G. and N.S.M. acknowledge James Byrnes, beamline scientist at NSLS-II beamline 16-ID for Life Science X-ray Scattering (LiX), for his assistance with conducting experiments at Brookhaven National Laboratory. The LiX beamline is part of the Center for BioMolecular Structure (CBMS), which is primarily supported by the National Institutes of Health, National Institute of General Medical Sciences (NIGMS) through a P30 Grant (P30GM133893), and by the DOE Office of Biological and Environmental Research (KP1605010). LiX also received additional support from NIH Grant S10 OD012331. As part of NSLS-II, a national user facility at Brookhaven National Laboratory, work performed at the CBMS is supported in part by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences Program under contract number DE-SC0012704.

Author contributions

M.J.T., R.A.P., C.H.B., S.K., M.A.W., and A.J.G. conceptualized the study. M.J.T., S.K., H.M., and R.U. performed physical experiments and analyzed the results. M.J.T., R.A.P., C.H.B., and M.A.W. developed all machine learning models and analyzed the results. The overall project was supervised by A.J.G., M.A.W, and N.S.M. Further, the manuscript was drafted by M.J.T., R.A.P., M.A.W., and A.J.G. with contributions from all authors.

Competing interests

M.J.T. and A.J.G. have filed a PCT patent application and are co-founders of Plexymer, Inc.