Data-driven de novo design of super-adhesive hydrogels

https://doi.org/10.1038/s41586-025-09269-4

Received: 20 November 2024

Accepted: 11 June 2025

Published online: 6 August 2025

Open access

Check for updates

Hongguang Liao[1,10], Sheng Hu[2,3,10], Hu Yang[4], Lei Wang[2,5], Shinya Tanaka[2,5], Ichigaku Takigawa[2,6] ✉, Wei Li[2,7] ✉, Hailong Fan[2,9] ✉ & Jian Ping Gong[2,8] ✉

Data-driven methodologies have transformed the discovery and prediction of hard materials with well-defined atomic structures by leveraging standardized datasets, enabling accurate property predictions and facilitating efficient exploration of design spaces[1–3]. However, their application to soft materials remains challenging because of complex, multiscale structure–property relationships[4–6]. Here we present a data-driven approach that integrates data mining, experimentation and machine learning to design high-performance adhesive hydrogels from scratch, tailored for demanding underwater environments. By leveraging protein databases, we developed a descriptor strategy to statistically replicate protein sequence patterns in polymer strands by ideal random copolymerization, enabling targeted hydrogel design and dataset construction. Using machine learning, we optimized hydrogel formulations from an initial dataset of 180 bioinspired hydrogels, achieving remarkable improvements in adhesive strength, with a maximum value exceeding 1 MPa. These super-adhesive hydrogels hold immense potential across diverse applications, from biomedical engineering to deep-sea exploration, marking a notable advancement in data-driven innovation for soft materials.

Designing soft materials, such as gels and elastomers, is a complex task. It requires selecting appropriate types and quantities of building blocks (for example, monomers) and determining their arrangement in the material, creating a gigantic design space with countless possible combinations. Moreover, soft materials exhibit intricate behaviours because of the interplay of weak molecular interactions and thermal fluctuations, resulting in complex structure–property relationships across multiple time and length scales, with mesoscale structures playing an important part[7].

These complexities hinder the development of accurate predictive theories or computational models, often rendering soft material discovery reliant on experimental trial and error. To reduce experimental demands, data-driven strategies are becoming increasingly essential[8,9]. Emerging tools, such as data mining (DM) and machine learning (ML), are transforming the field by advancing the analysis of complex behaviours, improving property predictions and driving theory and modelling development[5,10–13].

Effectively integrating these tools into an end-to-end design framework is important for accelerating soft material discovery. An important first step is the creation of high-quality datasets, which is complicated by the several potential material designs and limited experimental throughput[14,15]. Adhesive hydrogels, for example, are a promising class of soft material widely sought for high-end applications. Yet achieving instant, strong and repeatable underwater adhesion remains a longstanding challenge[16,17]. Previous studies on this material have identified several monomer types, making it difficult to form a consistent dataset or forge a simple design principle for optimizing performance[16].

Biological soft tissues, as naturally evolved soft materials, exemplify complex structures tailored for specific functions[18]. Studying these systems can help reduce the design space for synthetic soft materials[19], such as gecko-inspired dry adhesives[20,21]. Particularly, adhesive proteins, found across diverse organisms (for example, archaea, bacteria, eukaryotes and viruses), enable adhesion in wet environments. Despite their diversity, these proteins share common sequence patterns that offer valuable insights into designing underwater adhesives[22]. However, identifying meaningful patterns, translating them into synthesis strategies and enabling extrapolative predictions by machine learning remain main challenges to achieving an end-to-end design model.

Here we introduce a new data-driven approach that integrates DM, experimentation and ML for the efficient development of high-performance underwater adhesive hydrogels (Fig. 1a). By mining adhesive protein databases, we extract characteristic sequence features to guide hydrogel design. These features are replicated in 180 synthetic hydrogels using random copolymerization and relative composition strategies, which strike a balance between biological fidelity and practical synthesis. Among these DM-driven hydrogels, several exhibit greater adhesive strength (F_a) than those reported in the literature (Fig. 1b). This set of 180 synthetic hydrogels forms a small yet high-quality dataset

[1] Graduate School of Life Science, Hokkaido University, Sapporo, Japan. [2] Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Sapporo, Japan. [3] Artificial Intelligence Research Center (AIRC-ISIR), Osaka University, Osaka, Japan. [4] School of Information, Central University of Finance and Economics, Beijing, People's Republic of China. [5] Department of Cancer Pathology, Faculty of Medicine, Hokkaido University, Sapporo, Japan. [6] Center for Innovative Research and Education in Data Science (CIREDS), Institute for Liberal Arts and Sciences, Kyoto University, Kyoto, Japan. [7] Suzhou Laboratory, Suzhou, People's Republic of China. [8] Faculty of Advanced Life Science, Hokkaido University, Sapporo, Japan. [9] Present address: College of Chemistry and Environmental Engineering, Shenzhen University, Shenzhen, People's Republic of China. [10] These authors contributed equally: Hongguang Liao, Sheng Hu. ✉e-mail: takigawa@icredd.hokudai.ac.jp; liw@szlab.ac.cn; fanhl@szu.edu.cn; gong@sci.hokudai.ac.jp

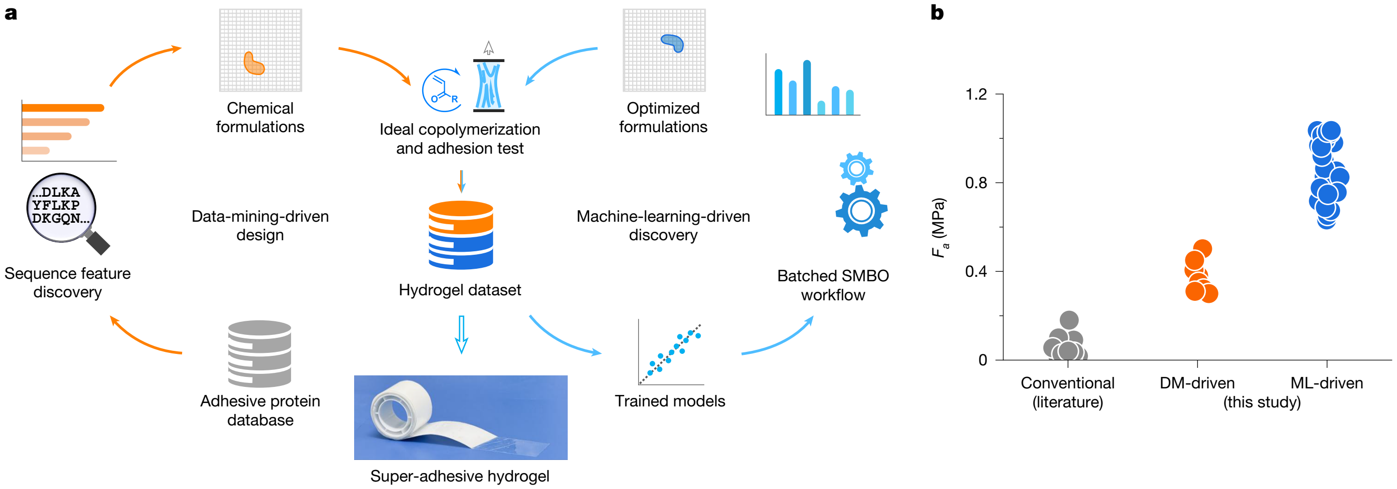


Fig. 1 | Data-driven de novo design of underwater adhesive hydrogels. **a**, Conceptual scheme of the proposed approach integrating DM, experimentation and ML to design high-performance adhesive hydrogels. **b**, Comparison of underwater adhesive strength (F_a) between previously reported hydrogels (Supplementary Table 1) and newly developed hydrogels in this study (DM-driven and ML-driven). F_a was measured using tack tests, and the testing conditions were optimized for maximum performance.

for further optimization by ML, leading to ML-driven hydrogels with underwater F_a exceeding 1 MPa—an order-of-magnitude improvement over previously reported underwater adhesive hydrogels and elastomers[16] (Supplementary Fig. 1).

The obtained super-adhesive hydrogels hold tremendous potential across a wide range of applications, offering reliable solutions for which traditional adhesives often fall short (Supplementary Fig. 1). They could improve medical procedures, advance biomedical engineering, support marine farming and enable deep-sea exploration. The substantial performance improvements showcase the success of our data-driven approach in designing high-performance hydrogels. Moreover, this approach is highly versatile and can be adapted to develop other types of functional soft materials, opening new possibilities in various fields.

DM of adhesive proteins

We compiled a dataset containing 24,707 adhesive proteins gathered from the National Center for Biotechnology Information (NCBI) protein database, using the keyword 'adhesive protein'. This dataset includes proteins from 3,822 different organisms across archaea, bacteria, eukaryotes, viruses and artificial proteins. Statistical analysis shows that the average length of those adhesive proteins ranges from approximately 300–500 amino acids (Supplementary Fig. 2).

To identify the most representative protein sequences and minimize the impact of individual variations, we ranked all species by the number of adhesive proteins they contain and selected the top 200 species for further analysis (Fig. 2a and Supplementary Fig. 3). We then performed multiple sequence alignment using Clustal Omega[23] to determine consensus sequences for each species (Extended Data Fig. 1), which are believed to play a crucial part in maintaining protein stability and adhesion throughout evolution[24,25].

To reduce the dimensionality of the variables, the 20 canonical amino acids were grouped into six classes based on their physicochemical properties: hydrophobic, nucleophilic, acidic, cationic, amide and aromatic (Supplementary Fig. 4). The consensus sequences were then encoded into functional class sequences. For consistency in the encoding, glycine, alanine and proline were excluded from the hydrophobic class because of their smaller side chains, which are proposed to have a less important role in interfacial contacts and interactions compared with other amino acids[26].

The block length of each functional class in the encoded sequences is typically less than three (Fig. 2b), indicating substantial sequence heterogeneity in adhesive proteins even at the coarse functional class level. Different species exhibited distinct patterns in the pairwise frequencies of these functional classes (Fig. 2c). This suggests preferences for specific functional class pairings within the sequences, hinting at an underlying order beneath the observed sequence heterogeneity.

Based on these insights, we devised a strategy for hydrogel design using six functional monomers to represent the six functional classes of amino acids. Although directly replicating functional class sequences offers a straightforward way to mimic protein primary structures and functions, achieving precise control over monomer sequences in synthetic polymers remains a marked challenge. Therefore, we aimed to statistically replicate the sequence features of functional classes through ideal random copolymerization of the six functional monomers, which has minimal composition drift during polymerization and enables statistically controlled sequences[19,27–29].

For this purpose, we used a relative composition approach to capture the neighbouring preferences of amino acid functional classes in the synthetic polymer chains. Specifically, we counted the occurrences of 21 distinct pair types for the six functional classes, denoted as n_{ij} (where $i, j = 1, \ldots, 6$), along the functional class sequences for each species and ranked them in descending order. The top five pairs, collectively accounting for approximately 50% of all occurrences, were used to compute the monomer proportions of each functional class as $\phi_i = N_i/\sum_i N_i$, where $N_i = \sum_j (n_{ij} + n_{ji})$ for each species (Extended Data Fig. 1 and Supplementary Data 1 and 2). These relative compositions served as descriptors for the corresponding species. From the top 200 species, we derived 180 unique compositions after removing 20 duplicates (Supplementary Table 2), which were then used for hydrogel synthesis.

Synthesis of DM-driven hydrogels

Six functional monomers (Fig. 3a), each representing one of the six functional classes of amino acids, were selected. Their pairwise reactivity ratios, determined by ^{1}H NMR analysis, were close to unity when copolymerized in the cosolvent dimethyl sulfoxide (DMSO) using free-radical polymerization (Supplementary Fig. 5 and Supplementary Table 3). These near-unity values indicate minimal composition drift during copolymerization in DMSO (Supplementary Figs. 6 and 7).

Monte Carlo simulations based on the Mayo–Lewis model were performed to analyse the sequence properties of the six functional monomers in the corresponding 180 heteropolymers, using the measured

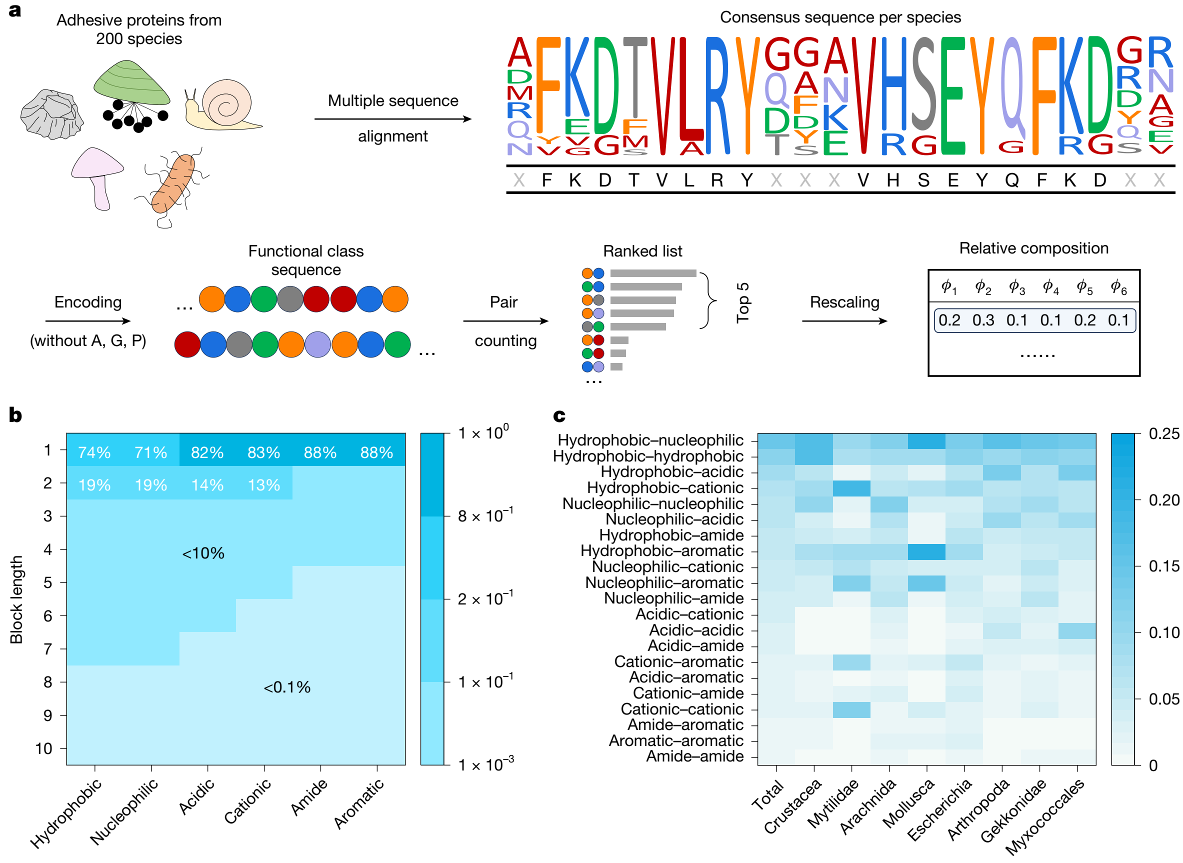


Fig. 2 | DM of adhesive proteins and formulation design. a, Schematic of the amino acids feature extraction process used to derive bioinspired formulations through adhesive protein DM, encoding and relative composition computation. **b**, Distribution of block length (that is, the number of consecutive residues from the same functional class) for the six functional classes, shown along the horizontal axis, based on the consensus sequences of the top 200 species. **c**, Pairwise frequency distribution of the 21 functional class pair types along encoded sequences, shown for the entire dataset and for eight representative species, shown along the horizontal axis, categorized by their biological classifications in the database.

reactivity ratios (Supplementary Table 3) and the derived monomer proportions (ϕ_i) (ref. 30) (Supplementary Table 2). The resulting distributions of monomer block lengths and pairwise frequencies (Fig. 3b,c) closely matched those observed in adhesive proteins (Fig. 2b,c), confirming that our synthesis protocol effectively captures key statistical features (Supplementary Fig. 8), such as sequence heterogeneity and neighbouring preferences.

Following the derived formulations, 180 DM-driven gels, labelled G-001 to G-180, were synthesized by one-pot free-radical copolymerization of the functional monomers with crosslinkers in DMSO (Methods and Supplementary Fig. 9). After solvent exchange from DMSO to normal saline (0.154 M NaCl), the hydrogels were characterized by volume swelling ratio, rheological behaviour and underwater adhesive strength (F_a). Adhesion was assessed using tack tests (Fig. 3d and Supplementary Fig. 10) on a glass substrate in normal saline, with a loading force of 10 N and a 10-s contact time applied for rapid screening.

Figure 3e shows the measured F_a for all 180 hydrogels (15 mm diameter, 0.3–0.8 mm thickness). Among them, 16 hydrogels exhibited robust adhesion with $F_a > 100$ kPa, and 83 hydrogels showed $F_a > 46$ kPa, surpassing the average reported in the literature (Supplementary Table 1). Notably, G-042 (derived from *Escherichia*, Supplementary Fig. 8), hereafter referred to as G-max, presented the highest adhesive strength of 147 kPa.

The high F_a values demonstrate the effectiveness of our data-driven approach in guiding the de novo design of adhesive hydrogels, highlighting two key insights. First, the functional class sequences extracted through DM capture the essential sequence features of adhesive proteins that are important for wet adhesion. Second, using ideal random copolymerization of functional monomers to statistically replicate these sequence features through relative compositions provides an effective strategy, bridging the gap between de novo design and material fabrication.

To validate the first insight, we examined the adhesion performance of hydrogels formulated using sequences derived from DM of resilin proteins. These hydrogels exhibited poor underwater adhesion (Extended Data Fig. 2 and Supplementary Table 4), underscoring the importance of specific sequence features from adhesive proteins for effective adhesion.

To validate the second insight, we analysed the adhesion performance of hydrogels synthesized by non-ideal copolymerization in dimethyl sulfide (DMS). In DMS, most pairwise reactivity ratios of monomers deviate significantly from unity (Supplementary Table 3), resulting in composition drift during polymerization and the formation of blocky sequences (Supplementary Figs. 6 and 7). Figure 3f compares two variants of G-004, showing that the variant synthesized in DMS appeared more translucent and exhibited markedly lower F_a than its counterpart with statistical sequences synthesized in DMSO. This finding underscores the important role of ideal random copolymerization of functional monomers (with near-unity reactivity ratios) in achieving the statistical sequence features essential for mimicking protein functions[19,27].

To improve F_a, we assessed the correlations between F_a and ϕ_i using Kendall's τ coefficients[31] and characterized the dependence of F_a on the swelling of hydrogels and rheological behaviours (Extended Data Fig. 3). We found that ϕ_{ATAC}, ϕ_{BA} and ϕ_{PEA} exhibit weak positive correlations with F_a, whereas ϕ_{HEA}, ϕ_{AAm} and ϕ_{CBEA} show weak negative

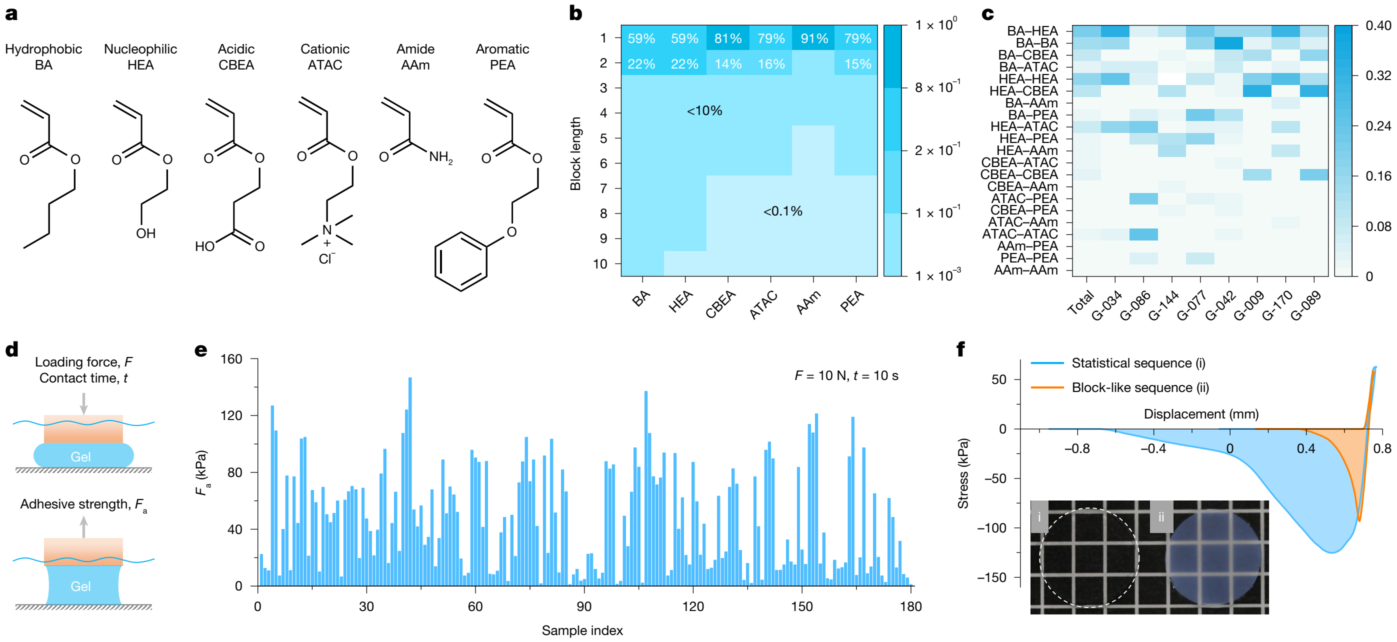


Fig. 3 | DM-driven hydrogels for underwater adhesion. a, Chemical structures of six functional monomers, each representing one of the six functional classes of amino acids. **b**, Distribution of monomer block lengths in heteropolymer sequences generated by Monte Carlo simulations based on experimentally determined reactivity ratios and derived formulations. **c**, Pairwise frequency distribution of monomer pairs in heteropolymer sequences obtained by Monte Carlo simulations, shown for all 180 derived formulations and for eight formulations (denoted by gel index) corresponding to the representative species shown in Fig. 2c. **d**, Schematic of the tack test for measuring underwater adhesion. **e**, Adhesive strength (F_a) of the 180 hydrogels. **f**, Stress–displacement profiles of two G-004 variants in the tack test: (i) statistical sequences synthesized in DMSO and (ii) block-like sequences synthesized in DMS. Inset images show the appearance of the two hydrogels. Adhesion tests were conducted under a 10-N loading force applied for 10 s on a glass substrate in normal saline (0.154 M NaCl). This test condition was used for rapid screening.

correlations. Nevertheless, these weak correlations, along with the intricate structure–property relationships (Extended Data Fig. 3), are insufficient to directly predict hydrogel formulations for optimal adhesion, highlighting the complex synergistic effects of monomer species, sequences and network structures.

Hydrogel optimization by ML

Next, we used ML to explore hydrogel formulations with enhanced adhesive strength, starting with the 180-hydrogel dataset. Among nine ML models benchmarked (Supplementary Tables 5 and 6), Gaussian process (GP)[32] and random forest regression (RFR)[33] emerged as the most effective base models for predicting F_a from ϕ_i, achieving low test error while minimizing overfitting (Extended Data Fig. 4).

Based on these models, we implemented sequential model-based optimization (SMBO)[33] to propose new hydrogel formulations, taking expected improvement (EI) as the acquisition function. To reduce the number of experimental rounds of hydrogel synthesis and characterization, we designed a batched SMBO workflow, which allows for multiple formulation proposals in a single round.

To enhance efficiency, we explored several batched SMBO methods, using trained base models as the hypothetical value providers (P) and GP, RFR, extra trees (ETR)[34] and gradient boosting machine (GBM)[35] as the EI maximizers (M), collectively denoted as P–M. We also implemented traditional Bayesian optimization methods, using kriging believer (GP_KB) (ref. 36), maximum and minimum constant liar (GP_CLmax, GP_CLmin) (ref. 36) and local penalization (GP_LP) (refs. 36,37) as heuristics for determining batch points. For validation, we selected the top 10 formulations (out of 40 proposed per batch), sorted by either EI magnitude or predicted F_a (PRED) as experimental test sets.

All validation followed the same protocol as for the training set to ensure data consistency. Figure 4a shows the true F_a values for formulations proposed by different SMBO methods (Supplementary Table 7). Non-SMBO baselines, GP_enu and RFR_enu, which selected the top five PRED from an enumeration of 10 million random formulations, failed to improve F_a beyond the training data. By contrast, all SMBO methods achieved higher F_a, with GP_KB and RFR-GP as the top performers, and RFR-GP yielding the highest F_a overall.

We further tested a 'warm-start' strategy using RFR-GP by adding 10 additional data points generated by RFR to the training set. This variant, termed RFR-GP*, exhibited the highest F_a among all models. Furthermore, formulations chosen through PRED sorting generally outperformed those selected by EI sorting. These findings demonstrate the effectiveness of batched SMBO and suggest the optimal models and strategies for improving workflow efficiency.

The validation outcomes expanded our hydrogel dataset. To assess the exploration abilities of RFR-GP and GP_KB within the SMBO framework, we conducted two additional rounds of ML optimization and experimental validation. Although new high-F_a formulations were identified, none surpassed the maximum F_a achieved in the first round (Extended Data Fig. 5). We suspect that the functionalities of the adopted monomer species may account for the observed performance plateau, and further optimization rounds were not pursued.

The relationship between F_a and ϕ_i in the final dataset (containing 341 hydrogels) is shown in Fig. 4b, using uniform manifold approximation and projection (UMAP)[38] for dimensional reduction (from six to two dimensions). Notably, formulations generated by RFR-GP and GP_KB show minimal overlap with the original 180-hydrogel dataset, indicating extrapolation during optimization. RFR-GP data points are more scattered than those of GP_KB, suggesting broader exploration compared with traditional Bayesian optimization.

To assess the influence of ϕ_i on F_a, we used SHAP (SHaply Additive exPlanations)[39] with the RFR model trained on the final 341-hydrogel dataset. The SHAP summary plot (Fig. 4c) shows that high values of ϕ_{BA} and ϕ_{PEA} significantly enhance F_a. This is because BA and PEA effectively expel water from the contact interface, and, when neighbouring with

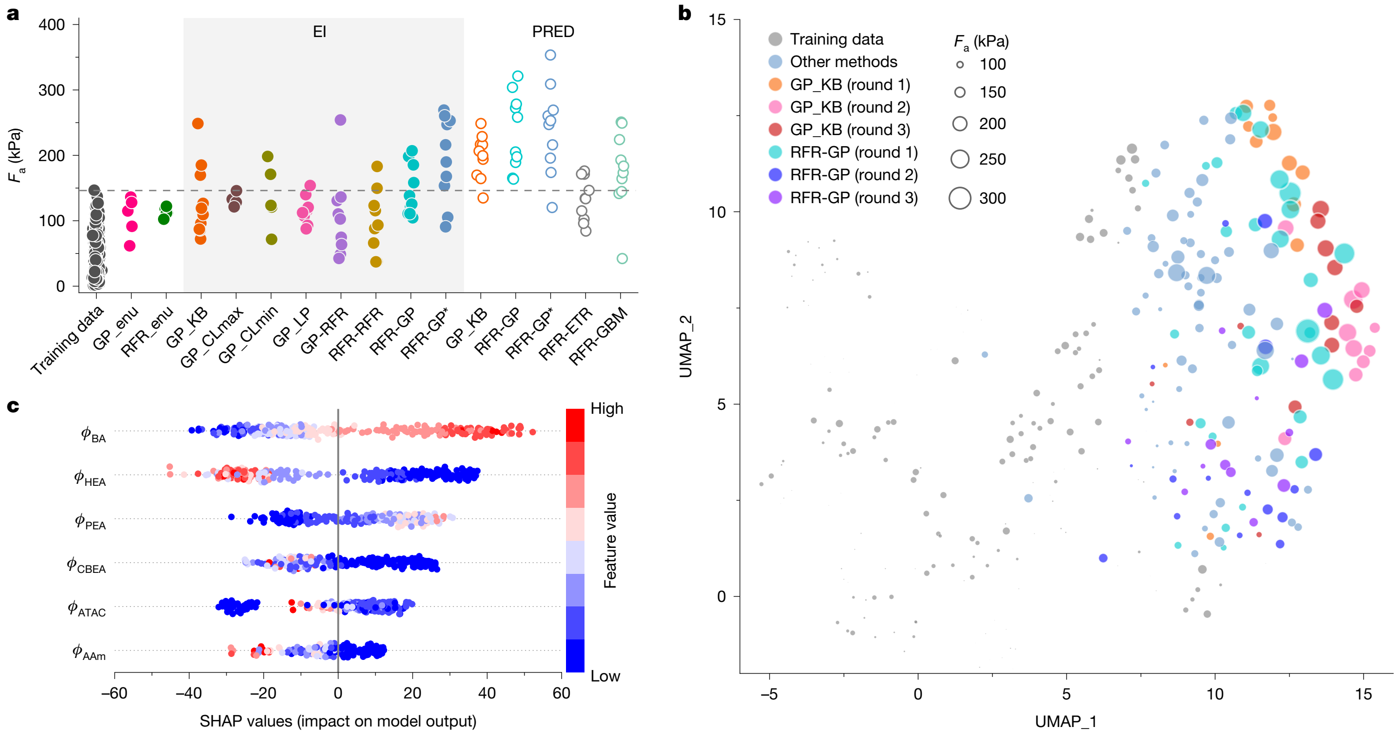


Fig. 4 | ML optimization of underwater adhesive hydrogels. a, Adhesive strength (F_a) of hydrogels fabricated based on predictions from various models trained on the 180-hydrogel dataset. The model nomenclature and detailed descriptions are provided in the Methods. All adhesion measurements were performed under the same test conditions as the training set: 10 N loading force, 10-s contact time, on a glass substrate and in normal saline. **b**, UMAP representation of the relationship between F_a and reduced monomer proportions (ϕ_i), highlighting the formulations proposed by GP_KB and RFR-GP (within the SMBO framework) across different rounds. Symbol size represents the magnitude of adhesive strength. **c,** SHAP beeswarm plot, ranked by mean absolute SHAP values, showing the influence of ϕ_i on F_a within the final dataset of 341 samples as analysed by the trained RFR model.

ATAC (Supplementary Fig. 11), they could enhance electrostatic interactions with the negatively charged glass surface[27,40–43] (Supplementary Fig. 12). By contrast, high values of ϕ_{HEA}, ϕ_{CBEA}, and ϕ_{AAm} tend to reduce F_a. Interestingly, ϕ_{ATAC} has a dual effect (Supplementary Fig. 13): low levels diminish electrostatic interactions, whereas excessive ϕ_{ATAC} increases hydrogel swelling, limiting polymer-surface contact and reducing F_a. Therefore, a moderate ϕ_{ATAC} is crucial.

These insights, consistent across all three ML rounds, establish a clear design principle for achieving strong underwater hydrogel adhesion to glass surfaces using the selected functional monomers: incorporating BA, PEA and ATAC is key. This combination leverages both hydrophobic effects and electrostatic interactions to enhance underwater adhesion to negatively charged surfaces. The hydrogels with the highest F_a from each ML round, denoted as R1-max, R2-max and R3-max, are exclusively composed of these three monomers (Fig. 5a) and share similar statistical sequence features as indicated by Monte Carlo simulations (Supplementary Figs. 11 and 14).

Performance of super-adhesive hydrogels

We conducted detailed studies on the three top-performing ML-driven hydrogels (R1-max, R2-max and R3-max) and compared them with the best DM-driven hydrogel (G-max) (Fig. 5, Extended Data Fig. 6 and Supplementary Table 8). In their as-prepared state, all gels were transparent and exhibited frequency-independent storage moduli (G') (Extended Data Fig. 6a), indicating negligible inter- or intramolecular aggregation in DMSO. Despite compositional differences, comparable G' values suggest similar network topologies.

On equilibration in normal saline, all gels underwent shrinkage (Extended Data Fig. 6c). In contrast to G-max, the ML-driven hydrogels exhibited increased opacity (Fig. 5b), stronger viscoelasticity and higher moduli (Extended Data Fig. 6b). This suggests that their higher hydrophobic BA and aromatic PEA content (Fig. 5a) promotes strong associations of copolymer strands in aqueous media, which facilitate energy dissipation. Moreover, the ML-driven hydrogels exhibited greater mechanical strength and toughness (Supplementary Video 1), as evidenced by the larger area under their stress–strain curves (Fig. 5c). The enhanced viscoelasticity and toughness contributed to their improved adhesion compared with G-max[44].

To comprehensively evaluate adhesive performance, we conducted tack tests across a range of test conditions, substrates and solution media. Generally, F_a increased with increasing loading force and contact time, eventually reaching a plateau (Fig. 5d and Extended Data Fig. 7), attributed to enhanced interfacial contact and water drainage at the hydrogel–substrate interface. These plateau values were used to compare maximum adhesion performance across substrates and solutions.

In normal saline, R1-max achieved a maximum F_a exceeding 1 MPa on glass (Fig. 5e) and maintained robust adhesion over 200 attachment–detachment cycles (Extended Data Fig. 8). It also demonstrated strong adhesion to a variety of substrates, including inorganic materials, plastics and metals, as confirmed by lap shear and peeling tests (Extended Data Fig. 9). Notably, R1-max sustained joints of plates made from different materials under a 1-kg shear load for over 1 year, showcasing exceptional durability (Fig. 5f and Supplementary Fig. 15).

In artificial seawater (0.7 M NaCl), all three ML-driven hydrogels exhibited similar levels of strong adhesion (Fig. 5g). In deionized water, however, R2-max outperformed the others, exhibiting cavitation during debonding (Supplementary Fig. 16). These results indicate that small compositional variations can affect adhesion performance in different environments, reflecting a principle observed in nature—adaptability over universal optimization—in which biological systems evolve to perform optimally in their specific environments.

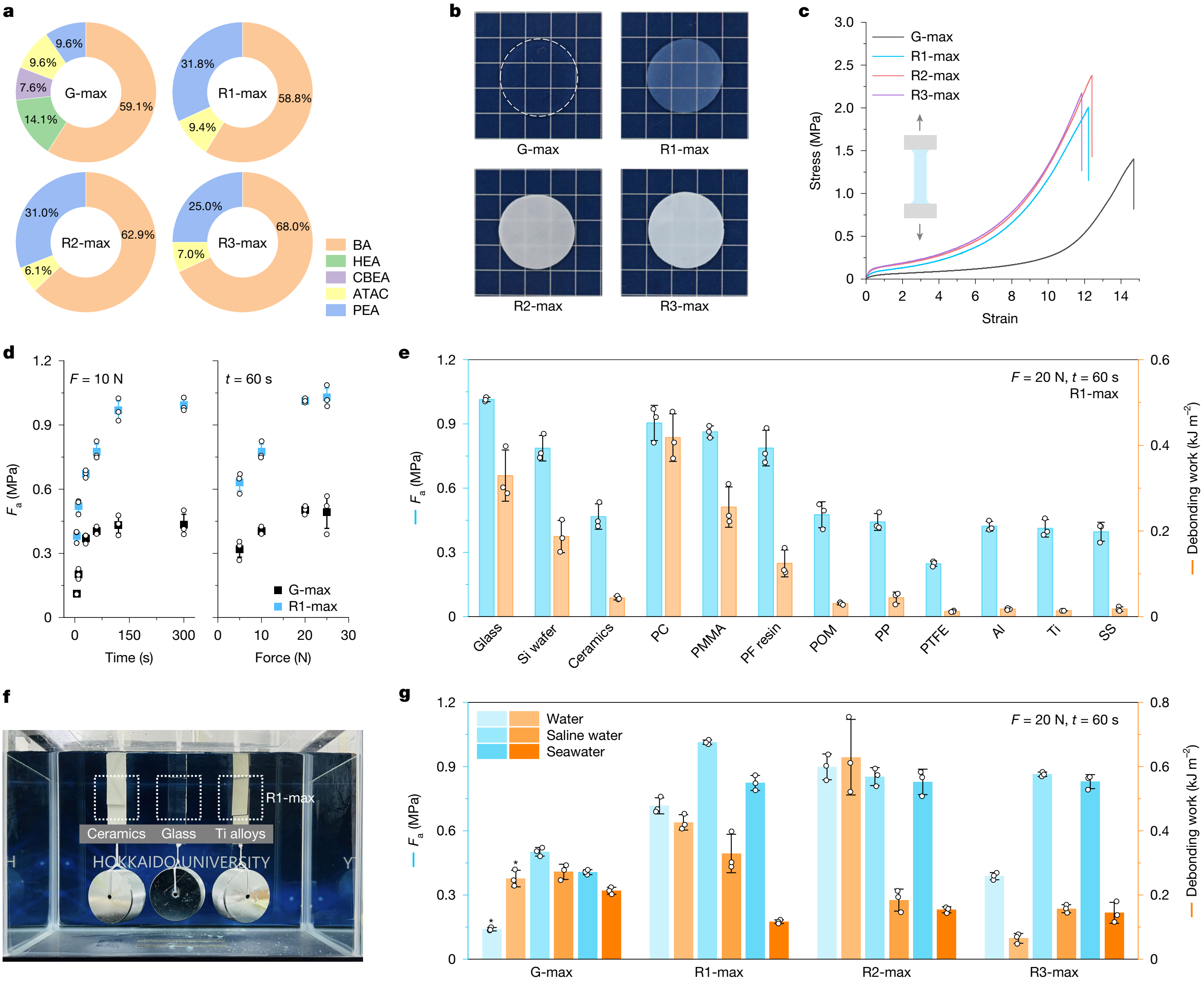


Fig. 5 | Characterization and performance of hydrogels identified by DM (G-max) and ML optimization (R1-max, R2-max and R3-max). **a**, Formulations of the gels. **b,** Photographic images showing the appearance of the gels. **c**, Uniaxial tensile stress–strain curves of the gels at a stretch rate of 100 mm min^{-1}. **d**, F_a of hydrogels as a function of contact time (left) and contact force (right) on glass in normal saline. **e**, F_a of R1-max on various substrates in normal saline. PC, polycarbonate; PMMA, poly(methyl methacrylate); PF, phenol formaldehyde; POM, polyoxymethylene; PP, polypropylene; PTFE, polytetrafluoroethylene; Al, aluminium alloy; Ti, titanium alloy; SS, stainless steel. **f**, Photographic image showing R1-max (25 mm × 25 mm in size, about 0.4 mm thickness) joining pairs of ceramics (left), glass (middle) and titanium (right) plates under a 1-kg load in normal saline for over 1 year. **g**, F_a on glass substrate in deionized water, normal saline and artificial seawater (0.7 M NaCl) for hydrogels equilibrated in the corresponding solutions. The asterisk on G-max indicates cohesive failure during testing. Error bars represent the standard deviation of N = 3 measurements.

This finding underscores the importance of ensuring data consistency in ML optimizations, as hydrogel performance varies with environmental conditions.

To demonstrate practical applicability, several case studies were conducted. R1-max was used to affix a rubber duck to a seaside rock (Extended Data Fig. 10a). Its strong adhesion in saltwater enabled the duck to withstand continuous ocean tides and wave impacts, revealing its suitability for harsh marine environments (Supplementary Video 2). R2-max, exhibiting the highest adhesion in deionized water (Fig. 5g), successfully sealed a 20-mm-diameter hole at the base of a 3-m-tall polycarbonate pipe filled with tap water (Extended Data Fig. 10b). It instantly stopped the high-pressure water leak (Supplementary Video 3), showcasing a level of performance that common adhesives cannot match (Extended Data Fig. 10c). Furthermore, all these hydrogels demonstrated good biocompatibility, as confirmed by subcutaneous implantation in mice (Supplementary Fig. 17), supporting their potential for biomedical applications.

In summary, we introduced a data-driven approach that integrates the extraction of valuable sequence information from proteins, scalable polymer synthesis and iterative ML to address longstanding challenges in the de novo design and development of soft materials. Beyond adhesive hydrogels, this data-driven design framework offers a systematic, scalable end-to-end approach for developing a wide range of functional soft materials. However, challenges remain, primarily because of limitations in monomer diversity, polymer synthesis technologies for controlling monomer sequences to a scale suitable for materials development and dataset scalability. Overcoming these challenges will require expanding modular monomer libraries, advancing polymerization techniques and developing physics-informed ML models that can generalize across sparse, multiscale datasets.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-09269-4.

1. Zeni, C. et al. A generative model for inorganic materials design. *Nature* **639**, 624–632 (2025).
2. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
3. Yao, Y. et al. High-entropy nanoparticles: synthesis-structure-property relationships and data-driven discovery. *Science* **376**, eabn3103 (2022).
4. Li, F. et al. Design of self-assembly dipeptide hydrogels and machine learning via their chemical features. *Proc. Natl Acad. Sci. USA* **116**, 11259–11264 (2019).
5. Xu, T. et al. Accelerating the prediction and discovery of peptide hydrogels with human-in-the-loop. *Nat. Commun.* **14**, 3880 (2023).
6. Tamasi, M. J. et al. Machine learning on a robotic platform for the design of polymer-protein hybrids. *Adv. Mater.* **34**, 2201809 (2022).
7. Fan, H. L. & Gong, J. P. Fabrication of bioinspired hydrogels: challenges and opportunities. *Macromolecules* **53**, 2769–2782 (2020).
8. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
9. Pollice, R. et al. Data-driven strategies for accelerated materials design. *Acc. Chem. Res.* **54**, 849–860 (2021).
10. Ferguson, A. L. & Brown, K. A. Data-driven design and autonomous experimentation in soft and biological materials engineering. *Annu. Rev. Chem. Biomol. Eng.* **13**, 25–44 (2022).
11. Jackson, N. E., Webb, M. A. & de Pablo, J. J. Recent advances in machine learning towards multiscale soft materials design. *Curr. Opin. Chem. Eng.* **23**, 106–114 (2019).
12. Li, Z. et al. AI energized hydrogel design, optimization and application in biomedicine. *Mater. Today Bio* **25**, 101014 (2024).
13. McDonald, S. M. et al. Applied machine learning as a driver for polymeric biomaterials design. *Nat. Commun.* **14**, 4838 (2023).
14. Gormley, A. J. & Webb, M. A. Machine learning in combinatorial polymer chemistry. *Nat. Rev. Mater.* **6**, 642–644 (2021).
15. Ferguson, A. L. Machine learning and data science in soft materials engineering. *J. Phys. Condens. Matter* **30**, 043002 (2018).
16. Fan, H. L. & Gong, J. P. Bioinspired underwater adhesives. *Adv. Mater.* **33**, 2102983 (2021).
17. Narayanan, A., Dhinojwala, A. & Joy, A. Design principles for creating synthetic underwater adhesives. *Chem. Soc. Rev.* **50**, 13321–13345 (2021).
18. Chen, Y. et al. Bioinspired multiscale wet adhesive surfaces: structures and controlled adhesion. *Adv. Funct. Mater.* **30**, 1905287 (2020).
19. Ruan, Z. et al. Population-based heteropolymer design to mimic protein mixtures. *Nature* **615**, 251–258 (2023).
20. Kim, Y. et al. Designing directional adhesive pillars using deep learning-based optimization, 3D printing, and testing. *Mech. Mater.* **185**, 104778 (2023).
21. Boesel, L. F., Greiner, C., Arzt, E. & del Campo, A. Gecko-inspired surfaces: a path to strong and reversible dry adhesives. *Adv. Mater.* **22**, 2125–2137 (2010).
22. Lee, B. P., Messersmith, P. B., Israelachvili, J. N. & Waite, J. H. Mussel-inspired adhesives and coatings. *Annu. Rev. Mater. Res.* **41**, 99–132 (2011).
23. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135–145 (2018).
24. Porebski, B. T. & Buckle, A. M. Consensus protein design. *Protein Eng. Des. Sel.* **29**, 245–251 (2016).
25. Chang, M. P., Huang, W. & Mai, D. J. Monomer-scale design of functional protein polymers using consensus repeat sequences. *J. Polym. Sci.* **59**, 2644–2664 (2021).
26. Jacob, J., Duclohier, H. & Cafiso, D. S. The role of proline and glycine in determining the backbone flexibility of a channel-forming peptide. *Biophys. J.* **76**, 1367–1376 (1999).
27. Fan, H. L. et al. Adjacent cationic-aromatic sequences yield strong electrostatic adhesion of hydrogels in seawater. *Nat. Commun.* **10**, 5127 (2019).
28. Panganiban, B. et al. Random heteropolymers preserve protein function in foreign environments. *Science* **359**, 1239–1243 (2018).
29. Jiang, T. et al. Single-chain heteropolymers transport protons selectively and rapidly. *Nature* **577**, 216–220 (2020).
30. Smith, A. A. A., Hall, A., Wu, V. & Xu, T. Practical prediction of heteropolymer composition and drift. *ACS Macro Lett.* **8**, 36–40 (2019).
31. Kendall, M. G. Rank and product-moment correlation. *Biometrika* **36**, 177–193 (1949).
32. Myers, R. H., Montgomery, D. C. & Anderson-Cook, C. M. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments* (Wiley, 2016).
33. Hutter, F., Hoos, H. H. & Leyton-Brown, K. Sequential model-based optimization for general algorithm configuration. In *Proc. International Conference on Learning and Intelligent Optimization* 507–523 (Springer).
34. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
35. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
36. Ginsbourger, D., Le Riche, R. & Carraro, L. A multi-points criterion for deterministic parallel global optimization based on Gaussian processes. In *Proc. Computational Intelligence in Expensive Optimization Problems* (eds Tenne Y. & Goh C. K.) 131–162 (Springer, 2010).
37. González, J., Dai, Z., Hennig, P. & Lawrence, N. Batch Bayesian optimization via local penalization. In *Proc. Artificial Intelligence and Statistics* 648–657 (PMLR, 2016).
38. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
39. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).
40. Fan, H. L., Cai, Y. R. & Gong, J. P. Facile tuning of hydrogel properties by manipulating cationic-aromatic monomer sequences. *Sci. China Chem.* **64**, 1560–1568 (2021).
41. Jin, Z. P. et al. Gluing blood into gel by electrostatic interaction using a water-soluble polymer as an embolic agent. *Proc. Natl Acad. Sci. USA* **119**, e2206685119 (2022).
42. Ou, X. et al. Structure and sequence features of mussel adhesive protein lead to its salt-tolerant adhesion ability. *Sci. Adv.* **6**, eabb7620 (2020).
43. Chang, H. et al. Short-sequence superadhesive peptides with topologically enhanced cation−π interactions. *Chem. Mater.* **33**, 5168–5176 (2021).
44. Creton, C. & Ciccotti, M. Fracture and adhesion of soft materials: a review. *Rep. Prog. Phys.* **79**, 046601 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025

Methods

Hydrogel fabrication

All copolymer gels were synthesized by one-step free-radical copolymerization of monomers with a chemical crosslinker. The crosslinker concentration was fixed at 0.1 mol% relative to the total monomer content to balance the elasticity and deformability of the gels[27]. DMSO solutions containing functional monomers (total concentration of 2.4 M) with compositions derived from DM and ML (Supplementary Tables 2 and 7), chemical crosslinker (glycerol 1,3-diglycerolate diacrylate, 2.4 mM), and UV initiator (2-oxoglutaric acid, 6 mM) were used. For example, to prepare the G-max gel, 1.819 g of BA, 0.413 g of HEA, 0.264 g of CBEA, 0.561 g of ATAC, 0.441 g of PEA, 8.4 mg of glycerol 1,3-diglycerolate diacrylate and 8.8 mg of 2-oxoglutaric acid were added to a 10 ml volumetric flask, followed by DMSO to reach 10 ml. The precursor solution was transferred to a glove box to remove oxygen, poured into a reaction cell (two 10 cm × 10 cm glass plates, 0.5-mm spacing) and irradiated with UV light (365 nm wavelength, 4 mW cm^{-2} intensity) for 8 h to form gels (Supplementary Fig. 9a). After UV irradiation, over 99% of the monomers were converted into polymers, as confirmed by NMR (Supplementary Fig. 9b).

The as-prepared organogels were then immersed in normal saline (0.154 M NaCl) to remove solvent and residual chemicals, with the saline exchanged every 12 h for at least 2 weeks until swelling equilibrium was reached. Hydrogels were stored in normal saline before use.

Underwater adhesion characterization

The tack test was conducted using a SHIMADZU tester (Autograph AG-X) equipped with Trapezium X software. Hydrogel (0.3–0.8 mm thickness) at swelling equilibrium was adhered to the probe using cyanoacrylate adhesive (super glue). For rapid screening, DM-driven hydrogels from the training round and ML-driven hydrogels from three optimization rounds, were prepared as 15 mm diameter samples. For detailed adhesion studies, 10 mm diameter samples were used to avoid exceeding the force range of the instrument. This change in diameter did not affect the adhesive strength results. The hydrogel on the probe was then immersed in a test solution (for example, normal saline) for 5 min to reach equilibrium. The probe descended towards the substrate at 1 mm min^{-1} until a loading force of 10 N was applied, maintained for 10 s and withdrawn at 10 mm min^{-1} (Supplementary Fig. 10). These test conditions were used as a standard protocol unless otherwise specified. For repeated adhesion tests, hydrogels rested underwater for 5 min between cycles, with glass substrates replaced every 100 tests. For prolonged attachment–detachment cycles (Extended Data Fig. 8), a 5 N loading force and a 10 s contact time were used to minimize gel fatigue. Each sample was tested at least three times. For hydrogel dataset construction, the highest adhesive strength recorded for each sample was reported as F_a, representing maximum adhesion performance under the specific conditions.

Lap shear adhesive strength was measured using a universal testing machine (UTM, INSTRON 5965). A hydrogel (10 mm diameter, area $A = 78.5$ mm^2) at swelling equilibrium was sandwiched between two glass slides, pressed at 20 N for 1 min in normal saline. Shear loading was applied at 50 mm min^{-1}. Shear adhesive strength (F_a) was calculated as $F_a = F_{max}/A$, where F_{max} is the maximum loading force. For adhesion durability tests (Supplementary Fig. 15), the sandwiched assembly was stored in normal saline for varying durations before testing.

Interfacial toughness was measured by 180° peeling tests using INSTRON 5965. Hydrogel strips (10 mm × 150 mm) were adhered to a glass substrate in normal saline using mild finger pressure, followed by a 2 kg hand roller applied in each direction for 1 min to ensure uniform contact. Polyethylene terephthalate (PET) films (50 µm thickness) served as a stiff backing. Peeling tests were conducted at 50 mm min^{-1}. Interfacial toughness (G_c) was calculated as $G_c = 2F_c/w$, where F_c is the plateau force and w is the sample width (10 mm).

DM of adhesive proteins

A comprehensive dataset of adhesive proteins was compiled from the NCBI protein database, using 'adhesive proteins' as the query keyword. A total of 24,707 protein sequences from 3,822 different organisms (bacteria, viruses, eukaryotes and animals) were collected without additional data cleaning. Based on taxonomy annotations, proteins were grouped by species, and a consensus sequence was generated for each species to capture common sequence patterns and reduce the influence of individual variations.

The dataset included 3,111 species, noting that taxonomic overlap results in protein counts not summing to 24,707. For robust analysis, the top 200 species, ranked by the number of distinct proteins identified per species, were selected for further study.

Protein sequences were exported in FASTA format[45] using the Bio.SeqIO interface in BioPython[46]. Consensus sequences were computed with Clustal Omega[23], which performs multiple sequence alignment by generating a distance matrix from pairwise alignments, constructing a guide tree based on evolutionary relationships and progressively aligning sequences from the closest to the most distant. The resulting alignment identifies the most frequent residues at each position, yielding a consensus sequence that highlights conserved regions.

Clustal Omega was executed with the command:

```
./clustalo -i "input_file" --outfmt=clu -o "output_aln_file" -v
```

where "input_file" and "output_aln_file" denote the input protein sequences and output consensus sequences, respectively. The 200 consensus sequences generated were used for subsequent sequence analysis and hydrogel formulation design.

ML methods

A six-dimensional feature vector, $\phi_i = [\phi_{\mathrm{BA}}, \phi_{\mathrm{HEA}}, \phi_{\mathrm{CBEA}}, \phi_{\mathrm{ATAC}}, \phi_{\mathrm{AAm}}, \phi_{\mathrm{PEA}}]$, was used to represent monomer proportions in hydrogels. The target variable was adhesive strength, F_a. To model the relationship between ϕ_i and F_a, we explored both linear and non-linear ML models (Supplementary Tables 5 and 6).

Linear models included least absolute shrinkage and selection operator regression (Lasso) and ridge regression (Ridge). Non-linear models comprised *k*-nearest neighbours (KNN), kernel ridge regression (KRR), support vector regression (SVR), random forest regression (RFR), gradient boosting regression with XGBoost (XGB), extra trees regression (ETR) and Gaussian process (GP) with a Matérn kernel[32,34]. These non-linear models encompass non-parametric (KNN), kernel-based (KRR, SVR and GP) and tree-ensemble (RFR, XGB and ETR) approaches, enabling a comprehensive comparison[34,35,47].

XGB was of v.1.6.2, whereas the other models were implemented using Scikit-learn (v.1.0.2) and Scikit-optimize (v.0.9.0). The hyperparameter n_estimators was tuned using Optuna[48], whereas others were optimized using grid search (Supplementary Table 6). A 10-fold cross-validation strategy was used to assess predictive performance on our dataset of 180 hydrogels, using root mean squared error (RMSE) as the metric. GP and RFR, with the lowest RMSE in training-test error using a 90%/10% train/test split (Extended Data Fig. 4), emerged as the top performer and runner-up, respectively, and were subsequently used as the base (surrogate) models.

To make extrapolative predictions, we tried three types of methods.

1. Exploitation-only enumeration:
   - GP_enu: random sampling in the input space using the fitted GP model.
   - RFR_enu: random sampling in the input space using the fitted RFR model.
   - Ten million ϕ_i vectors were generated from a uniform distribution [0, 1.0) for each monomer, normalized to sum to 1.0. The top five

  vectors, ranked by predicted F_a from each model, were experimentally validated.

2. Batched BO:
   - GP_KB: used GP predictions as the hypothetical values for selecting the next data points maximizing EI.
   - GP_CLmax: used the maximum F_a (y_max) from the training set as a hypothetical value for selecting the next data points with EI maximums.
   - GP_CLmin: used the minimum F_a (y_min) for selecting the next data points with EI maximums.
   - GP_LP: incorporated a locally penalized term in EI calculation[37].
   - GP_KB, GP_CLmax and GP_CLmin simplified the joint q-EI probability calculation[36] by using the GP prediction value as a hypothetical value for selecting the next data points with EI maximums. A batch size of $q = 10$ was selected.
3. Batched sequential model-based optimization (SMBO):
   - GP-RFR: GP as the hypothetical value provider and RFR as the EI maximizer.
   - RFR-RFR: RFR as both the hypothetical value provider and the EI maximizer.
   - RFR-GP: RFR as the hypothetical value provider and GP as the EI maximizer.
   - RFR-GP*: RFR-GP with a warm start, 10 RFR-generated points were added to the real dataset for GP regression.
   - RFR-ETR: RFR as the hypothetical value provider and ETR as the EI maximizer.
   - RFR-GBM: RFR as the hypothetical value provider and GBM as the EI maximizer.
   - SMBO iteratively updates the surrogate model while exploring promising data points[33]. GP and RFR, when used as the hypothetical value providers, balance exploitation and exploration, whereas GP_CLmax and GP_CLmin emphasize exploitation and exploration, respectively[49].

SMBO (Supplementary Algorithm 1) consists of four components: the true function (f), global domain (X), acquisition function (S) and surrogate model (M). Initial training data (D) are sampled from X, and experimental F_a values are obtained (line 1). The surrogate model M is fitted to D (line 3) and S (EI) identifies the next data point based on predictive uncertainty (line 4). This data point is subsequently validated experimentally (line 5), updating D (line 6) for T iterations (line 2).

EI quantifies expected improvement, $\int_{y*}^{\infty}(y-y^*)p(y)\mathrm{d}y$, over the current best target (y^*). Owing to the time-intensive nature of hydrogel fabrication (each takes about 2 weeks), GP and RFR were used as the hypothetical value providers, enabling the maximization of the joint q-EI probability without requiring new experiments per iteration. EI maximizers (GP, RFR, ETR and GBM) used hyperparameters from Scikit-optimize (v.0.9.0).

For GP as the EI maximizer, the limited-memory Broyden–Fletcher–Goldfarb–Shannon (L-BFGS-B) algorithm[50] was executed 20 times per iteration (40 iterations total) to identify the point with the highest EI, updating the GP prior. For the other three EI maximizers (RFR, ETR and GBM), 10,000 points were randomly sampled per iteration, as numerical optimization is more suitable for tree-ensemble models lacking gradient information. SMBO ran for 40 iterations with each EI maximizer, selecting two sets of 10 data points in each iteration: the top 10 ranked by EI values (batch size $q = 10$), and the top 10 ranked by predicted F_a values for experimental validation. These two sets may overlap, and the total number of data points may be less than 20.

For BO methods (GP_KB, GP_CLmax, GP_CLmin and GP_LP), the procedure was similar, except that the hypothetical value provider was either GP itself (GP_KB and GP_LP) or constant values (y_max for GP_CLmax and y_min for GP_CLmin).

After the first round, 109 validated points expanded the dataset to 289 hydrogels. The second and third rounds added 27 and 25 points, respectively, resulting in a final dataset comprising 341 hydrogels.

Data availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Information. The data that support the findings of this study are available online at GitHub (https://github.com/sheng-hu/hydrogels).

Code availability

ML algorithms and Python codes that support the findings of this study are available online at GitHub (https://github.com/sheng-hu/hydrogels).

45. Pearson, W. R. Finding protein and nucleotide similarities with FASTA. *Curr. Protoc. Bioinformatics* **53**, 3.9.1–3.9.25 (2016).
46. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
47. Jones, D. R., Schonlau, M. & Welch, W. J. Efficient global optimization of expensive black-box functions. *J. Glob. Optim.* **13**, 455–492 (1998).
48. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: a next-generation hyperparameter optimization framework. In *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2623–2631 (Association for Computing Machinery, 2019).
49. Chevalier, C. & Ginsbourger, D. Fast computation of the multi-points expected improvement with applications in batch selection. In *Proc. Learning and Intelligent Optimization* (eds Nicosia, G. & Pardalos, P.) 59–69 (Springer, 2013).
50. Zhu, C., Byrd, R. H., Lu, P. & Nocedal, J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.* **23**, 550–560 (1997).

Acknowledgements This research was supported by the JSPS KAKENHI grants (JP21K17745, JP21K14676, JP22K21342, JP22H04968 and JP24K17728). The Institute for Chemical Reaction Design and Discovery (WPI-ICReDD) was established by the World Premier International Research Initiative (WPI), MEXT, Japan. We thank Y. Katsuyama for the preparation of the experimental devices and MT AquaPolymer and Osaka Organic Chemical for providing the monomers.

Author contributions H.F., S.H., H.L., W.L., I.T. and J.P.G. conceived the presented idea. S.H. and H.F. performed the DM. H.L., H.F. and J.P.G. designed the experiments. H.L. and H.F. performed the experiments. S.H. and I.T. designed the ML strategy. S.H. implemented ML. W.L. carried out the simulations. L.W. and S.T. performed the biological experiments. H.L., H.F., S.H., W.L., H.Y. and J.P.G. contributed to data preparation, analysis and manuscript drafting with inputs from all authors. I.T., H.F. and J.P.G. provided supervision and resources for this study. All authors discussed the results and contributed to the final paper.

Competing interests H.L., S.H., I.T., W.L., H.F. and J.P.G. are inventors of a patent application (2024-134812) titled 'Random copolymers and adhesives' submitted by Hokkaido University, which covers the composition of underwater adhesive materials in this study.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-09269-4.

Correspondence and requests for materials should be addressed to Ichigaku Takigawa, Wei Li, Hailong Fan or Jian Ping Gong.

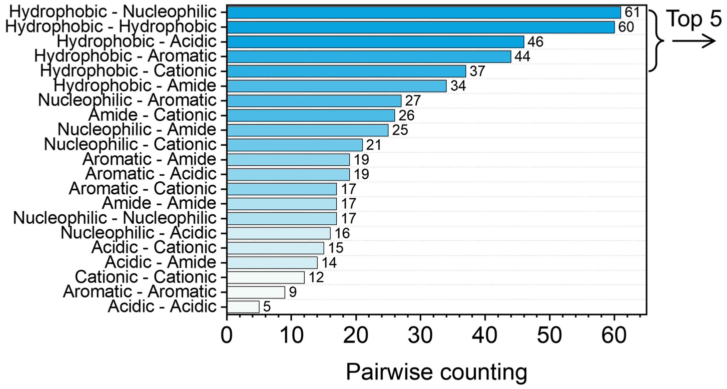
Peer review information *Nature* thanks Laura Russo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at http://www.nature.com/reprints.

a

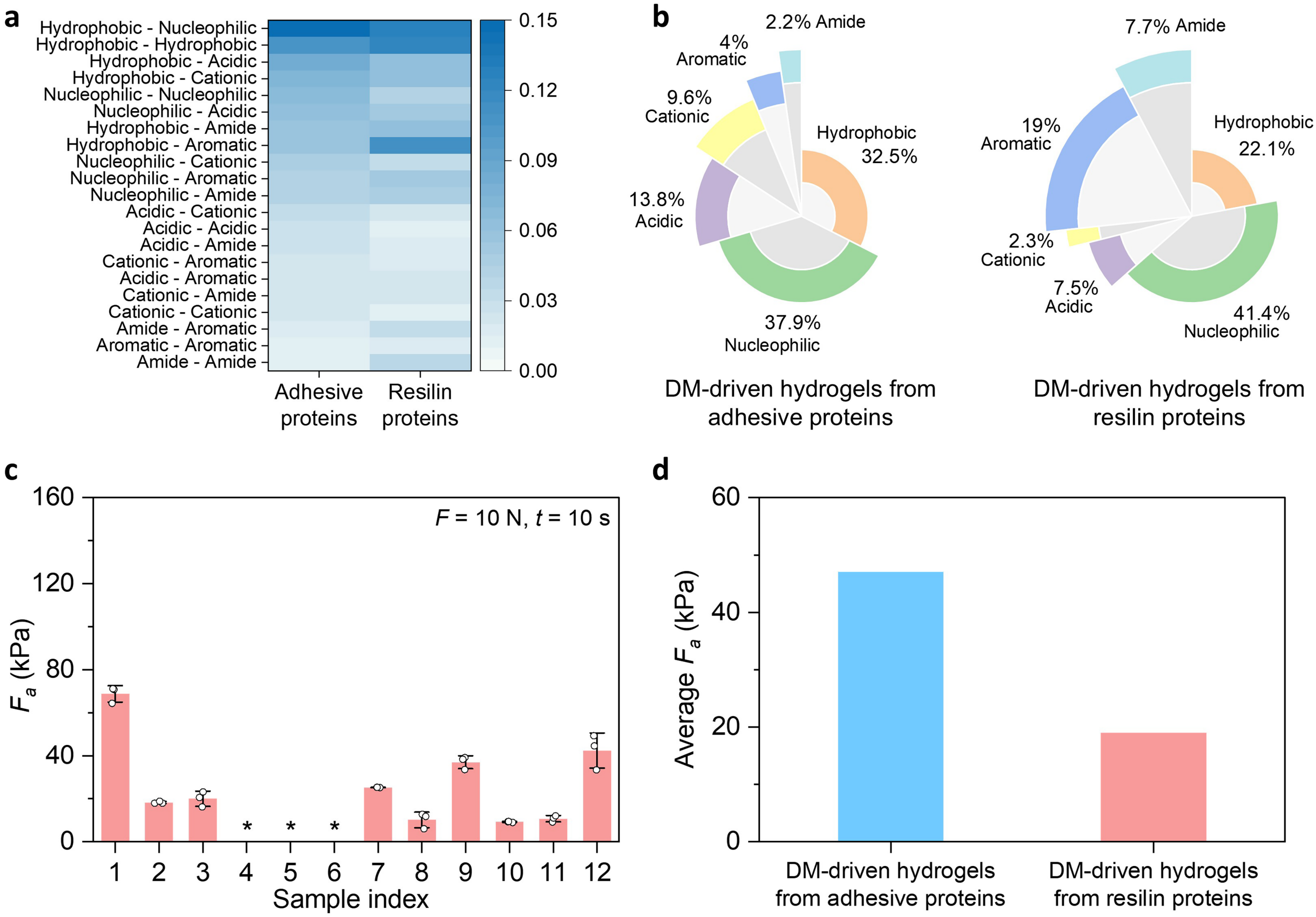
```
MNRIYSLRYSAVARGFIAVSEMXXXXXXXXXXXXXXXXXXXXXXXLLSXXXXXXXXAXEXNXXXGXXXXXXFAXXXGXXXXXXXXXXXYXXXGXXXX
XXXXXAMPDFSAVDSEIGVATLXXGQYXXDVXVNGXXXXXSXGXXXNRAXLXIXXXXPXLSXXXXXXXXXXXXXXXXXXXXXXXXAXMSEYMKNEIL
EFLNRHNGGKTAEIAEALXXXXXPXXGAXXLXXXXXXAVTDYQARYYLLLLEKXGXVXXXXXXXXXXXXXGLXLDYXXXXXXXXXXXXGDQRSPXXX
XXXXLTLRRGMATYWXLKGEXQAGQXCSSXXTTXXXIPLDMXXXXXXXXXXXXXXFXXXXXXXXTWXXXSSXXXXXXTQGTTTYAMHGQQGNDLNAG
KNLIFQGQNGQXXLXXAIXXXXXXLTLGXNXXXXTXXFXXWXXXGXXLXSXXXXXXXXXGXXXXXXXXKIXLGTLXXXXTGXXXGXXXXXXXRVVVX
QQGXXXXXXXXXXXXXXXXXXXFXXXXXXXXXXXXXXXXXXWXXXXXXLXVXXXGXXXXXHRIKXTPTXXXXXNNXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXFXXXXXXXXLGXXXXGXXXQXXXXXXXXXXXXXXXQXNCNXXXXXXXXXXXXDVXNXXXLFXGXXXGXLXXXXRLPEGX
XXXXLXSXXXDIXXXYXQXSXNXKXXXIXXXXXXXXXXXXXXYXXLXXXGXXXXXXXXXXXXXXXXXXXXXXXXXXXXNXXXXXXGXXXXXNSARVR
XXXSXXXXXXXXXXXXXGGMARXXXXXXXXXXXGXXXXXXXXXXXXXXXXYYQXGRXDRXXXXXXXXGXXXXXXXXXXXIXXXXXGXXVXXXXXXXX
XXXXXXXXXXXXXXXXLAXXXGXRXGTTQSXXXXXXXXXXXXXXDINELKXXXXXPXXXXXTXFXXXXXXXXXXXXXXWXXXXXXXXXXXXXLXXXX
XXXXXXXIXEXXXXXXEXXXPXXXXXAEXXXXXXXXXXXXXXLALLVNEMPGXXWDVXXXXGXIXNXXXRYXXXQXNHKSXXNXXXXXPXXXXXXXQ
XXXXXXXXXQXXXXXXXXXXXXXXXXXXXXXXXXXXXLXYEKEGLTNXXXXXXXXXGXXXXXXXXNMSLRKLLTLFIVSXXXXXXXXXLMALGTTSS
XXXXXXXXXXXXXXDMVVXXXXXPPDLPVGSVILTRXWXXXXPGGASXYSLGXVNSYXXXXGXXNKIXXVVITQRPQFITSWRPGDXXGXXSASXXI
ATVTWNQCNGPXFADGSWAYYREYIAWVVFPKKVMTXNGYPLFIEVHNKGXXXXXXXXXCPXXXXXXXXXXXXXXXXXKXYXXXERAFDNGXXADNV
XXXXLXXXXXXXXXXXXRFXXXXXXXXXXXXXXXXXXXXGEXXXLXXXXXXXXFXXXXXXXXXXKGDYSVXIPYXGXVXXXXXXXXTXXGXXXXELL
NSLAAVKSGXKAKRAQRPAXXXXGKXXXXXSYXXXXXXXPXXEXKXXXXXXXLQIALXVAXXXXYXXXXXXXXXXXXXXXXXGXXXXXXXXSXXXXI
XXXXXXXNXXGXKXFNIXXXANXXXXXXXXXXXXXXXXXXFLDEXXXXXXXXXXXXXXAXXXXXXXXXXXXXXXXXXXXXXXXXLXXXXXXXXXXXX
XXXXRNTTPXXXXXXFXXXXXCXVXAXXXXXXXXXXXXXGXXIXXXXXXXXXXXXXXXXTYSHGKKFSVGLXXGWDSIVXXXGXLVDVXDPLQFNXX
YTXGXLXXQXXPXXXXRXETRYITXXXXXRXYXXGTQNLTIGSRLYGXXXXXXXGXIESSKIQPGVLSGXXXXXXMIXPLTXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXNXXXXLXGXXXXXXXXXXXXXXXXXXXXXXXXXXXXGXXXXXXXXXXXXXEQXXXXXXPXXXLXXXXXEXXXXXXXXXXEXXX
XLFGLSVXRLXXXXXFXXXXEFGXXXXXXXXXGXXXXXXLXXXXXDXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXRXXXXXXXEGVEXXXXXXXXX
XXXXXXXQGXGEXXADXXPXPXXXXXXXXXXXXXXXXRHNGXIXXXMPXX
```

b

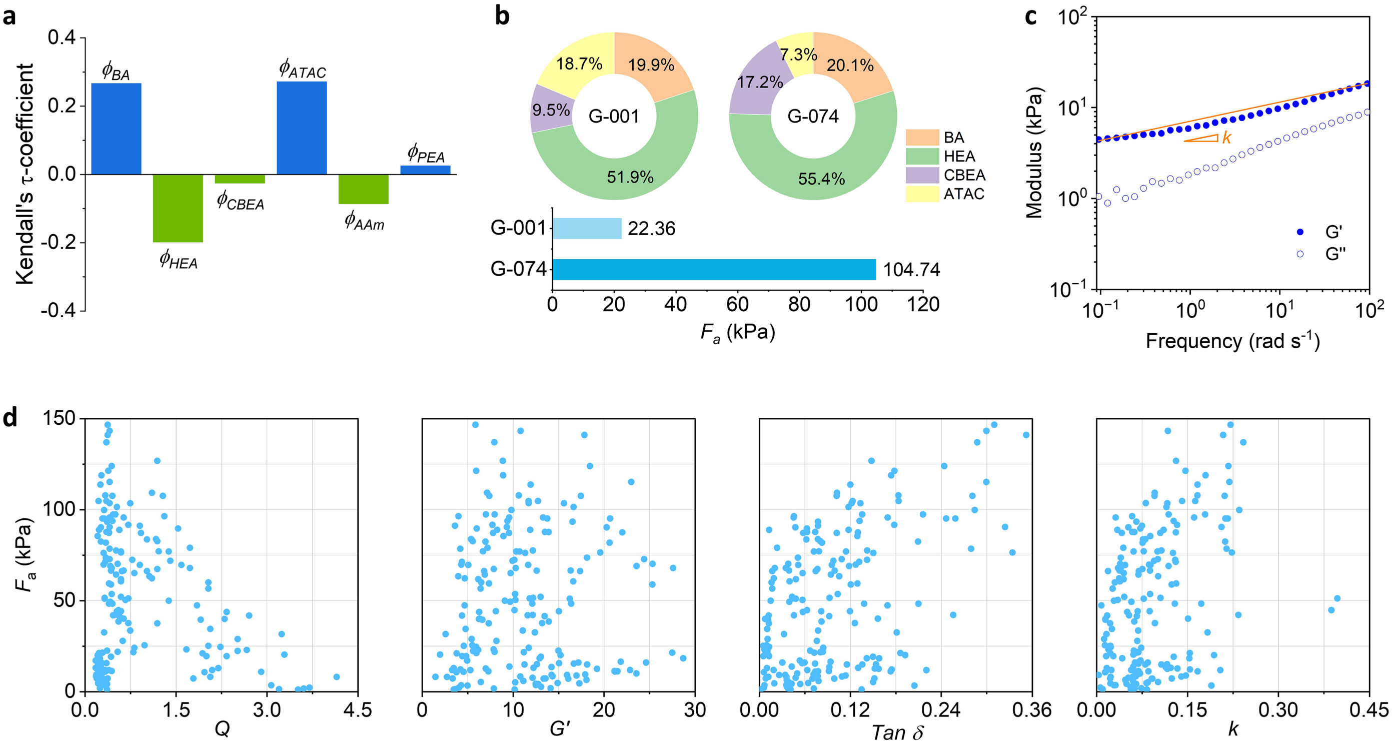


Functional class	Counting number $N_i = \sum_j (n_{ij} + n_{ji})$	Relative composition $\boldsymbol{\phi_i = N_i / \sum_i N_i}$
Hydrophobic	308	0.621
Nucleophilic	61	0.123
Acidic	46	0.093
Aromatic	44	0.089
Cationic	37	0.074
Amide	0	0

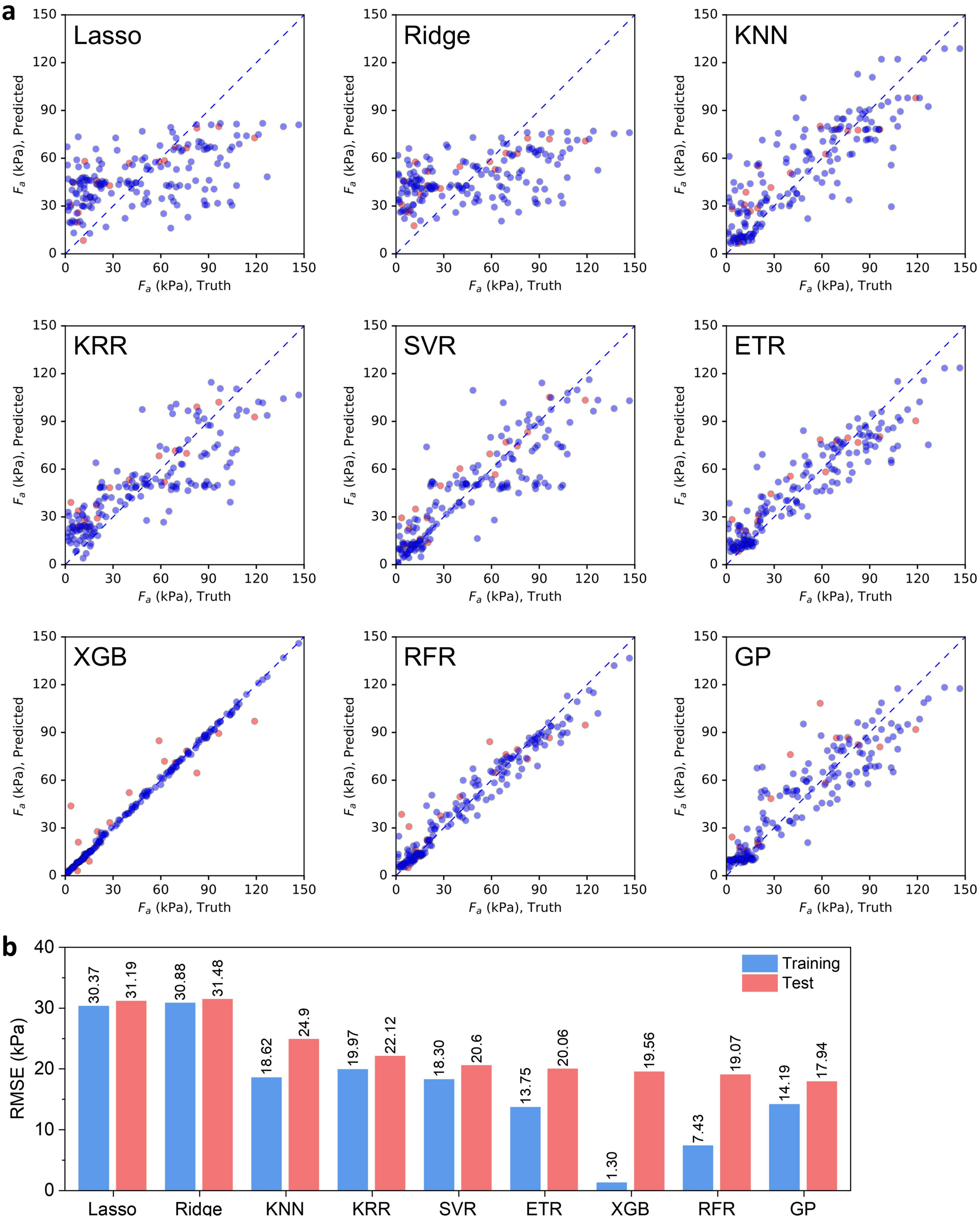
Extended Data Fig. 1 | Sequence analysis of Enterobacteriaceae adhesive proteins. (a) Consensus sequence fragment. (b) Pairwise functional class counts within the consensus sequence fragment. Complete data for the top 200 species are provided in Supplementary Data 1 and Supplementary Data 2.



Extended Data Fig. 2 | Data mining-driven hydrogels based on the resilin protein database. (a) Pairwise frequency distribution of 21 functional class pair types in encoded consensus sequences for adhesive and resilin proteins from data mining (DM). The resilin dataset comprises 2,537 proteins sourced from the NCBI protein database using the keyword "resilin." (b) Average monomer proportions in formulations derived from adhesive and resilin protein databases. (c) Adhesive strength (F_a) of DM-driven hydrogels derived from the resilin protein database. Asterisks indicate significant gel shrinkage during solvent exchange, making adhesion testing unfeasible for these samples. Detailed formulations are included in Supplementary Table 7. Adhesion tests were performed on a glass substrate in normal saline using a tack test with a 10 N loading force and a 10-s contact time, consistent with conditions used for DM-driven hydrogels derived from adhesive proteins. Error bars represent the standard deviation of $N = 3$ measurements. (d) Comparison of average F_a between DM-driven hydrogels derived from adhesive and resilin protein databases.

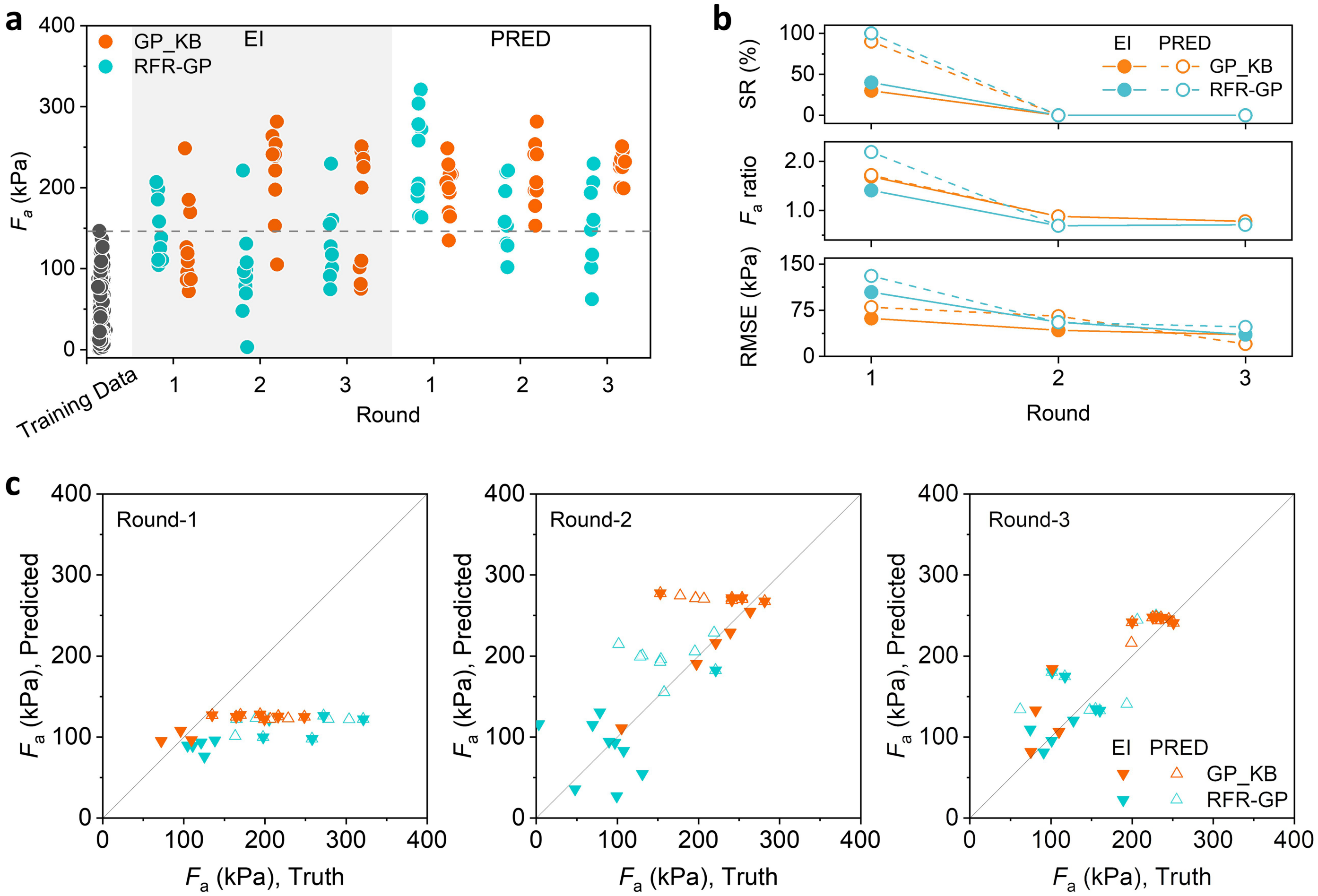


Extended Data Fig. 3 | Analysis of correlations between adhesive strength (F_a) and properties of 180 bioinspired hydrogels. (a) Correlation between monomer proportions (ϕ_i) and F_a, captured by Kendall's τ coefficients. These coefficients reveal that ϕ_{BA}, ϕ_{ATAC}, and ϕ_{PEA} have weak positive correlations with F_a, while ϕ_{HEA}, ϕ_{AAm}, and ϕ_{CBEA} show weak negative correlations. (b) Example illustrating the complex interplay between ϕ_i and F_a, due to the synergistic effects of different monomers. Comparing G-074 to G-001, ϕ_{BA} is roughly the same, ϕ_{HEA} and ϕ_{CBEA} increase, and ϕ_{ATAC} decreases. Despite Kendall's τ coefficients suggesting that the F_a of G-074 should be lower than that of G-001, the actual F_a of G-074 is about five times higher. (c) Angular frequency dependence of the storage modulus (G') and loss modulus (G'') of the G-042 hydrogel as an example. The slope (k) of G' is calculated from the line connecting G' values at frequencies of 0.1 and 100 rad s^{-1}. A larger slope indicates greater viscoelasticity of the hydrogel. (d) F_a as a function of network properties for the 180 bioinspired hydrogels. Q, G', $Tan\,\delta = G''/G'$, and k represent the volume swelling ratio, storage modulus at a frequency of 10^0 rad s^{-1}, loss factor at a frequency of 10^0 rad s^{-1}, and the slope of the G' curve, respectively. These results suggest that hydrogels with shrinking behavior ($Q < 1$), moderate G', high $Tan\,\delta$, and moderate k tend to exhibit higher F_a values. All characterizations were performed on gels equilibrated in normal saline (0.154 M NaCl). Adhesion tests were conducted using a tack test with a 10 N loading force and a 10-s contact time on a glass substrate to enable rapid screening and ensure consistent comparisons across the dataset.

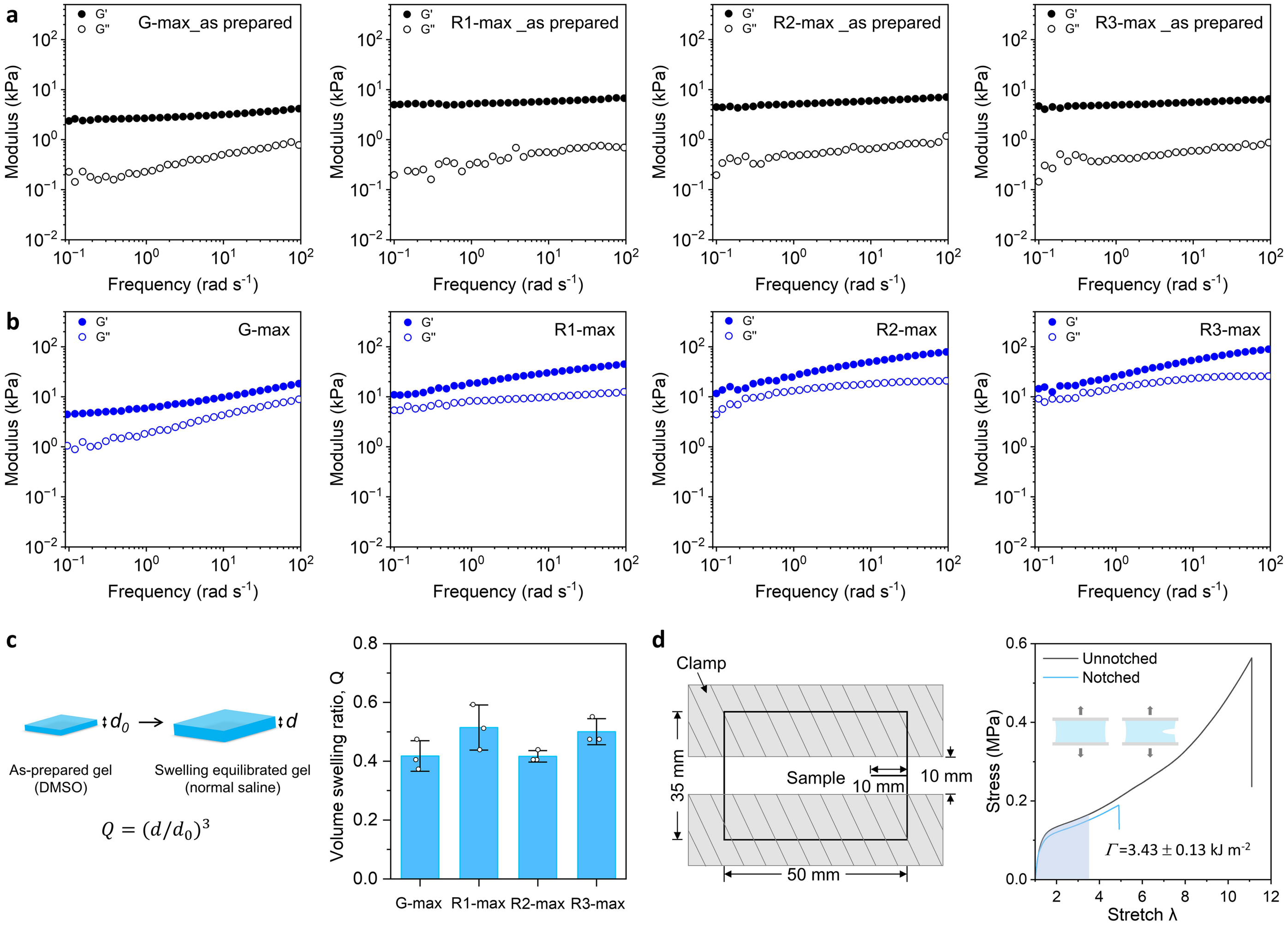


Extended Data Fig. 4 | See next page for caption.

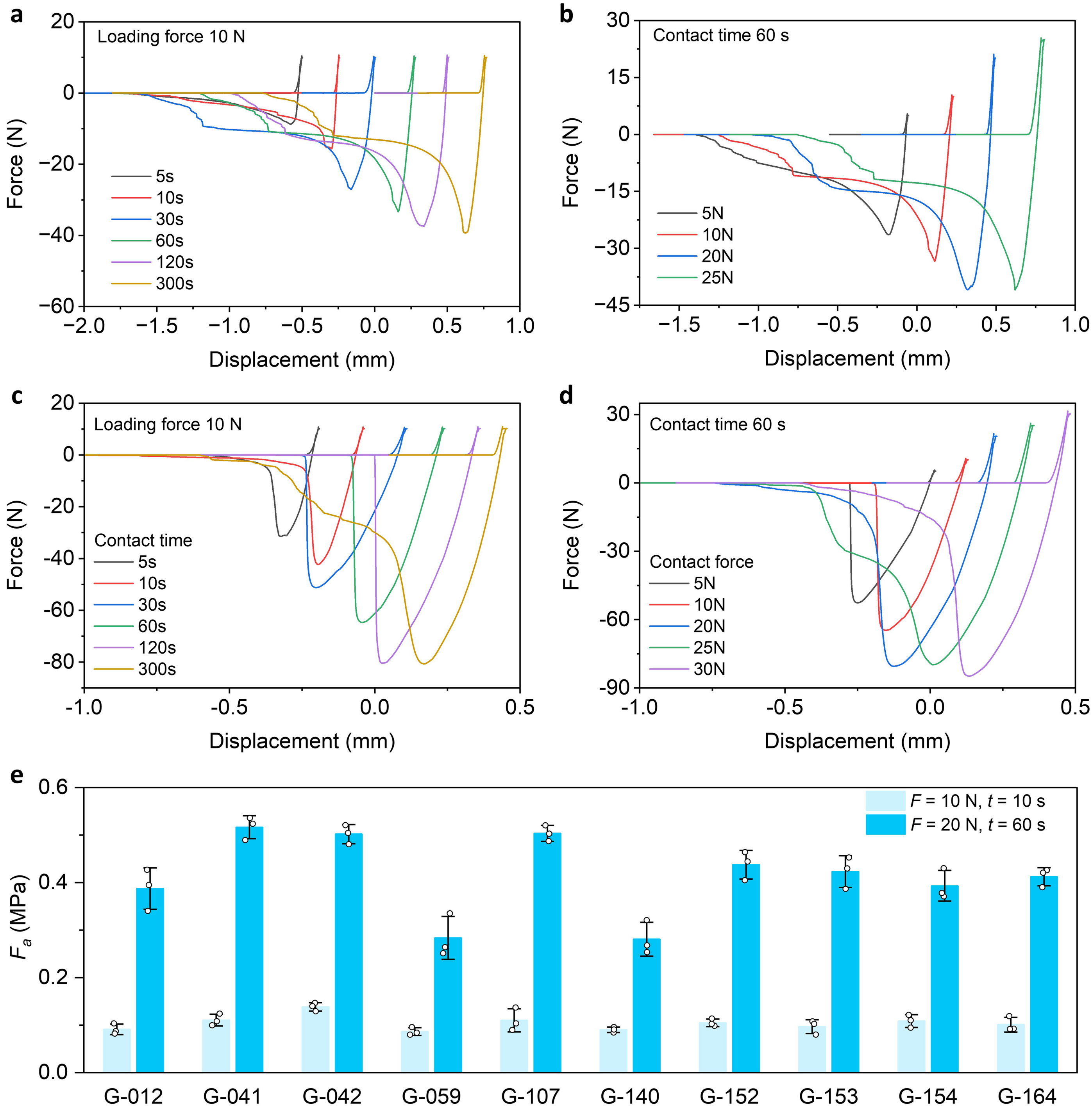
Extended Data Fig. 4 | Machine learning (ML) trained models. (a) Error plots for nine ML models using a 90%/10% training-test split. Training data points are represented by blue dots, while test data points are shown in red. The dashed line indicates where the predicted values match the experimental data (truth). (b) Root mean squared errors (RMSEs) depicting the prediction accuracy across the nine ML models trained on the dataset of 180 bioinspired hydrogels, assessed via 10-fold cross-validation. A lower test error, combined with minimized overfitting (i.e., a smaller gap between training and test errors), indicates a more effective regression model.



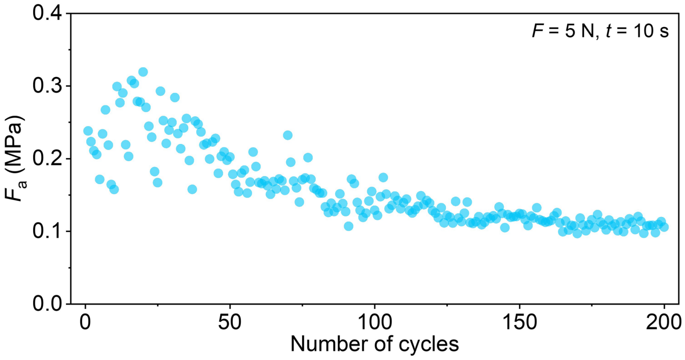
Extended Data Fig. 5 | Machine learning-driven optimization and experimental validation in three consecutive rounds. (a) Adhesive strength (F_a) of hydrogels fabricated in experiments according to the formulations proposed by GP_KB and RFR-GP models. (b) Variations in performance metrics, including: (i) successful rate (SR), defined as the fraction of the test set with higher true F_a than the training set; (ii) ratio of maximum true F_a between the test and training sets; and (iii) root mean squared errors (RMSEs) of the test sets. The success rate and F_a ratio decrease from the first round to the second round and level off in the third round, implying convergence toward the global optimum via SMBO. Meanwhile, the RMSE decreases continuously over the three rounds, indicating that expanding the training dataset improves the accuracy of regression models. (c) Parity plots comparing ML predicted F_a versus true F_a.



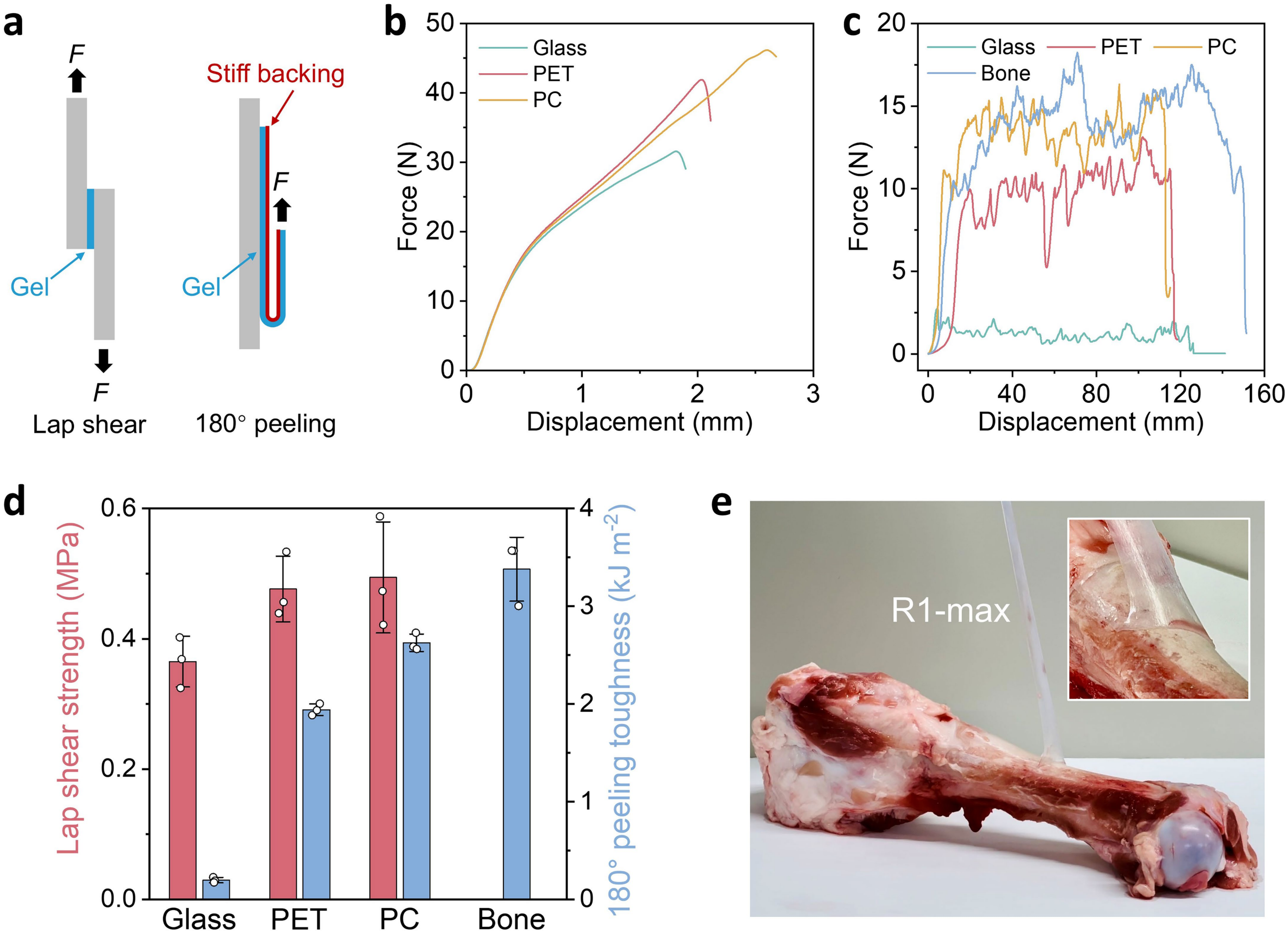
Extended Data Fig. 6 | Properties of top-performing hydrogels from machine learning and data mining approaches. (a) Angular frequency dependence of storage modulus (G') and loss modulus (G'') for the top-performing machine learning-driven gels (R1-max, R2-max, R3-max) and the top-performing data mining-driven gel (G-max) in DMSO. (b) Angular frequency dependence of G' and G'' for the four hydrogels equilibrated in normal saline (0.154 M NaCl). (c) Volume swelling ratio (Q) of the four hydrogels equilibrated in normal saline relative to their as-prepared state in DMSO. (d) Pure shear stress-stretch ratio curves for the R1-max hydrogel (equilibrated in normal saline) with and without a notch, measured at a stretch rate of 100 mm min^{-1}. The notched sample exhibited crack propagation at a critical stretch ratio (λ_c) of 3.4. The fracture energy (Γ) estimated from the pure-shear test is shown. Experimental details are provided in the Supplementary Materials. Error bars represent the standard deviation of $N = 3$ measurements.



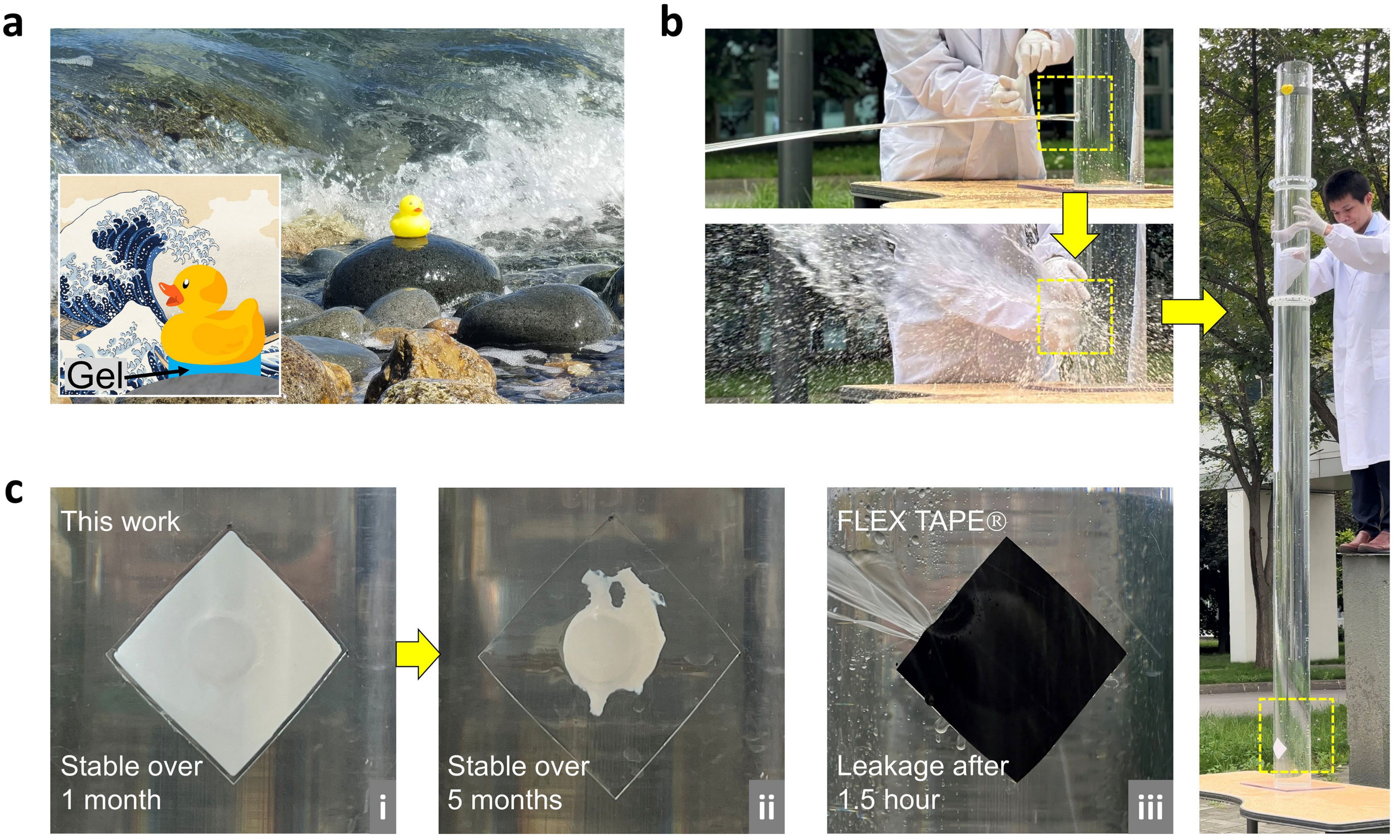
Extended Data Fig. 7 | Adhesive strength (F_a) of hydrogels under different tack test conditions. (a, b) Force-displacement curves of G-max hydrogel: (a) at a fixed loading force of 10 N with varying contact times, and (b) at a fixed contact time of 60 s with varying loading forces. (c, d) Force-displacement curves of R1-max hydrogel: (c) at a fixed loading force of 10 N with varying contact times, and (d) at a fixed contact time of 60 s with varying loading forces. (e) F_a of the 10 samples that exhibited high adhesions in Fig. 3e measured at different loading force and contact time. All adhesion tests were conducted in normal saline on glass substrates. Error bars represent the standard deviation of $N = 3$ measurements.



Extended Data Fig. 8 | Repeated adhesion of the R1-max hydrogel. Adhesion stability of the R1-max hydrogel (10 mm diameter, ~0.4 mm thickness) over 200 attachment-detachment cycles on a glass substrate in normal saline. Testing was conducted under a 5 N loading force and a 10-s contact time.



Extended Data Fig. 9 | Adhesion performance of the R1-max hydrogel on various substrates. (a) Schematic illustration of lap shear and 180° peeling tests for adhesion assessment. (b) Force-displacement curves from lap shear tests of R1-max (10 mm diameter, ~0.4 mm thickness) adhering to glass, PET, and PC substrates. (c) Force-displacement curves from 180° peeling tests of R1-max (10 mm × 150 mm strips, ~0.4 mm thickness) adhering to glass, PET, PC, and pork bone surfaces. (d) Lap shear adhesive strength and 180° peeling interfacial toughness of R1-max on various substrates. (e) Photographic images (from different perspective angles) showing R1-max (25 mm × 150 mm strips, ~0.4 mm thickness) being peeled away from a pork bone surface. All hydrogels were equilibrated in normal saline before testing. Error bars represent the standard deviation of $N = 3$ measurements. Experimental details are provided in the Methods section. These results highlight the exceptional adhesion performance of R1-max on various surfaces.



Extended Data Fig. 10 | Demonstration of data-driven hydrogels in practical applications. (a) Photographic images of R1-max adhering a rubber duck to a seaside rock, withstanding ocean tides. (b) Photographic images of R2-max (6 cm × 6 cm in size, ~0.37 mm thickness) sealing a 20-mm-diameter hole at the base of a 3-meter-tall PC pipe to halt high-pressure water leakage (burst flow rate at the outlet of the hole was ~5.4 m s^{-1}). (c) Photographic images show (i) R2-max successfully repairing a 20 mm hole at the base of a 3-meter-tall polycarbonate pipe filled with tap water; (ii) no water leakage was observed for over 5 months in air, with the gel becoming transparent upon drying, and the opaque region indicating water penetration only around the hole; (iii) in contrast, commercial FLEX TAPE® failed under the same conditions, with water leakage occurring within 1.5 h. These findings highlight the exceptional wet adhesion performance of the R2-max hydrogel.