AI-enabled Alkaline-resistant Evolution of Protein to Apply in Mass Production

Liqi Kang[†a], Banghao Wu[†b], Bingxin Zhou[c,d], Pan Tan[f], Yun (Kenneth) Kang[g], Yongzhen Yan[g], Yi Zong[g], Shuang Li[g], Zhuo Liu,[*c,d,e,f] and Liang Hong*[a,c,d,e,f]

a. School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, China.

b. School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China.

c. Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai 200240, China.

d. Shanghai National Centre for Applied Mathematics (SJTU Center), MOE-LSC, Shanghai Jiao Tong University, Shanghai 200240, China.

e. Zhangjiang Institute for Advanced Study, Shanghai Jiao Tong University, Shanghai 201203, China.

f. Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

g. Changchun GeneScience Pharmaceuticals Co., Ltd.

†These authors contributed equally to this work.

Corresponding author: hongl3liang@sjtu.edu.cn, judeliu@sjtu.edu.cn

Keywords: Deep learning, Large language model, Protein engineering, Antibody, Alkali resistance

Abstract: Artificial intelligence (AI) models have been used to study the compositional regularities of proteins in nature, enabling it to assist in protein design to improve the efficiency of protein engineering and reduce manufacturing cost. However, in industrial settings, proteins are often required to work in extreme environments where they are relatively scarce or even non-existent in nature. Since such proteins are almost absent in the training datasets, it is uncertain whether AI model possesses the capability of evolving the protein to adapt extreme conditions. Antibodies are crucial components of affinity chromatography, and they are hoped to remain active at the extreme environments where most proteins cannot tolerate. In this study, we applied an advanced large language model (LLM), the Pro-PRIME model, to improve the alkali resistance of a representative antibody, a VHH antibody capable of binding to growth hormone. Through two rounds of design, we ensured that the selected mutant has enhanced functionality, including higher thermal stability, extreme pH resistance and stronger affinity, thereby validating the generalized capability of the LLM in meeting specific demands. To the best of our knowledge, this is the first LLM-designed protein product, which is successfully applied in mass production.

1. Introduction

Protein engineering, situated at the nexus of molecular biology, bioinformatics, and biotechnology, focuses on the design of proteins to introduce novel functionalities or enhance existing attributes[1-3]. With the exponential growth of biological data and computational power, protein engineering has experienced a significant shift towards advanced computational methodologies, particularly deep learning, to expedite the design process and unravel complex protein-function relationships[4-9]. However, a significant challenge in industrial protein engineering is designing proteins with inherent resistance to extreme conditions, such as high temperature and extreme pH environments (acidic or alkaline)[10, 11]. Unlike proteins in natural ecosystems, those used in industrial processes often encounter harsh physical and chemical conditions, necessitating exceptional resilience to maintain functionality[12, 13]. Previous efforts to enhance protein resistance have often relied on rational design and mutant library screening. These methods are typically labor-intensive, inefficient, and yield limited improvements[14-17]. Consequently, the industrial demand for proteins resilient to harsh environments poses a notable absence within the training datasets of Artificial Intelligence (AI) models. Exploring whether AI can achieve the evolution of protein resistance to extreme environments is crucial for broadening protein applications and improving modification efficiency.

Recent advances in large-scale protein language models (LLMs) have enabled zero-shot predictions of protein mutants based on self-supervised learning from natural protein sequences[18-21]. Although AI-guided protein design has been applied to predict the mutants with greater thermostability and higher activity[22-24], it is unexplored whether these models based on the natural protein information can find the mutants that adapt the unnatural extreme environments, such as the alkaline solution with the pH value higher than 13.

Here, we employed a LLM (large language model) developed by our group, the Pro-PRIME model[25], to predict dozens of mutants of a nano-antibody against growth hormone (a VHH antibody), and examined their fitness,

including alkali resistance and thermostability, to evaluate their performance under extreme environments. We utilized the Pro-PRIME model to score saturated single-point mutations of the VHH in a zero-shot setting, and selected the top 45 mutants for experimental testing. Some mutants exhibited improved alkali resistance, while others demonstrated higher thermal stability or affinity. Subsequently, we fine-tuned the Pro-PRIME model to predict dozens of multi-point mutations. As a result, we obtained three multi-point mutants with enhanced alkali resistance, higher thermostability, as well as strong affinity to the targeted protein. Also, the dynamic binding capacity of the selected mutant did not show significant decline after more than 100 cycles, making it suitable for practical application in industrial production. The selected mutant has been used in practical production and lower the cost for over one million dollars in a year. To the best of our knowledge, this is the first protein product developed by a LLM that has been successfully applied in mass production. Due to the Pro-PRIME model's ability to achieve precise predictions of multi-point mutations with reliance on a small amount of experimental data, our two-round design process involved experimental validation of only 65 mutants in two months, demonstrating remarkable high efficiency. Furthermore, we performed a systematic analysis of these findings and determined that the model can yield more valuable predictive outcomes while remaining consistent with rational design principles. Specifically, within the framework of multi-point combinations, the model's incorporation of negative single-point mutations into the combinatorial space led to exceptional results, showcasing its capacity to capture epistatic interactions. Notably, in striving for global optimum, deep learning methods offer distinct advantages over traditional rational design approaches.

2. Results and Discussion

2.1. Wet experimental testing of single-point mutants

Currently, the prevailing approach in protein engineering involves rational design or high-throughput screening to identify positive single-point mutations, followed by their combination into more effective multi-point mutations through greedy search methods[26]. However, these approaches are inclined to trap in a local optimum and unable to avoid the negative epistasis[27, 28]. To overcome these limitations, we employed the Pro-PRIME model to screen single-point mutations in the first round, enabling rapid inference of scores for saturated mutations. An efficient strategy involves validating only a select number of top-ranked mutants, thereby saving considerable time and reducing economic costs compared to rational design and high-throughput experiment. Moreover, we posit that deep learning models can capture nuanced features during training, and recommend single-point mutations that expert experience may not cover. These mutants expand possibilities for subsequent combination of multi-point mutations and mitigate the risk of converging into local optima.

Specifically, we made use of the Pro-PRIME model to score the single-site saturated mutants, and selected the top 45 single-point mutants of the VHH antibody for testing. Pro-PRIME is a deep learning-based methodology developed to guide protein engineering, utilizing masked language modeling (MLM) and multi-task learning to study and comprehend the semantic and grammatical features inherent in protein sequences, and further capture the temperature traits associated with these sequences. A higher score corresponds to a greater appearance probability of the residue at this site in the Pro-PRIME model[25]. The alkali resistance of a mutant was determined by its half maximal effective concentration (EC50) after treatment with alkali. A lower EC50 represents that antibodies have stronger affinity. Firstly, the 45 single-point mutants were treated with 0.3 M NaOH for 24 hours, and evaluated by their thermal stability and affinity. The results revealed that 15 mutants exhibited higher alkali resistance, and 35 mutants displayed higher melting temperatures (T_m) (Figure 1A). We also noted that 8 out of 45 single-site mutants

90 designed by Pro-PRIME showed improved affinity as compared to the WT before alkali treatment (Figure S1). It is
91 worth noting that six mutants (A57D, A15P, V113D, P117Q, R20T, and L12K) exhibited enhancements in all three
92 properties, i.e., higher alkali resistance, greater thermal stability, and stronger affinity before alkali treatment.
93 Additionally, eight mutants (P29T, N85Q, G134R, N78S, R110E, L49V, T58K, and W112F) exhibited improved alkali
94 resistance and T_m, albeit at the moderate cost of affinity. This trade-off is acceptable, as excessively high affinity could
95 complicate the separation of growth hormone from the VHH. Furthermore, we subjected single-point mutants with
96 remarkable alkali resistance to treatment with 0.5 M NaOH, and measured their affinity after 24 hours. Our findings
97 demonstrated that four mutants (L49V, A15P, W112F, and R20T) exhibited higher affinity than the wild type following
98 the same treatment protocol (Figure 1B). Interestingly, these four single-point mutations all showed improvements in
99 T_m, and their affinities before alkali treatment were either similar to or lower than those of the wild type. However,
100 this does not imply that enhancing thermal stability will concurrently improve alkali resistance. The spearman
101 correlation between the EC50 values after 0.3 M NaOH treatment and T_m values of single-point mutants is -0.29,
102 indicating that these two properties of the antibody are only weakly correlated (Figure S3). Therefore, achieving
103 multi-point mutants that enhance both alkali resistance and thermal stability remains a challenge.

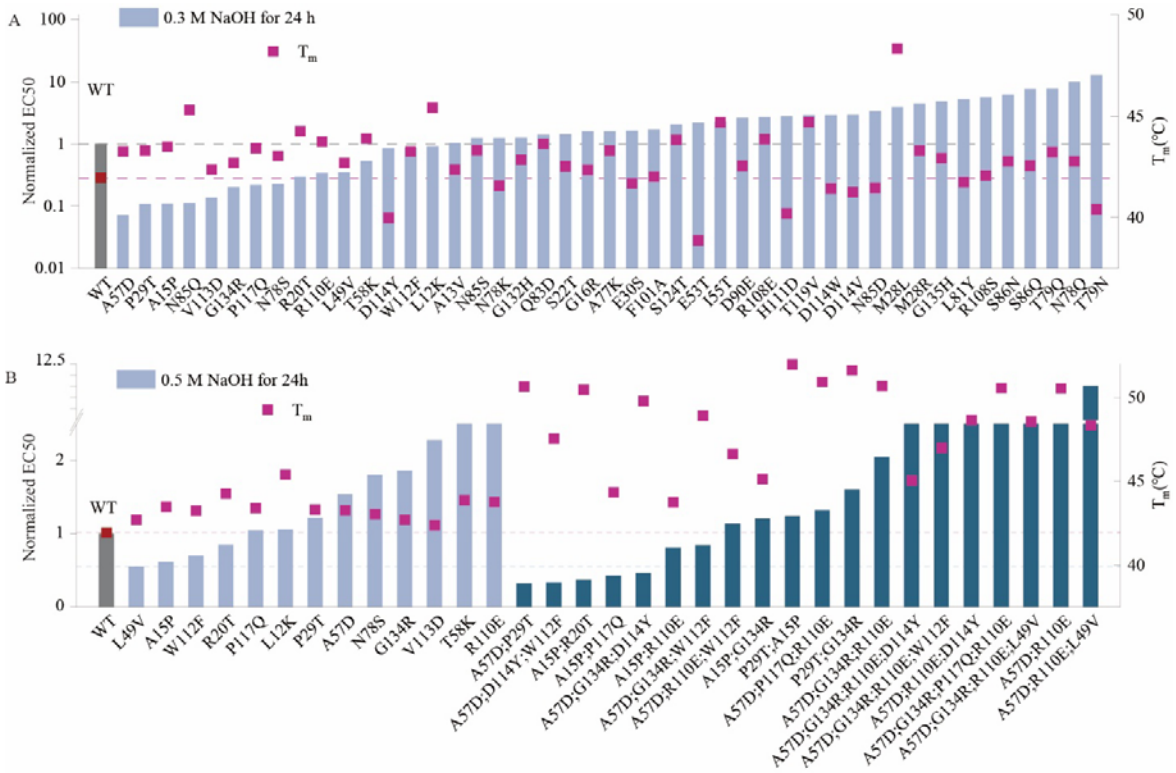

Figure 1. The experimental results from two rounds of design. (A) Experimental results of single-point mutants. The melting temperatures of mutants are shown as red squares. The affinity of mutants after treatment with 0.3 M NaOH for 24 hours are shown as blue bars. (B) Experimental results of multi-point mutants and single-point mutants. The melting temperatures of mutants are shown as red squares. The affinity of single-point mutants after treatment with 0.5 M NaOH for 24 hours are shown as blue bars, while those of multi-point mutants are shown as green bars. The affinity values were normalized in the graph, with the wild-type EC50 set as 1.

2.2. Multi-point mutation design driven by experimental data

In the process of combining multiple single-site mutations, common greedy algorithms typically proceed by sequentially adding the most effective single-point mutations and gradually increasing the number of stacked mutations. However, this approach often demonstrates low efficiency and is susceptible to becoming trapped in local optima. Importantly, the most effective multi-point mutations in practical scenarios may not necessarily include the most effective single-point mutations[28]. Moreover, in the traditional approach of incrementally stacking mutations from single points to multiple points, only paths that yield better results at each step are retained[27]. Nevertheless, combinations that are discarded due to decreased effectiveness may significantly enhance performance in subsequent stacking step, a phenomenon unforeseeable in traditional methods.

Although the model inference is fast, it is not feasible to explore all possible mutations when designing multi-point mutants due to the exponential increase in the number of potential combinations. To manage this challenge, we constructed a mutant library based on a two-stage design process. In the first stage, we scored all single-point mutations using the model, and in the second stage, we combined experimentally validated single-point mutations to create the multi-point mutant library. This approach ensures that even when designing multi-point mutants (e.g., five-point mutants), the number of mutants to score remains in the millions, which is computationally efficient and practical. The number of single-point mutations selected for the multi-point mutant library is a key factor influencing both the computational load and the scope of the design space. To maintain a balance between efficiency and accuracy, we limited the number of single-point mutations to between 30 and 50. This strategic approach allows us to achieve both scalability and precision in our protein engineering tasks.

To better search the global optimium, we chose the Pro-PRIME model to predict the results of multi-point mutants due to its remarkable performance on the design of diverse proteins[25]. We trained two Pro-PRIME models using the alkali resistance and T_m data of single-point mutants, respectively, to score these two properties of mutants. During the prediction of multi-point mutations, we prioritized mutants with higher alkali resistance scores while ensuring T_m scores were not lower than those of the wild type. Ultimately, we selected 20 multi-point mutants for experimentation (see details in Supporting Information, SI). Experimental results indicated that the alkali resistance of five multi-point mutants (A57D;P29T, A57D;D114Y;W112F, A15P;R20T, A15P;P117Q, and A57D;G134R;D114Y) surpassed that of the best single-point mutant. The T_m of multi-point mutants was generally higher, with the highest (P29T;A15P) exceeding that of the wild type by approximately 10 °C (Figure 1B). Although we did not optimize for the affinity of the VHH antibody, certain multi-point mutants (A15P;R20T, A15P;G134R, and P29T;G134R) exhibited affinity levels close to or even exceeding those of the wild type (Figure S2). Although experimental results from single-point mutations suggested that simultaneously improving the alkali resistance and thermal stability of VHH

might be challenging, we successfully designed multi-point mutants that balance multiple properties. This demonstrates the excellent multi-objective optimization capability of the Pro-PRIME model.

Note that we employed different strategies for designing single-point and multi-point mutants, specifically using a zero-shot approach for single-point mutations and fine-tuning the model for multi-point mutations. These choices were made based on the distinct characteristics of the two tasks and the availability of experimental data. For single-point mutations, the number of possible mutations is relatively limited, and at the outset, there were no experimental data available. In such cases, the zero-shot setting was chosen because it allows the model to predict the fitness of mutants based solely on the information learned during pre-training on a large protein sequence dataset. Since single-point mutations are computationally manageable, this approach was deemed appropriate to generate initial predictions for protein engineering. However, when designing multi-point mutants, the situation changes significantly. The potential combinations of mutations increase exponentially, and without prior data, it becomes computationally infeasible to evaluate every possible combination within a reasonable timeframe. Moreover, by the time we reached the multi-point mutation design stage, experimental data for several single-point mutations had already been obtained. This data enabled us to fine-tune the model to better capture the specific structural and functional features that contribute to protein stability and resistance, especially in the context of multiple interacting mutations. Fine-tuning improves the model's accuracy by adjusting its parameters to align more closely with the experimental data, ensuring that the predicted multi-point mutants are more likely to meet the desired engineering goals. After the second round of design, the fitness of the mutants was further improved. In improving alkali resistance, experimental results showed that 15 of the 45 designed mutants exhibited positive responses, yielding a success rate of 30%, close to the 35% success rate achieved in the second round. Compared to the wild type, the best single-point mutant improved alkali resistance by approximately 44.7%, while the best multi-point mutant achieved a 67.7% increase. For thermal stability enhancement, the success rate in the first round

was 77.8%, rising to 100% in the second round. The top single-point mutant exhibited a Tm increase of 6.37°C over the wild type, while the best multi-point mutant had a Tm increase of 10.02°C. We also tested the performance of the zero-shot approach for multi-point mutants, and the results showed that this method did not yield satisfactory predictions. The Spearman correlation coefficient between the zero-shot predictions and experimental results for multi-point mutants was -0.71, indicating a significant discrepancy. This further highlights the importance of fine-tuning the model for multi-point mutations, as the fine-tuned model provided more accurate and reliable results. In summary, the choice of zero-shot for single-point mutations and fine-tuning for multi-point mutations was driven by practical considerations regarding computational feasibility and the availability of experimental data. Fine-tuning the model significantly enhances its predictive performance, particularly for complex multi-point mutations where multiple residues interact. We believe this strategy strikes an optimal balance between computational efficiency and predictive accuracy, making it well-suited for practical protein engineering applications.

2.3. Complex epistatic effects generated by the combination of single-point mutations

To understand the principle of the evolution, we analyzed the improvement in multi-point mutants across three different indicators. As illustrated in Figure 2B, the evolution of thermal stability in the second round was successful, with nearly all multi-point mutants exhibiting an increase in T_m compared to the best single-point mutant they contain. Three multi-point mutants demonstrated higher affinity than the wild type, but only P29T and G134R exhibited positive epistasis, resulting in the creation of a double-point mutant with superior affinity. Although A15P;R20T and A15P;G134R showed higher affinity than the wild type, these combinations did not contribute to affinity improvement because the A15P single-point mutation they contained exhibited stronger affinity compared to these double-point mutants.

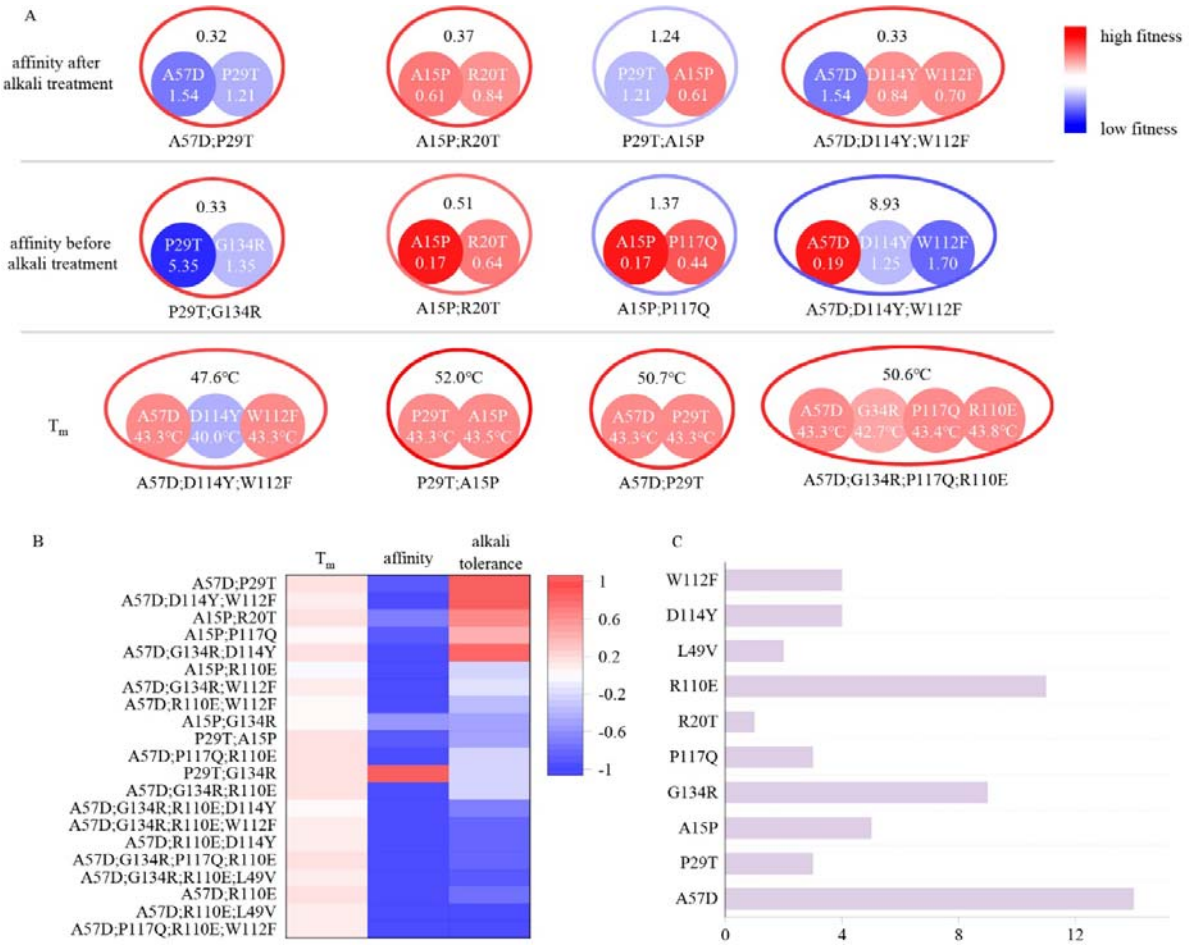

Figure 2. Schematic diagram illustrating the combined effects of single-point mutations. The colors (outer contours representing multi-point mutations and inner solid circles representing the included single-point mutations) and numerical values represent the EC50 values after alkali treatment with 0.5 M NaOH for 24 hours, the EC50 values before alkali treatment, and the Tm values. Blue indicates mutations inferior to the wild type, while red indicates mutations superior to the wild type. The EC50 values before and after alkali treatment are normalized, with the wild type set to 1. (B) Experimental elvalutation of multi-point mutations. Different colors represent the proportions of improvement or decline in T_m, affinity, and alkali resistance of multi-point mutations compared to the corresponding best single-point mutations they include. (C) Distribution of the occurrences of single-point mutations included in the 20 multi-point mutation variants. The length of the bar represents the frequency of occurrence of each mutation.

As shown in Figure 2A, single-point mutations produced complex results during the combination process due to epistatic effects, which were unpredictable using traditional methods. A57D;P29T exhibited remarkably high alkali resistance, despite A57D and P29T being negative mutants individually. This "double negative yields positive" phenomenon also occurred in terms of affinity, where the EC50 value of P29T;G134R was approximately one-third of that of the wild type, despite both P29T and G134R caused a decrease in affinity individually, with the EC50 value of

P29T being larger than that of the wild type by a factor of five. These mutation pairs are distantly located on the VHH antibody structure, making it challenging to infer the reasons for the enhanced properties after combination (Figure 3B). Not all combinations resulted in improved mutant effects. For instance, while A15P showed high affinity and alkali resistance, adding R20T or P29T led to decrease in affinity or alkali resistance, respectively. Additionally, combining A15P with P117Q, two single mutations capable of individually enhancing affinity, resulted in a decrease in affinity.

In traditional methods, only positive single-point mutations are selected to compose the multi-point mutants and negative single-point mutations are typically avoided in multi-point combinations. However, the number of negative mutations often far exceeds that of positive mutations, thus limiting the sequence space that can be explored by traditional methods. Large language models can leverage negative mutations to generate positive multi-point mutants, surpassing the capabilities of rational design and significantly expanding the design space in protein engineering. Therefore, the Pro-PRIME method manifests significant advantages in exploring sequence space, being less susceptible to local optima, and having greater potential to find the global optimum. As depicted in Figure 2C, the 20 multi-point mutations we identified consisted of combinations of nine different single-point mutations. Although the distribution was biased, we incorporated some negative mutations that experts might consider disadvantageous, implying that the Pro-PRIME method balances the exploitation of local optima and the exploration of global optimum.

2.4. Reevaluating the AI predictions from a rational design perspective

In addition to interpret the multi-point mutations based on the single-site mutations, we investigated the design of VHH antibodies from a rational design perspective to understand the AI predictions. The experimental results from SDS-PAGE revealed a significant degradation of the VHH antibody following alkali treatment, with small bands accounting for up to 61.1%. This phenomenon could be attributed to the poor alkali resistance of the VHH antibody.

Mass spectrometry analysis identified specific breakage sites at Q2, Q4, G9, G10, G16, S22, S23, S26, A57, N78, N85, F101, and G134, highlighting these as ideal targets for rational design. Introducing mutations at these sites to prevent the breakdown of VHH antibodies could effectively enhance their alkali resistance. The predictions of the Pro-PRIME model align with this rationale, with 4 out of the top 10 performing single-point mutations occurring at these breakpoint sites. Experimental evidence showed that our single-point mutations could effectively reduce the proportion of small bands, and not all positive mutations locate at the breakage sites, such as P29T and A15P (Figure 3A). Mutations identified by the model at other locations may have non-local effects, contributing to overall protein stability. In the second round of experiments, multi-point mutants exhibited an overall lower proportion of small bands, reaching as low as 13.6% (A57D;G134R;D114Y), though we did not use the experimental results of SDS-PAGE to train the model and make predictions. Thus, the improvement in the second round suggests a positive correlation between the alkali resistance of the VHH antibody and the degree of breakage after alkali treatment. Employing a rational design approach, it is plausible to identify single-point mutants with improved alkali resistance at these breakpoints. However, determining the optimal mutated residues requires conducting multiple single-point saturation mutation experiments, which is time consuming and costly. Deep learning methods can directly predict mutation types at each site, enhancing design efficiency while remaining aligned with rational design principles.

Furthermore, we analyzed the VHH antibody and homologous sequences from the perspective of evolutionary sequence alignment (Figure 3C). The results revealed that some critical single-point mutations (P29T, L49V) changed original residues to more conservative ones, aligning with the "consensus sequence" method commonly used in rational design[29]. The Pro-PRIME model's ability to predict such mutations is expected, given its training set contains information from homologous proteins. However, the advantage of deep learning methods extends beyond providing suggestions aligned with rational design principles; they can predict mutations contrary to expert experience. For instance, both A15 and R20 are conservative residues according to homologous sequence alignment (see Figure 3C). Moreover, mutating A15 to proline reduces hydrogen bonds, and changing R20 to threonine

contradicts empirical physicochemical properties, since arginine is more likely to form salt bridges and hydrogen bonds, benefiting stability. However, both A15P and R20T contribute to increased alkali resistance and play positive roles in subsequent multi-point combinations. This exceptional capability beyond the traditional rational-design principles expands the exploration space of deep learning models, aiding to approach the global optimum.

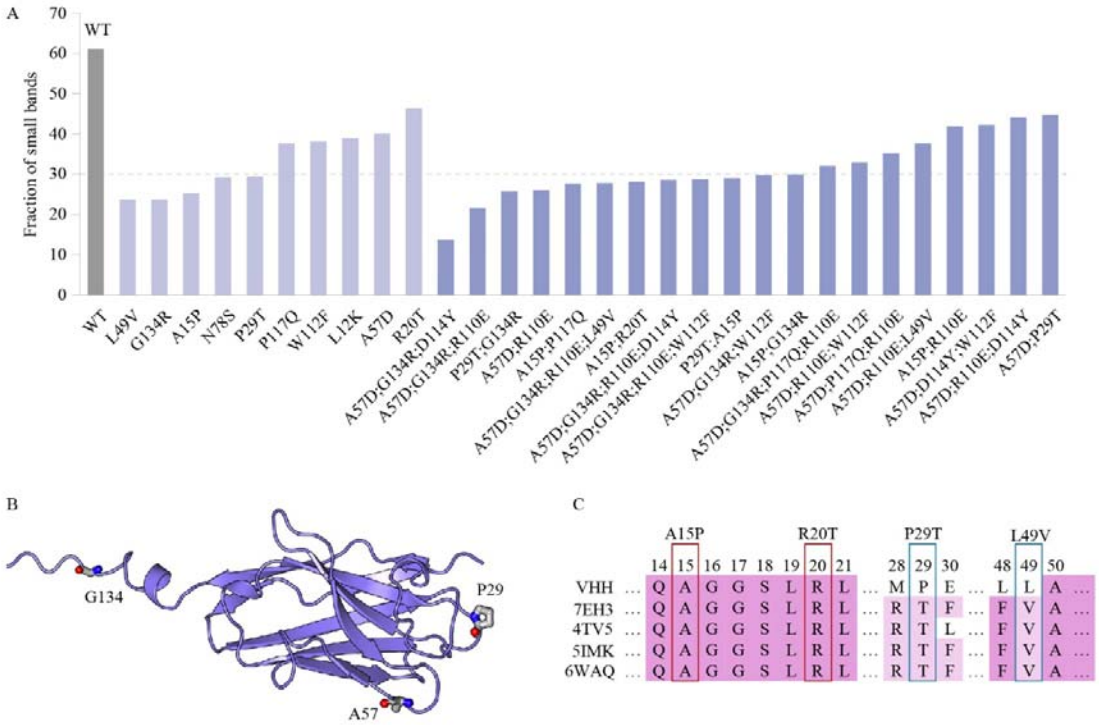

Figure 3. (A) The SDS-PAGE experimental results depict the proportion of small bands observed after alkali treatment for multi-point mutations and certain single-point mutations exhibiting relatively higher alkali resistance. (B) Structure of the VHH antibody. (C) The multiple sequence alignment of the VHH and several homologous sequences.

In order to gain deeper insights into the mechanisms by which the identified mutations enhance protein properties, we performed molecular dynamics (MD) simulations on the best alkali-resistant mutant. The simulation results revealed several key observations that help explain the observed improvements in protein stability and alkali resistance. As shown in Figure S4A, the two-point mutant of A57D;P29T has a Tm increase of around 8 ℃ and a much stronger binding affinity than the WT. Our analysis of the MD trajectories indicates that the A57D;P29T mutant

has a more rigid structure than that of WT due to its lower root mean squared deviation (RMSD) of protein (Figure S4B). Furthermore, we calculated the root mean squared fluctuation (RMSF) for each residue, and realized that the mutant displayed less fluctuation at residue 29 but similar flexibility at residue 57. Interestingly, residues at positions 10, 108 and 118 which spatially distant from residues 29 and 57 in the mutant exhibited remarkable weakened fluctuations than those in the WT (Figure S1C), implying a more rigid structure of the mutant contributing to its improved resistance on high temperature and strong alkalinity. However, Figure S4D shows the AlphaFold3 predicted structures of the WT and the mutant are quite similar. To unveil the origin of change on structural flexibility, we computed the intramolecular interactions, such as salt bridges and hydrogen bonds for both WT and the mutant. We observed that the mutations increased the number of hydrogen bonds between the mutation sites and the rest of the protein (Figure S4E). However, the overall structure of the mutant did not show significant changes, which is also evident from the solvent-accessible surface area (SASA) analysis (Figure S4F). We also analyzed changes in salt bridges and found that although residue 57 mutated to Histidine, no new salt bridges were formed. Additionally, RMSF results showed that residues 10, 108, and 118 became more rigid, but further analysis revealed that there were no significant changes in hydrogen bonds or other interactions in these regions. Taken together, these findings suggest that the enhanced alkali resistance of the mutant is likely due to an overall increase in protein stability, rather than a dramatic change in its structural conformation. The MD simulation results, which are detailed in Figure S4, provide a deeper understanding of how specific mutations can improve protein properties and offer valuable insights for future protein engineering applications.

2.5. Improved acid and salt resistance simultaneously

According to the rational design principles, a protein with enhanced intramolecular interactions can resist various stress, including high temperature, strong alkali, strong acid, concentrated salt, etc.[30, 31] Hence, we expected that the mutants with higher thermal stability and alkali resistance can tolerate acidic and saline environments as well.

We evaluated the binding affinity of the mutants to the target protein under elution with saline (1 M NaCl) or acidic (20 mM citric acid) solutions to characterize their salt-induced and acid-induced dissociation abilities. Additionally, we assessed the acid resistance of the mutants by measuring their affinity after a 48-hour treatment with 1 M ethanoic acid (see details in SI). Figure 4 shows the acid and salt resistance of the selected multi-point mutants with higher thermal stability and stronger alkali resistance. It is noteworthy that the Pro-PRIME model can provide us the mutants, such as A57D;P29T and A15P;R20T, with strong resistance on acid and salt (Figure 4A) as well as high temperature and alkali resistance (Figure 2B). Our design enhances protein stability while retaining other properties such as affinity, thereby endowing the mutant with the potential for application in industrial production.

2.6. Application in industrial production

The Pro-PRIME method can predict multi-site mutants using a small amount of experimental data, and the entire design process can be completed within two months. This method has the potential to empower industrial production due to its relatively low economic and time costs. Our mutants have already been widely applied in the purification process of growth hormones. Figure 4B shows the ratio of the residual dynamic binding capacity (DBC) after 0.5 M NaOH treatment for 6 hours and 24 hours to the residual DBC before treatment for various mutants and the wild type. The wild type experienced a 74.1% loss in the residual DBC after 6 hours of alkali treatment, and only retained 15.2% of its residual DBC after 24 hours of alkali treatment. In contrast, the designed multi-site mutants retained 60-90% of their residual DBC after 6 hours of alkali treatment, and even after 24 hours of treatment, some mutants maintained over 50% of their residual DBC, which is remakably more stable than the wild type in the alkaline environment of industrial production. Figure 4C shows the variation in yield during affinity chromatography across multiple cycles after employing our mutant. After more than 140 cycles, the yield of our mutant did not exhibit a significant downward trend. In contrary, the binding affinity of wild type is unable to sustain after 60 cycles. Therefore, our designed mutants maintain activity after more cycles of reuse compared to the wild type,

substaintially reducing the production cost of growth hormones. Consequently, the selected mutant designed by the LLM has been applied in mass production scale up to 5,000 liters.

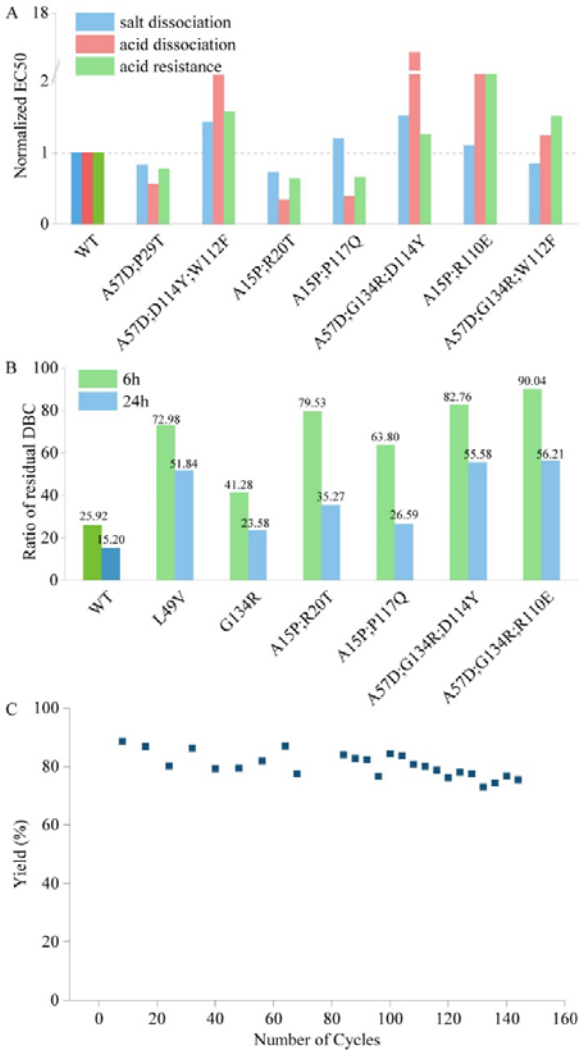

Figure 4. (A) The tolerance of acid dissociation and salt dissociation, and the acid resistance of multi-point mutations with strong alkali resistance. All EC50 values are normalized, with the wild type set to 1. (B) Experiment results of residual dynamic binding capacity (DBC) at 10% breakthrough. The bars represent the ratio of the residual DBC after 0.5 M NaOH treatment for 6 hours and 24 hours to the residual DBC before treatment for multi-site mutants and the wild type. (C) Yield in affinity chromatography and the corresponding number of cycles. This figure illustrates the variation in yield during affinity chromatography across multiple cycles.

3. Conclusions

Through two rounds of evolution, we successfully designed a VHH antibody with strong resistance to extreme environments and enhanced affinity using the Pro-PRIME model. Although rare case can tolerate the extreme pH and

saline conditions in our pre-training dataset, the Pro-PRIME model showed impressive performance after supervised learning with limited data, especially on capturing the epistatic effects. The analysis of these 65 mutants revealed that the Pro-PRIME model is adept at exploring the large space of protein fitness, being less susceptible to local optima, and having greater potential to find the global optimum. Our efficient method of designing mutants that consider multiple properties improvement holds promise for industrial application of proteins. Specifically, the VHH antibody has been deployed in practical production and significantly enhancing the efficiency of the entire production line after our design. While the Pro-PRIME model itself has been reported[25], this work demonstrates its first-time application to the challenge of designing proteins with alkali resistance and other extreme properties that are not found in natural proteins, nor have previous studies addressed or provided data for such applications. This shift from optimizing existing protein properties to engineering entirely new, unnatural traits is a significant advance in the field. This study shows that the AI models, such as Pro-PRIME, can not only guide the evolution of protein thermal stability, enzymatic activity, ligand affinity but also enable to develop the mutants adapting the harsh unnatural environments, such as extreme pH and concentrated salt, largely expanding its application. The novelty of this work lies in the ability to design and engineer proteins with novel properties, specifically alkali resistance, which is an unprecedented achievement in AI-assisted protein engineering. The great potential of AI model is expected to significantly accelerate the development of proteins for diverse applications in medicine, agriculture, bioengineering, etc.

CRediT authorship contribution statement

Liqi Kang: Writing – original draft, Data curation, Conceptualization, Methodology. **Banghao Wu:** Data curation, Methodology. **Bingxin Zhou:** Writing – review & editing. **Pan Tan:** Writing – review & editing. **Yun (Kenneth) Kang:** Data curation. **Yongzhen Yan:** Data curation. **Yi Zong:** Data curation. **Shuang Li:** Data curation. **Zhuo Liu:** Writing – review & editing, Supervision. **Liang Hong:** Writing – review & editing, Supervision.

Conflicts of interest

Patents have been filed in 2023 for all these mutations in the present work. GeneScience Pharmaceuticals, a company that discovered the native sequence of VHH, own the patent of VHH.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (12204302), the Computational Biology Program of Shanghai Science and Technology Commission (23JS1400600), the Shanghai Pujiang Program (Grant No. 22PJ1406900), the Startup Fund for Young Faculty at SJTU (SFYF at SJTU), the Oceanic Interdisciplinary Program of Shanghai Jiao Tong University (Project No. SL2022MS018), the Natural Science Foundation of Shanghai (Grant No. 23ZR1431700), Shanghai Jiao Tong University Scientific and Technological Innovation Funds (21X010200843), Science and Technology Innovation Key R&D Program of Chongqing (CSTB2022TIAD-STX0017), the Student Innovation Center at Shanghai Jiao Tong University, and Shanghai Artificial Intelligence Laboratory.

References

[1] S.L. Lovelock, R. Crawshaw, S. Basler, C. Levy, D. Baker, D. Hilvert, A.P. Green, The road to fully programmable protein catalysis, Nature 606(7912) (2022) 49-58.

[2] N. Tokuriki, D.S. Tawfik, Protein dynamism and evolvability, Science 324(5924) (2009) 203-207.

[3] S. Lutz, S.M. Iamurri, Protein engineering: past, present, and future, Protein Engineering: Methods and Protocols (2018) 1-12.

[4] H. Narayanan, F. Dingfelder, A. Butté, N. Lorenzen, M. Sokolov, P. Arosio, Machine learning for biologics: opportunities for protein engineering, developability, and formulation, Trends in pharmacological sciences 42(3) (2021) 151-165.

[5] W.D. Jang, G.B. Kim, Y. Kim, S.Y. Lee, Applications of artificial intelligence to enzyme and pathway design for metabolic engineering, Current Opinion in Biotechnology 73 (2022) 101-107.

[6] Y. Qiu, G.-W. Wei, Artificial intelligence-aided protein engineering: from topological data analysis to deep protein language models, Briefings in Bioinformatics 24(5) (2023) bbad289.

[7] Z. Zhou, L. Zhang, Y. Yu, B. Wu, M. Li, L. Hong, P. Tan, Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning, Nature Communications 15(1) (2024) 5566.

[8] Y. Tan, M. Li, P. Tan, Z. Zhou, H. Yu, G. Fan, L. Hong, PETA: Evaluating the impact of protein transfer learning with sub-word tokenization on downstream applications, arXiv preprint arXiv:2310.17415 (2023).
[9] M. Li, L. Kang, Y. Xiong, Y.G. Wang, G. Fan, P. Tan, L. Hong, SESNet: sequence-structure feature-integrated deep learning method for data-efficient protein engineering, Journal of Cheminformatics 15(1) (2023) 1-13.
[10] M.M. Pinney, D.A. Mokhtari, E. Akiva, F. Yabukarski, D.M. Sanchez, R. Liang, T. Doukov, T.J. Martinez, P.C. Babbitt, D. Herschlag, Parallel molecular mechanisms for enzyme temperature adaptation, Science 371(6533) (2021) eaay2784.
[11] R. Jaenicke, Protein stability and molecular adaptation to extreme conditons, European Journal of Biochemistry 202(3) (1991) 715-728.
[12] M.T. Reetz, Z. Sun, G. Qu, Enzyme engineering: selective catalysts for applications in biotechnology, organic chemistry, and life science, John Wiley & Sons2023.
[13] Y. Xia, X. Li, L. Yang, X. Luo, W. Shen, Y. Cao, L. Peplowski, X. Chen, Development of thermostable sucrose phosphorylase by semi-rational design for efficient biosynthesis of alpha-D-glucosylglycerol, Applied Microbiology and Biotechnology 105 (2021) 7309-7319.
[14] K. Minakuchi, D. Murata, Y. Okubo, Y. Nakano, S. Yoshida, Remarkable alkaline stability of an engineered protein A as immunoglobulin affinity ligand: C domain having only one amino acid substitution, Protein Science 22(9) (2013) 1230-1238.
[15] M. Linhult, S. Gülich, T. Gräslund, A. Simon, M. Karlsson, A. Sjöberg, K. Nord, S. Hober, Improving the tolerance of a protein a analogue to repeated alkaline exposures using a bypass mutagenesis approach, Proteins: structure, function, and bioinformatics 55(2) (2004) 407-416.
[16] S. Gülich, M. Linhult, S. Ståhl, S. Hober, Engineering streptococcal protein G for increased alkaline stability, Protein engineering 15(10) (2002) 835-842.
[17] B. Palmer, K. Angus, L. Taylor, J. Warwicker, J.P. Derrick, Design of stability at extreme alkaline pH in streptococcal protein G, Journal of biotechnology 134(3-4) (2008) 222-230.
[18] A. Madani, B. Krause, E.R. Greene, S. Subramanian, B.P. Mohr, J.M. Holton, J.L. Olmos Jr, C. Xiong, Z.Z. Sun, R. Socher, Large language models generate functional protein sequences across diverse families, Nature Biotechnology (2023) 1-8.
[19] B. Hie, D. Zhong Ellen, B. Berger, B. Bryson, Learning the language of viral evolution and escape, Science 371(6526) (2021) 284-288. https://doi.org/10.1126/science.abd7331.
[20] R.M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, A. Rives, MSA transformer, International Conference on Machine Learning, PMLR, 2021, pp. 8844-8856.
[21] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C.L. Zitnick, J. Ma, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proceedings of the National Academy of Sciences 118(15) (2021) e2016239118.
[22] L.M. Blaabjerg, M.M. Kassem, L.L. Good, N. Jonsson, M. Cagiada, K.E. Johansson, W. Boomsma, A. Stein, K. Lindorff-Larsen, Rapid protein stability prediction using deep learning representations, Elife 12 (2023) e82593.
[23] S. Wang, H. Tang, Y. Zhao, L. Zuo, BayeStab: Predicting effects of mutations on protein stability with uncertainty quantification, Protein Science 31(11) (2022) e4467.
[24] F. Li, L. Yuan, H. Lu, G. Li, Y. Chen, M.K. Engqvist, E.J. Kerkhoven, J. Nielsen, Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction, Nature Catalysis 5(8) (2022) 662-672.
[25] F. Jiang, M. Li, J. Dong, Y. Yu, X. Sun, B. Wu, J. Huang, L. Kang, Y. Pei, L. Zhang, A general temperature-guided language model to design proteins of enhanced stability and activity, Science Advances 10(48) (2024) eadr2641.
[26] M.S. Weiß, U.T. Bornscheuer, M. Höhne, Solid-phase agar plate assay for screening amine transaminases, Protein Engineering: Methods and Protocols (2018) 283-296.
[27] O. Khersonsky, R. Lipsh, Z. Avizemer, Y. Ashani, M. Goldsmith, H. Leader, O. Dym, S. Rogotner, D.L. Trudeau, J. Prilusky, Automated design of efficient and functionally diverse enzyme repertoires, Molecular cell 72(1) (2018) 178-186. e5.
[28] D.M. Weinreich, N.F. Delaney, M.A. DePristo, D.L. Hartl, Darwinian evolution can follow only very few mutational paths to fitter proteins, science 312(5770) (2006) 111-114.
[29] M. Sternke, K.W. Tripp, D. Barrick, Consensus sequence design as a general strategy to create hyperstable, biologically active proteins, Proceedings of the National Academy of Sciences 116(23) (2019) 11275-11284.
[30] H.J. Wijma, R.J. Floor, D.B. Janssen, Structure-and sequence-analysis inspired engineering of proteins for enhanced thermostability, Current opinion in structural biology 23(4) (2013) 588-594.

[31] Z. Xu, Y.-K. Cen, S.-P. Zou, Y.-P. Xue, Y.-G. Zheng, Recent advances in the improvement of enzyme thermostability by structure modification, Critical reviews in biotechnology 40(1) (2020) 83-98.