MATERIALS SCIENCE

ForceGen: End-to-end de novo protein generation based on nonlinear mechanical unfolding responses using a language diffusion model

Bo Ni[1], David L. Kaplan[2], Markus J. Buehler[1,3]*

Through evolution, nature has presented a set of remarkable protein materials, including elastins, silks, keratins and collagens with superior mechanical performances that play crucial roles in mechanobiology. However, going beyond natural designs to discover proteins that meet specified mechanical properties remains challenging. Here, we report a generative model that predicts protein designs to meet complex nonlinear mechanical property-design objectives. Our model leverages deep knowledge on protein sequences from a pretrained protein language model and maps mechanical unfolding responses to create proteins. Via full-atom molecular simulations for direct validation, we demonstrate that the designed proteins are de novo, and fulfill the targeted mechanical properties, including unfolding energy and mechanical strength, as well as the detailed unfolding force-separation curves. Our model offers rapid pathways to explore the enormous mechanobiological protein sequence space unconstrained by biological synthesis, using mechanical features as the target to enable the discovery of protein materials with superior mechanical properties.

INTRODUCTION

Proteins present an elegant yet complex and rich design platform. The various functions and outstanding properties of proteins can be attributed to the folded three-dimensional (3D) structures, encoded by the underlying one-dimensional (1D) primary sequences consisting of about 20 naturally occurring amino acids (*1*). Through evolution, nature has demonstrated great success in "designing" proteins as a set of critical building blocks that constitute fundamental functions of all life, and specifically remarkable biomaterials, ranging from the structural hierarchies in collagens, complex assemblies such as silk, to tissue assemblies such as muscle and skin (*2*–*5*). In these various tissues and systems, the detailed mechanical signature—often, their response to mechanical pulling—is an essential feature for mechanobiology (*6*–*9*).

At the same time, there remains vast design space of mechanically optimized proteins yet unexplored by nature given the enormous possibilities of protein sequences (*10*). Hence, inspired by nature, discovering de novo proteins may unlock potentially unprecedented properties and functions (*3*, *10*–*16*). However, this enormous design space and costs associated with experimental testing present great challenges in finding effective tools to design de novo protein sequences that meet a set of interested functions or properties (*14*–*19*).

In recent years, the development of deep learning approaches and their applications to proteins have provided fast avenues for protein study and design. For forward problems focused on structure identification, deep learning–based tools such as AlphaFold2 (*20*) and RoseTTAFold (*21*) represent a breakthrough in achieving competitive accuracy with experimental methods in predicting 3D folded structures based on protein sequences at a much reduced cost. (*22*) Built upon these approaches, other protein folding tools [e.g., Omegafold (*23*), RGN2 (*24*), HelixFold-single (*25*), and ESMFold (*26*)] have been exploring the application of large language models. By removing dependence on multiple sequence alignments (MSAs) as the input, improvements in further reducing computational costs and achieving better predictions for orphan and rapidly evolving proteins have been demonstrated (*23*, *24*, *26*, *27*).

End-to-end models based on deep learning that predict various structural features [e.g., secondary structure type and content (*28*–*33*), binding sites (*34*), and surfaces (*35*)] and properties [e.g., solubility (*16*, *36*, *37*), melting temperature (*38*), natural vibrational frequencies (*39*, *40*), and strength (*41*)] for given sequences have also been reported. At the sample time, the inverse design of de novo proteins that meet desired structural or property features presents a more challenging task. On one hand, facing the enormous sequence space, search algorithms teamed with efficient deep learning–based forward predictors (*30*, *42*, *43*) may still suffer from inefficient exploration and the design accuracy and varieties of the discovered sequences are not easily controlled. On the other hand, recently emergent generative models (*44*–*49*) provide a direct map from the desired characteristics to potential designs and are becoming an emerging paradigm for various materials research and design (*50*–*55*), including proteins. For example, using an attention-based diffusion model trained on secondary structure data, de novo protein sequences can be generated based on secondary structure design objectives (*56*). However, these generative design models often focus on structural level design [such as secondary structures (*56*) or detailed protein backbone shapes (*57*–*60*)]. In contrast, development of generative models aimed at end-to-end design from property of interest to protein sequence remains rare (*61*).

Here, we focus on nanomechanical properties (*62*–*65*) of proteins. Thanks to the advent of single-molecule technology (*66*) [e.g., atomic force microscopy (AFM) (*67*, *68*), optical tweezers (*69*, *70*), and magnetic tweezers (*71*, *72*)], the measurement of protein unfolding under an applied mechanical force provides a unique molecular basis for understanding protein deformation (elasticity/plasticity) and

[1]Laboratory for Atomistic and Molecular Mechanics (LAMM), Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA. [2]Department of Biomedical Engineering, Tufts University, Medford, MA 02155, USA. [3]Center for Computational Science and Engineering, Schwarzman College of Computing, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA.
*Corresponding author. Email: mbuehler@mit.edu

fracture (*64*) and can play key roles in affecting some macroscopic mechanical properties of protein-based materials due to the inherent structural hierarchy. For example, via experimental measurements and theoretical analysis, it has been demonstrated that toughness of synthetic protein hydrogels can be correlated to the mechanical unfolding responses of the protein molecules and that mechanically strong folded proteins can result in tough hydrogel designs (*73*). Therefore, generating de novo proteins that meet desired mechanical unfolding responses can represent a key molecular level design step in protein-based material designs. Compared with previous protein design cases, this problem presents some unique challenges. First, this is a property-to-sequence end-to-end design task bypassing the structure level, which is expected to be more difficult than previous structure-to-sequence design tasks (*56*, *60*). Second, the available or affordable data on mechanical unfolding responses of known proteins (*74*) are rare when compared to those for protein structures (*75*) or sequences (*76*). Besides mechanical properties, we expect that these two challenges are also shared by many other property-to-sequence design tasks in proteins.

To address this problem, here, we combine an attention-based (*77*) diffusion model (*56*) with a pretrained large language model (*26*) for proteins to construct a generative deep learning model that predicts amino acid sequences and 3D protein structures based on mechanical unfolding responses as design objectives. In a singular workflow (Fig. 1), we start with performing a large series of full-atom molecular dynamics (MD) to simulate the mechanical unfolding process of Protein Data Bank (PDB) (*75*) proteins and recording the force responses (Fig. 1A). Then, we construct a protein language diffusion model (pLDM) by translating the protein sequences into a word probability latent space using a pretrained protein language model (pLM) and training a diffusion model to learn the map between sequence representations and the force-separation responses (Fig. 1B). At deployment, the trained pLDM predicts sequence candidates based on the given unfolding force conditions and the integrated folding algorithm (i.e., OmageFold) (*23*) determines the 3D structures of the resulting sequences. For validation, we compare the designed sequences with known proteins to analyze novelty (Fig. 1C) and test the designed proteins using MD to compare the mechanical properties and unfolding responses with input conditions. To prepare the design pipeline for further experimental validation, other properties key to experimental synthesis and testing, such as solvent accessible surface area (SASA) (*78*), solubility, or stability (*36*, *79*), can be estimated using available predicting tools (*36*, *78*, *79*) to further screen for preferred protein candidates (Fig. 1D). Through well-controlled comparisons, we demonstrate that our pLDM outperforms the vanilla diffusion model with or without an iterative design scheme. Built upon the property-to-sequence generation capability of our model and the broad potential of protein materials in achieving superior mechanical

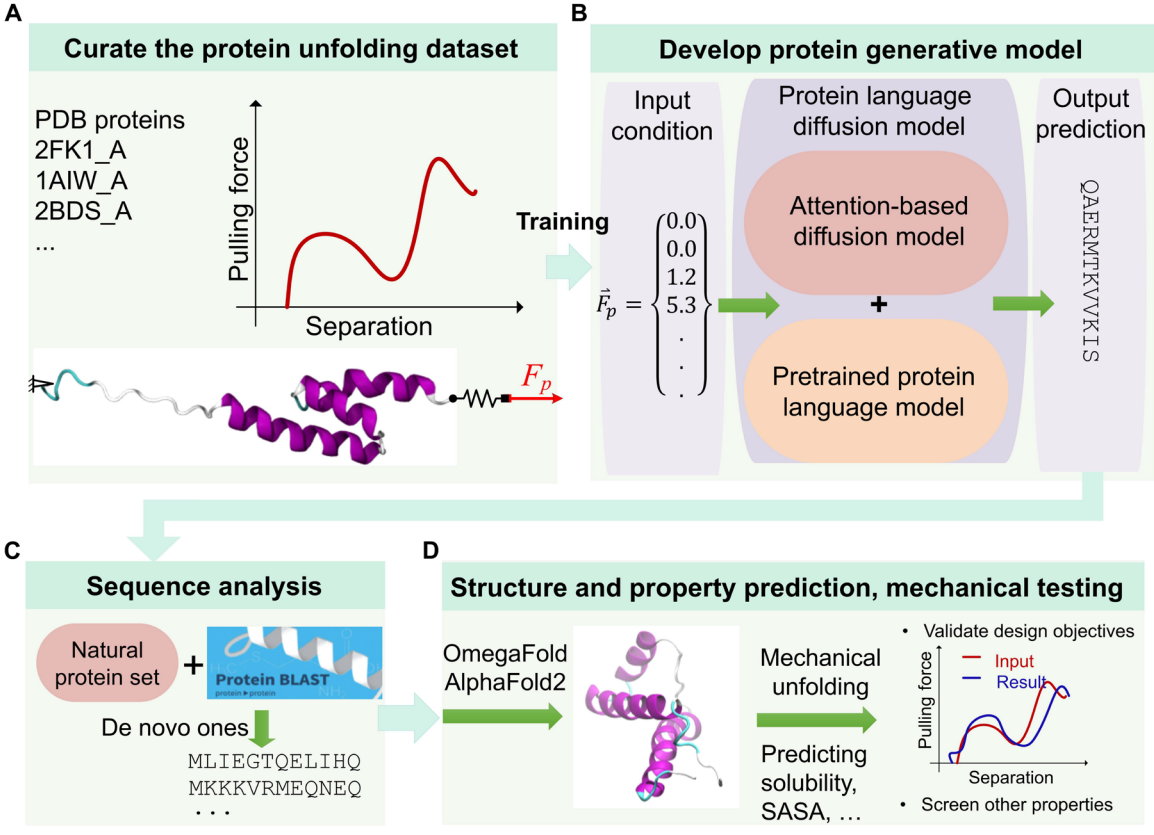

Fig. 1. Workflow of developing the end-to-end protein generation model. (**A**) Curating a PDB protein dataset on their mechanical properties by unfolding protein chains by force in MD simulations. (**B**) Overview of the conditioned protein language diffusion model (pLDM) developed here. (**C**) Analyzing the novelty of the generated protein sequences via protein-protein BLAST tests. (**D**) Validating the mechanical properties of the designed protein candidates using folding tools and mechanical unfolding tests and predicting other properties (e.g., solubility or stability) for further screening of the desired protein candidates.

properties, as well as other interesting properties (*80*–*83*) [e.g., optical (*82, 83*), electronic (*80*), energy storage (*81*), etc.], we expect that our end-to-end design model can be useful in numerous biological and engineering applications for the property-targeted generative design of various protein material systems.

RESULTS

Full-atom modeling of unfolding proteins by force

Inspired by single-molecule force spectroscopy (*67*), we simulate the unfolding process of protein chains under mechanical force to understand their mechanical properties at the molecular level. As shown in Fig. 2A, we start with PDB proteins with experimentally measured 3D structures. Using full-atom MD with the CHARMM force field (*84*) and a generalized Born implicit solvent model (*85*), we first relax the protein molecule at body temperature (i.e., 310 K) to reach equilibrium conformation. Then, we stretch the protein chain of N amino acids along the direction connecting the two chain ends (i.e., a and b) by fixing one end and steering the other with a spring (i.e., the segment between b and c in Fig. 2A) of a force constant $k = 0.5$ kcal/(mol Å^2) at a constant velocity $v = 0.1$ Å/ps. The pulling force, F_p, is recorded every 0.2 ps until the distance between two pulling ends, L_{ac}, reaches the contour length, L_{con}, of the protein chain, where we assume the average length of each amino acid is 3.6 Å (*86*) and $L_{con} = N \times 3.6$ Å. Further details on the MD simulations can be found in Materials and Methods. Movies of the unfolding trajectory of some selected PDB protein examples can be found in the Supplementary Materials.

In Fig. 2B, we smooth the raw force response (the red curve) to get rid of high-frequency fluctuations and get the unfolding response, $F_p(L_{ac})$, of the protein chain (the blue curve), from which we can identify the toughness and strength of the protein molecule using the unfolding energy T and the maximal value of force F_{max} defined as the following.

$$T = \int^{L_{con}} F_p dL_{ac} \tag{1}$$

$$F_{max} = \max_{L_{ac} \le L_{con}} \{F_p\} \tag{2}$$

To curate a dataset based on naturally existing proteins, we use the Biomolecule Stretching Database (BSDB) (*74*) as guidance and select 7026 PDB proteins that have no gaps in their experimentally determined structures and consist of no more than 128 amino acids.

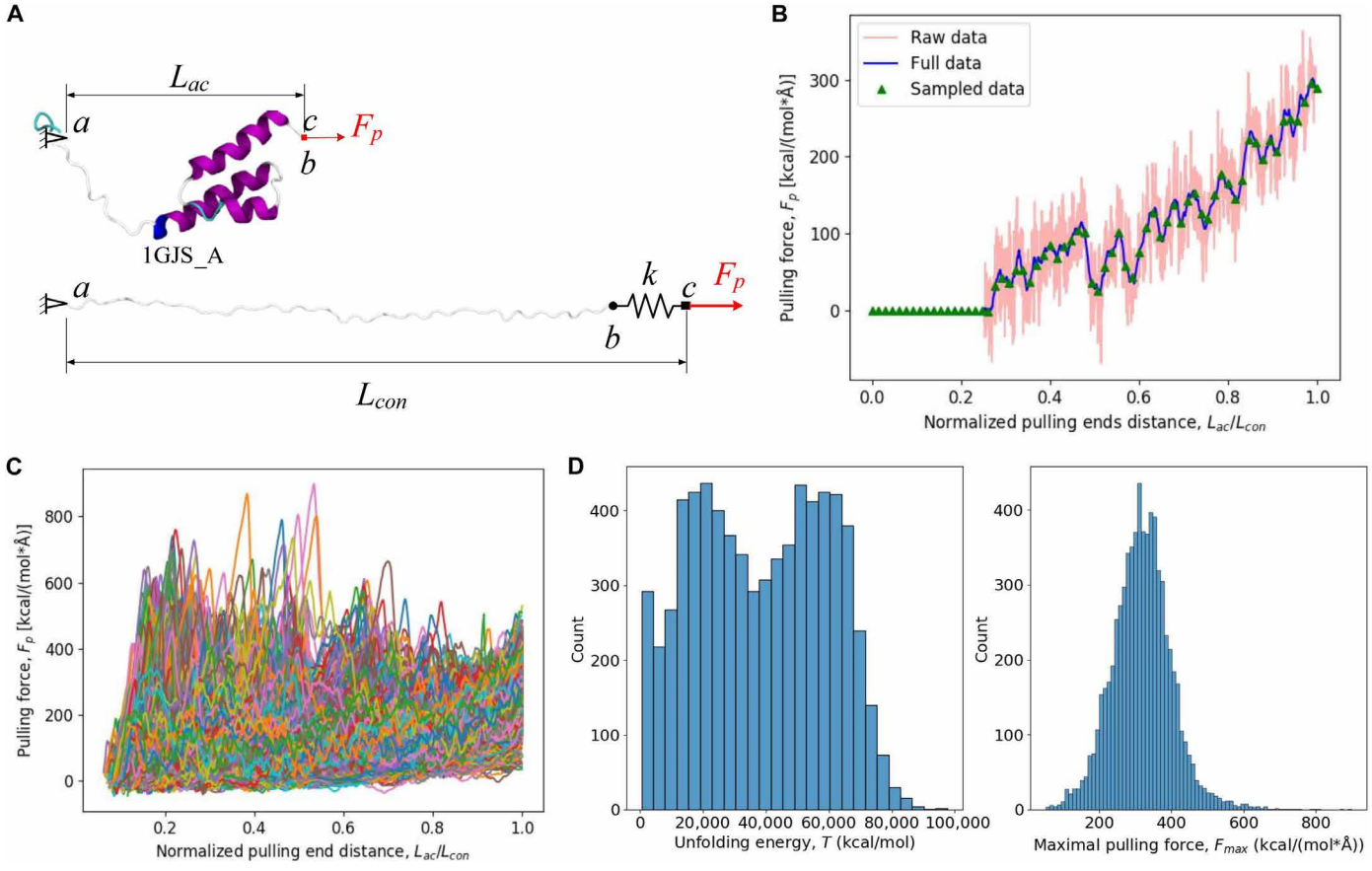

Fig. 2. Mechanical unfolding of proteins and mechanical properties dataset curation. (**A**) Full-atom simulation of mechanically unfolding a PDB protein chain using steered MD. (**B**) Collecting (red data) and smoothing (blue data) the pulling force history during the whole unfolding process and converting it into a vector representation (green triangle dots). (**C**) Collecting pulling force curves during mechanically unfolding for a large member of PDB proteins. (**D**) The distributions of the unfolding energy (left) and maximal pulling force (right) for the PDB protein training set developed here.

Then, we collect their structures directly from the PDB (*75*) and apply the protocol above to test their mechanical unfolding responses. An overview of the distributions of the unfolding responses and mechanical properties are shown in Fig. 2 (C and D). Specifically, in Fig. 2D, the unfolding energy or toughness shows a bimodal distribution while the strength presents a unimodal one; in Fig. 2C, one can observe that there exist various unfolding responses among proteins. For example, the maximal force may appear as the peak in the middle of unfolding process or near the end when reaching the contour length, which may indicate very different deformation mechanisms. An in-depth study of these mechanical properties and their dependence on the structural features and sequences for a large number of proteins is of great research interest and should be carried out in the near future. Further insight can also be obtained from numerous experimental studies using AFM-based force generative outcomes during unfolding states of proteins. Here, we focus on applying these freshly collected data to develop protein generation models. To efficiently label the pulling force responses during the full unfolding process, we introduce a pulling force vector $\vec{F}_p$ (green triangles in Fig. 2B) to represent the full response as the following

$$\vec{F}_p = \{F_p(L^i_{ac}): i = 0, 1, 2, \ldots, N\} \quad (3)$$

where N is the sequence length of the protein chain and we sample the pulling force when the distance between the pulling ends reaches $L^i_{ac} = i \times L_{con} / N$. For L^i_{ac} that is smaller than the value of the initial equilibrium conformation, we simply define the force values as zeros. Such a vector representation can adjust to the protein sequence length automatically; that is, longer/shorter proteins with potentially more/fewer unfolding details have more/fewer sampling points evenly distributed. Next, we develop DL models to generate protein sequences that meet the given mechanically unfolding responses represented in terms of the pulling force vector $\vec{F}_p$.

pLDM and inverse design for mechanical signatures

To solve the conditioned protein design task, here we develop a pLDM by combining a pretrained pLM (*26*) and an attention-based diffusion model (*56*). Figure 3A depicts an overview of the model developed in the present work. The pLM (on the right of Fig. 3A) is pretrained on large amounts of protein sequences data (*76*) to form

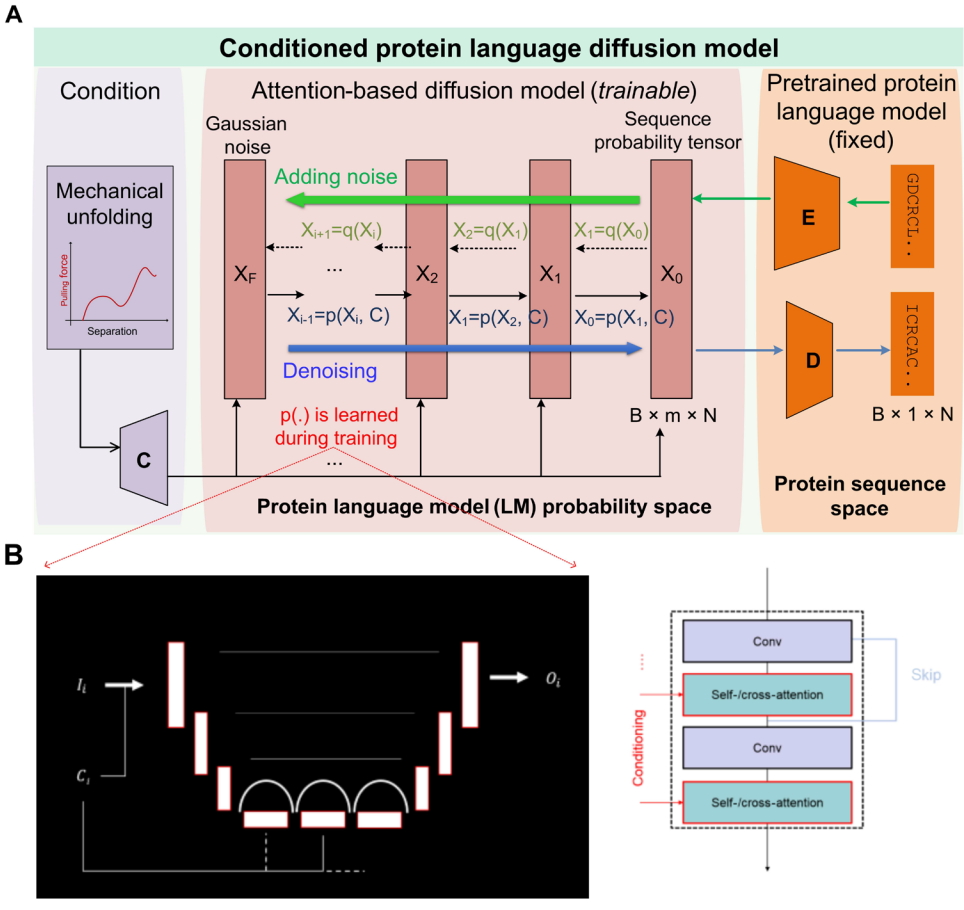

Fig. 3. Overview of the pLDM. (**A**) Structure of the developed model, pLDM. It combines a protein language model (pLM) pretrained on large protein sequence data and a trainable attention-based diffusion model. We use the pretrained pLM to translate protein sequence representations between the tokenized sequence space and the word probability latent space. The diffusion model, with a building block of a 1D U-net, is trained to predict the noise added at each diffusion steps, thus gradually removing them to generate meaningful sequence representations at deployment. (**B**) Depiction of the 1D U-net architecture that translates an input I_i into an output O_i under a condition set C_i. The model features 1D convolutional layers, as well as self-/cross-attention layers as shown on the right.

internal representations that better understand not only sequences but also structures and properties of proteins (*33*). We leverage this knowledge by applying pLMs to translate protein sequences from the tokenized sequence space into the word probability latent space. Then, we train a diffusion model developed in the previous work to operate in this probability latent space. The diffusion model is built upon a 1D U-Net architecture with attention mechanisms (Fig. 3B). At deployment, starting with the given condition (on the left of Fig. 3A) and random signal seed, the diffusion model predicts and removes the noise at each step and produces meaningful sequence probability tensors, which are then translated back into protein sequence using the fixed pLM. There are multiple choices for the pretrained pLM and larger pLMs require higher computing resource and cost (*26, 33, 87*). For computational efficiency, in the following, we focus on the results that adopt a medium-sized pretrained model with 150 million parameters from the ESM-2 series (we find that this yields improved performances as is shown in the discussion in the next subsection) (*88*).

Once the model has been trained, we demonstrate the performance of the developed pLDM by testing it with various mechanical unfolding responses, including those that come from naturally existing proteins and those that are de novo. The generated sequences are folded into 3D structures using OmegaFold (*23*) and then undergo the same mechanical unfolding tests using full-atom MD simulations. With protein BLAST (*89*) test and comparing pulling force responses with the input, we examine the novelty of the generated sequences and the accuracy of the protein design.

For protein design with the mechanical unfolding responses that correspond to naturally existing proteins, we test the model with the pulling force records of PDB proteins in the test set, with which the model has not been trained. Figure 4 shows some examples of the designed proteins and their mechanical unfolding responses. In terms of the design target, the conditioned pulling force paths (red curves) in Fig. 4 (A to F) represent a variety of different patterns, including simple ones that show that the pulling force nearly keeps increasing during the unfolding process (Fig. 4D), the examples that

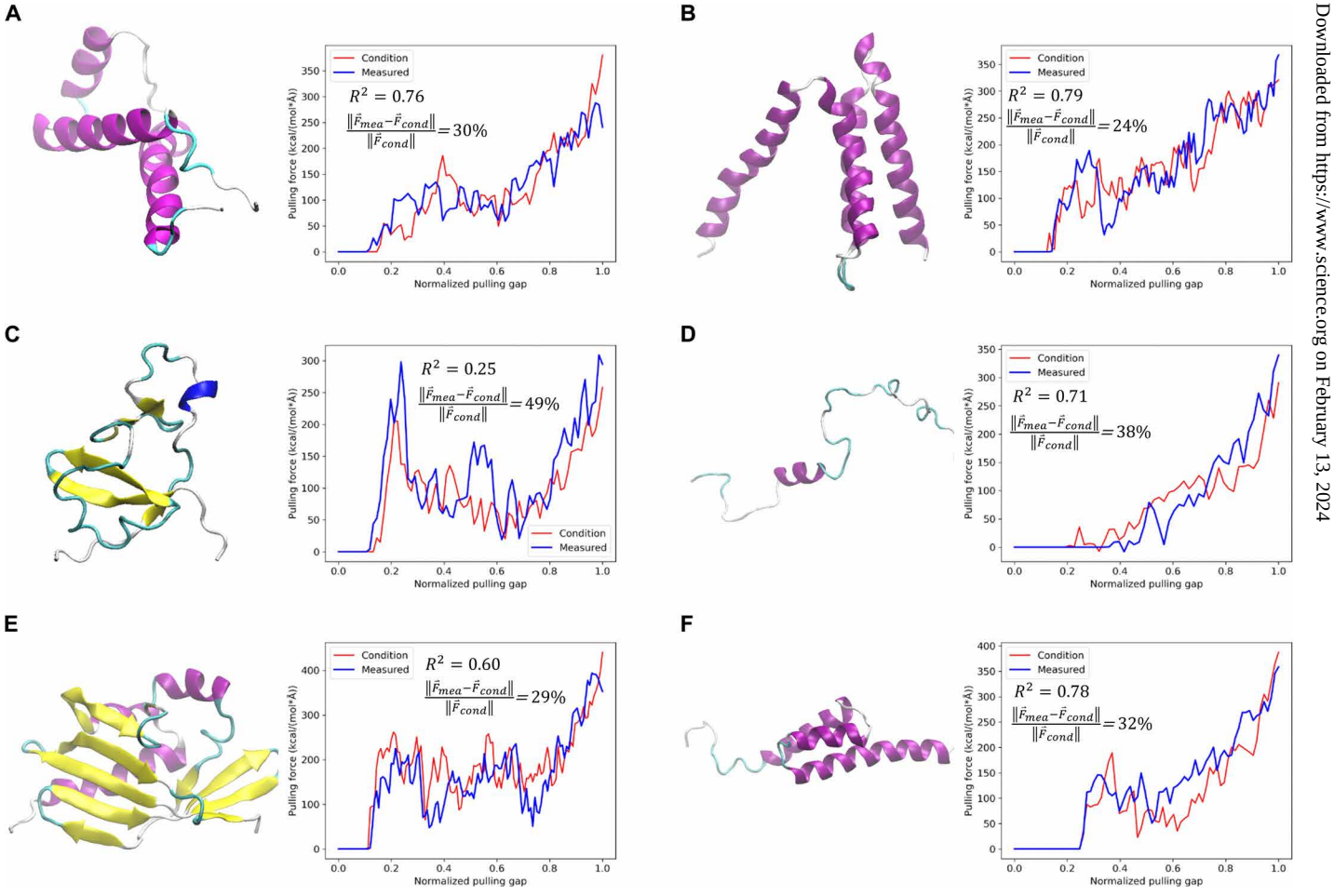

Fig. 4. Results for protein generation based on mechanical unfolding responses of naturally existing proteins. (**A**) to (**F**) show a variety of representative cases of different unfolding force paths (red curves), including the one that nearly keeps increasing (D), the ones that show local peaks during an overall increasing trend (A, B, and F), the one that meets an oscillating plateau and then increases (E), and the one that reaches a high peak in the early stage (C). The proteins generated by our model demonstrate pulling force patterns (blue curves) that follow the trend of design objectives. Because of the complex and highly oscillating nature of pulling force response during mechanical unfolding of proteins, we use R^2 value and relative L_2 error (listed in each panel) to measure the accuracy of the design in following the overall trend and quantitative values. Corresponding to the various pulling force responses, the generated proteins show a variety of structures, including high α helix content (A, B, and F), a mix of β sheets and random coils (C), a mix of an α helix segment and random coils (D) and a mix of β sheets and α helices (E).

show local peaks in early stages of unfolding during an overall increasing trend of the pulling force (Fig. 4, A, B, and F), the ones that reach an oscillating plateau and then increase (Fig. 4E), and those that achieve high peak in the initial stage of unfolding (Fig. 4C). Despite this complexity of protein unfolding scenarios, and the oscillating nature of pulling force, the proteins generated by our model demonstrate pulling force responses (blue curves) that in general closely follow the design objectives. We use multiple metrics, including R^2 and relative L_2 error (see Materials and Methods for details), to measure the accuracy of design in meeting the overall trend as well as quantitative values of the mechanical responses. Corresponding to the various patterns of pulling force responses, the generated proteins also show a variety of internal structures. For example, a relatively simple combination of an α helix segment and random coils without complex entanglement or close spatial packing in Fig. 4D produces an unfolding process with consistently increasing pulling forces with few oscillations. An entangled mix of α helices and β sheets in Fig. 4E yields an unfolding pulling force path with a plateau region of strong oscillations. Unfolding of the β sheets in Fig. 4C can be related to the high local peak in pulling force history. In Fig. 4 (A, B, and F), α helix segments of different lengths and spatial arrangements can produce similar trends but different quantitative pulling force records. More details on the mechanical unfolding process of some of these cases, including cases A, C and E, can be found in the MD trajectory movies in the Supplementary Materials.

On the novelty of these generated proteins, we apply basic local alignment search tool (BLAST) analysis (*89*) to the predicted amino acid sequences to access whether, and to what extent, they represent de novo sequences or closely related forms of known proteins. Table 1 shows the results of the BLAST analysis for the various cases listed in Fig. 4. We find that even though the input design targets are from existing PDB proteins, many of the generated protein sequences (cases shown in Fig. 4, A to D and F) do not match any sequences in the database of known proteins with standard BLSAT analysis (i.e., returning "no significant similarity found" in protein BLAST test) and are de novo ones. The model can also produce sequences (e.g., case E in Fig. 4) that show some similarity to the existing proteins. However, the most similar example found (i.e., 8CH0) is not included in the training and testing set. While the model is only trained on a very small portion of PDB proteins, with the pulling force corresponding to existing PDB proteins as an input, we expect the possibility of the model "rediscovering" sequences that show some similarities to the

Table 1. Results of the BLAST analysis and predicted solubility for the various generated proteins (from Fig. 4) based on existing mechanical unfolding responses. Given the pulling force vectors of existing proteins as the design condition, the model still shows high probability in predicting sequences that show little similarity to existing proteins as can be seen from the BLAST results (A to D and F). For other cases, sequences with some similarity to known proteins can be predicted (E). The predicted solubility is scaled with the experimental dataset with a population average of 0.45. The listed solubility values are all larger than 0.45, thus predicting to have a higher solubility than the average soluble *E. coli* proteins in the experimental dataset (*90*).

Case	Sequence	BLAST result: the sequence producing the most significant alignment		Predicted solubility
		Among PDB proteins	Beyond PDB proteins	
A	MLIEGTQELIHQKLAKGKT-VLVQRYVAKGLQVDDNTEDL-LANAKNYLNPDQIERSIAYAQK-IEEMEGDDMFKVALV	–	No significant similarity found (NSSF)	0.939
B	MKKKVRMEQNEQKKQVY-QELNDKVENDEALAPKS-VALYIAALKEKEEAGKIPHHF-NLLERLKLTITSCRFFLLKIQN-NDTKLQKRRKFIDETIQLAREI-YEKQDNK	–	NSSF	0.876
C	MGKITPVVLAGGKQK-EDEETLDGGEILTKDG-KTLKLISDAQVAVMN-VKQVQEGTYEGSQVIEEDG-VRGNYVSYVGK	–	NSSF	1.000
D	GSSGSSGRDVTQQTNKCCR-RCSRKPHCCIKAWRPRSSD-LYYHEKHTHSGPSSG	–	NSSF	0.676
E	MNTPEHMTAVVQRYVAAL-NGGDLDGIVALFADDATVED-PVGFQNVSGKAADANFYESP-GFLDLVKALTGPVRAFGNEK-FFAMIVFFEYEGTKTVVAGI-DHIRFNGAGKVVSMRAYF-DEKNIHASA	100% query cover, 73.6% identical with 8CH0	–	0.774
F	MPWHHHGSSGLVQTG-MAATGLKDFIVEAYPKKPD-DIIKVCRSEPSAGYWWCED-VQNEVKQKCLSKKQRQVKAQ	–	NSSF	0.606

known proteins. Further measures may be utilized to boost the novelty of design for such cases, including multi-shot design and selecting the best one based on BLAST results. In current work, we focus on understanding the performance of the current model.

Besides examining individual cases, we also show the distributions of design accuracy and novelty for a larger number of testing cases. Figure 5 demonstrates the results of 187 generated proteins based on various pulling force conditions from the standalone test set. On the mechanical unfolding responses, the R^2 and relative L_2 error between the measured pulling force vectors of the generated proteins and the input conditions among cases show unimodal distributions with median values of 0.56 and 0.36 (Fig. 5, A and B), respectively. The distributions indicate that for many of the cases, the designed proteins follow the input conditions in terms of the trend and value reasonably well during the whole mechanical unfolding process. However, as demonstrated in individual cases (Fig. 4), it remains challenging for designed proteins to precisely follow the input pulling force values at each unfolding step. This can also be seen by comparing components of pulling forces of all cases together in Fig. 5C. While conditioned input values of pulling force components and the measured ones based on the generated proteins in general share the same trend (that is, the distribution centers at the $y = x$ dashed line as the visual guide in Fig. 5C), the finite width of the distribution cloud deviating away from the ideal case indicates that for individual components, there could be considerable mismatches.

The limited component-wise accuracy demonstrates the difficulty and challenge of designing proteins based on detailed mechanical unfolding responses, even with the current model. At the same time, the proteins generated by our model still show reasonable agreement between the achieved and the conditioned mechanical properties, including toughness (Fig. 5D) and strength (Fig. 5E). Strength defined as the maximum of the pulling force shows an R^2 value of 0.41 (Fig. 5E), slightly smaller than that of the pulling force components (0.54 as listed in Fig. 5C). At the same time, an R^2 value of 0.93, much higher than that of pulling force components (Fig. 5C), was observed for toughness (Fig. 5D), which is defined as the unfolding energy over the whole unfolding process (Eq. 1). This difference in R^2 values indicates that when the entire unfolding process is considered, the component-wise error tends to cancel each other and the designed proteins follow the input conditions in terms of toughness more sensitively. On the novelty of the designed proteins, Fig. 5F shows a bimodal distribution of the highest percent identity found via protein BLAST analysis for all the generated sequences. The highest peak (on the left in Fig. 5F) corresponds to the cases where the generated proteins have little similarity to the existing/known ones and are totally de novo. There also exists the other weaker peak

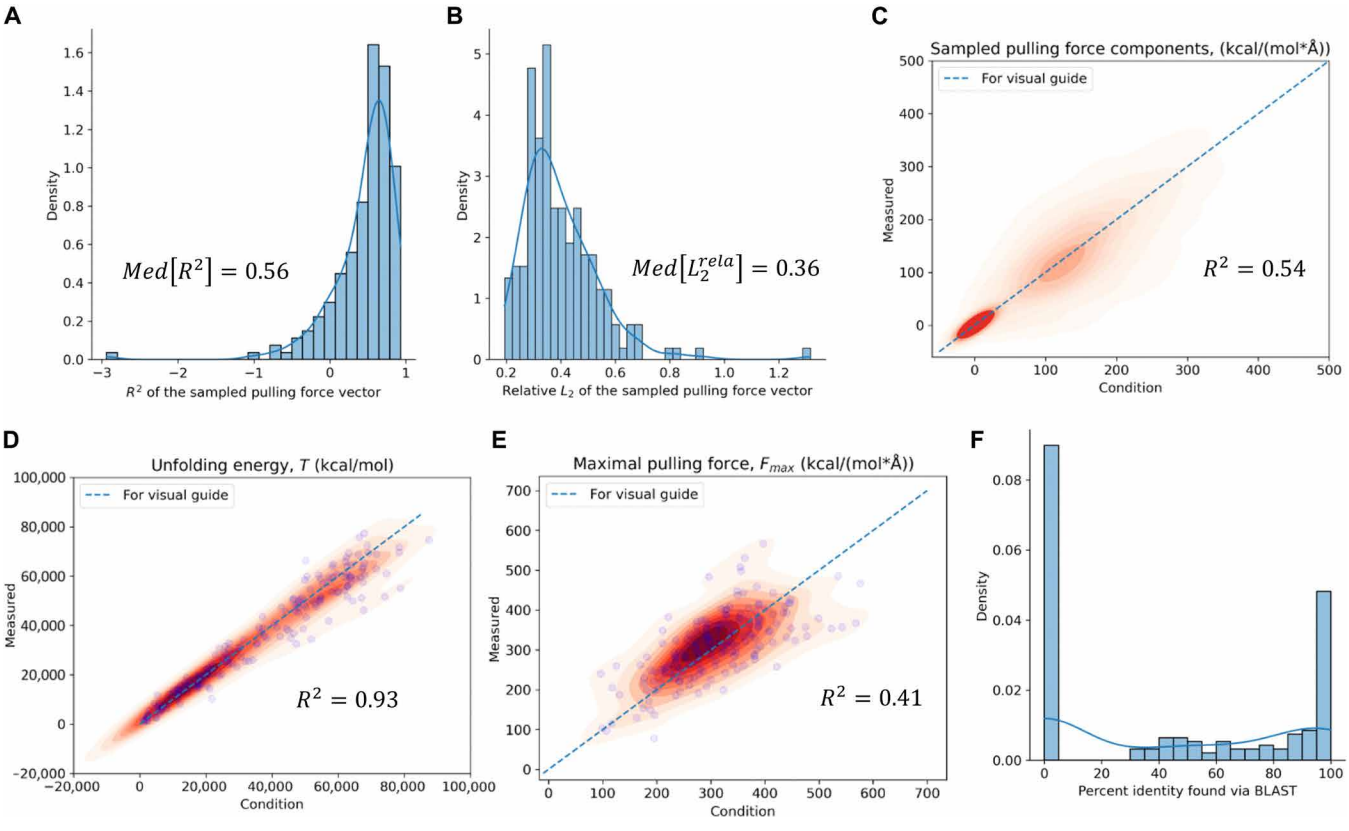

Fig. 5. Overall quality of generating proteins based on mechanical unfolding responses that correspond to naturally existing proteins in the test set. We test the model with mechanical unfolding responses from 187 proteins in the standalone test set. On the pulling force response, (**A**) and (**B**) show the distributions of R^2 (A) and relative L_2 error (B) for comparing the pulling force response of each designed protein with the input condition while (**C**) shows the comparison in terms of pulling force components for all testing cases. On the overall mechanical properties, (**D**) and (**E**) compare the designed proteins with input conditions in terms of unfolding energy (i.e., toughness) and maximal pulling force (i.e., strength). On the novelty of the designed sequences, (**F**) shows the distribution of the highest percent identity found via BLAST test.

on the right for cases in which the generated proteins are similar to the known ones. The bimodal distribution echoes the result of individual cases listed in Table 1, and the relative height of the two peaks indicates that our model has a stronger tendency in generating de novo sequence designs.

To further boost the novelty of designed sequences, different measures could be taken, including how the model is applied and how the input conditions are constructed. Here, we discuss one possibility of using de novo mechanical unfolding responses, i.e., pulling force vectors that do not correspond to existing proteins, as the input. There is still a lack of complete rules on how to identify or construct physically achievable pulling force vectors for all possible amino acid sequences. To explore the de novo mechanical unfolding responses, here we start with existing pulling force vectors and mute them using a mixing scheme. As shown in Fig. 6A, two pulling force vectors from PDB proteins 2AAN and 1KN7 were chosen and show different patterns during the unfolding process, one undergoing an oscillating plateau in the early stages while the other continues to increase. These differences in pulling force can be attributed to the different internal protein structures and sequence length. 2AAN shows a high β sheet content in a closely packed conformation. In comparison, 1KN7 has a mix of α helix segments and unstructured coils with a more open spatial arrangement. We mix the pulling force vectors of each of the proteins with respect to the same normalized pulling end gap at different ratios (i.e., 1/3 to 2/3 for mix 1 and 2/3 to 1/3 for mix 2) and then use these as the de novo mechanical unfolding response to generate proteins with our model. In addition, in our model, the length of the generated sequence (i.e., N) can be controlled through the length of the input pulling force vector (i.e., $N + 1$). Here, we intentionally choose $N = 99$, which is different from that of both base proteins when constructing the de novo input pulling force vectors. The results of the proteins designed with these de novo pulling force vectors are shown in Fig. 6 (B and C). Similar to previous designs listed in Fig. 4, the mechanical unfolding responses of the designs follow the input conditions, even though this time they come from a mix of proteins of different structures and mechanical properties. Moreover, the designed proteins adopt internal structures of a mix of α helices with different compactness, which show little similarity to those of the base proteins. Another set of examples is shown in Fig. 6 (D to F) with base proteins of different internal structures. Again, the designed proteins fulfill the targeted de novo mechanical unfolding responses reasonably well.

Comparing the structures of the generated proteins including the ones in Fig. 6, one can raise intriguing questions about the underlying relation between protein structure and mechanical properties. For example, as seen in in Fig. 6, none of the generated proteins have β sheets even though some of the base proteins do. Instead, many of the designed proteins are composed of α helices and β turns and packed in a distinct conformation. As the unfolding force response may be related to individual secondary structures unfolding, as well as the interactions between secondary structures, one may examine whether the model has learned to search for certain topological features, like different compactness, to generate the desired force-separation patterns. To investigate this possibility, we calculate the SASA using dr-sasa (*78*) based on the relaxed protein structure. This helps us to assess the compactness of the folded structures for both generated and base proteins. The SASA and their values normalized with respective to the sequence length are listed in table S1. Among the base proteins in the mixing pair (Fig. 6, A and D), the one with a higher unfolding force level (base protein 1 in Fig. 6A or base protein 3 in Fig. 6D) includes closely packed β sheet structures with a normalized SASA smaller than that with a lower unfolding force (base protein 2 in Fig. 6A or base protein 4 in Fig. 6D), and with loosely packed internal structures. Among the mixtures (Fig. 6, B and C, or Fig. 6, E and F), similar trends are observed: A higher unfolding force level indicates a relatively smaller SASA and more

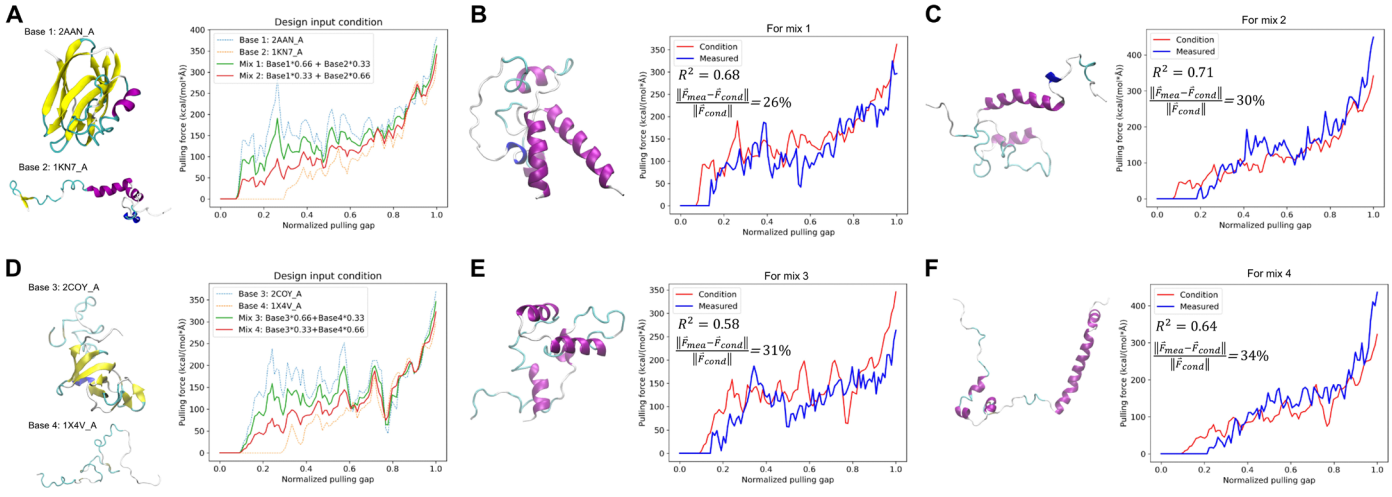

Fig. 6. Results for protein generation based on de novo mechanical unfolding responses. For de novo design input, in (**A**) and (**D**), we construct mechanical unfolding responses by mixing existing ones at a different ratio and changing the targeted sequence length. Here, we intentionally choose pulling forces of different patterns, one that monotonically increases (base 2 in A and base 4 in D) and the other that first reaches a plateau then increases (base 1 in A and base 3 in D) as the bases to mix. By choosing the mixing ratio, the de novo inputs cover a transition between these two pulling force patterns. (**B**), (**C**), (**E**), and (**F**) show the results of the generated proteins. On the pulling force, the generated proteins follow the design input closely in terms of both the overall pattern and quantitative values throughout the unfolding process. On the structure, interestingly, the generated proteins show a mix of α helix and random coil segments, which is different from some of the base proteins of high β sheet content. On the sequence length, the generated proteins have a length that is different from both base proteins.

closely packed internal structures (Fig. 6, B or E). However, between the mixes and the base pairs of proteins, no clear relation in terms of SASA or normalized SASA can be identified. The mix can have a normalized SASA falling in between the base proteins or can be larger than both bases. Therefore, systematic generation of more examples and in-depth study of the patterns and relationship of their structural and mechanical properties would likely be required to gain additional understanding of protein sequence-structure-property relationship as well as the learned tendency of our model, which can be done in future work, along with additional experimental validation.

With the obtained de novo pulling force responses as the input, it is interesting to investigate the novelty of the generated proteins. As shown in Table 2, all of the designs have no significant similarity among known protein sequences according to protein BLAST analysis. The case study shown here demonstrates that our generative model can also be used as an effective tool to explore the unknown possibilities in property space of mechanical unfolding responses by constructing de novo input conditions. In return, it is more probable to discover de novo designs of proteins.

Besides using the mixing scheme demonstrated above, other ways can be used to create or discover de novo mechanical unfolding responses as the design input. One possibility is to train a generative model (e.g., a diffusion model) on the pulling force data curated here. At deployment, mechanical unfolding responses can be generated with or without input conditions. Because the possible pulling force vectors should in principle be determined by the sequence-structure-property relationship of proteins, in-depth investigation on this process deserves a separate study, where the generation model developed here can serve as an effective tool in navigating the complex design space for proteins.

To prepare the protein design pipeline for further experimental testing, we can also integrate available predicting tools (*36, 78, 79*) to estimate properties of generated proteins that are key to experimental synthesis, manipulation, and test. For example, by applying the solubility predictor based on protein sequences, we can screen generated proteins for desired solubility. As listed in the last column of Tables 1 and 2, surprisingly, we find many of the generated proteins are predicted to have a solubility higher (i.e., larger than 0.45) than the average soluble *Escherichia coli* protein from the experimental solubility dataset (*90*), which indicates encouraging potential for successful experimental synthesis, use, and manipulation.

Benefits of using pLDM in small-data generation tasks

Combining the above results of protein designs based on existing or de novo mechanical unfolding responses, we have demonstrated that the pLDM developed here can achieve reasonably good design accuracy for both detailed unfolding pulling force history and overall mechanical properties. At the same time, many of the generated sequences are totally de novo, having no significant similarity among any existing or known proteins. This is surprising considering the

Table 2. Results of the BLAST analysis and predicted solubility for the various generated proteins (from Fig. 6) based on de novo mechanical unfolding responses. Given the de novo pulling force vectors as the design condition, the model shows high probability in predicting de novo sequences that show little similarity to existing proteins as can be seen from the BLAST results (for Mix 1 to 4). The predicted solubility is scaled with the experimental dataset with a population average of 0.45. The listed solubility values are all larger than 0.45, thus predicting to have a higher solubility than the average soluble *E. coli* proteins in the experimental dataset (*90*).

Case	Sequence	BLAST result: the sequence producing the most significant alignment		Predicted solubility
		Among PDB proteins	Beyond PDB proteins	
For mix 1	MDDALLLKAMQQLLLA-PIRVKEDDPLVRRDAAIG-FAPDGVRVDFEYTAKVD-LAKATLDEVGLKGANTTQP-PRPIANKLPPPIVVLASKLLEI-YKELKQL	–	No significant similarity found (NSSF)	0.742
For mix 2	GSSGSSGYGRQVTTR-SPRETTSLSFDIDREPMEFN-QLKAQELMMPFNLKALDT-GRFNRPLQFVEQAKGK-MEKALLKKATDPVQALPK-KDLSEGISPKMG	–	NSSF	0.814
For mix 3	MMDTPKLMDELKDYAPQPAR-RALNLTNPRTAAVPEKTG-DDVAPFFDHAAEKENLGF-HHEVANDNWESEAKFLK-LTKVPVSPQVIYNAAGLL-FEAARKTP	–	NSSF	0.761
For mix 4	GSSGSSGPSTYKNPG-DRFFTTSYFTDPELEAGQFE-VRTKDKMLNGITLLQQKP-CGKSCELFVDQNKKAVEEK-KKKLMLTQMAAYYQQDDLT-MASGPSSG	–	NSSF	0.657

fact that (i) the design task is challenging, and (ii) the training data used are relatively small compared with previous work (*56*). Specifically, the design task here is to map detailed pulling force responses directly to protein sequences, bypassing predicting the 3D structural transition of protein chains during the unfolding process. On the basis of MD simulations and experimental studies, it is clear that the unfolding process of protein structures under mechanical forces corresponds to complex uncoiling of the backbone and breaking of hydrogen bonds within the 3D hierarchical structures of proteins. Theoretically, we can expect that this design task could be more challenging than designing protein sequences based on structural conditions (e.g., secondary structure types). At the same time, for the pLDM discussed above, which designs sequences with a length, N, smaller than or equal to 126, it was trained with a dataset of 6863 proteins, mainly limited by the high computational cost to curate the data on mechanical properties. The size of the dataset is small compared with the PDB proteins of known structures (i.e., about 13,644 for single-chain protein with $N \leq 126$), and even further dwarfed by the number of all possible protein sequences (i.e., more than 20^{126}).

Given such a challenging protein design problem with a limited small set of labeled data, we discuss the strength of the current pLDM by comparing it with the protein diffusion models (pDMs) developed previously (*56*). The structure of the adjusted pDM is summarized in fig. S1A. The main difference between pLDM (Fig. 2B) and pDM is about the space where the diffusion/denoising processes occur during the training/inferring. While the pDM operates in the tokenized protein sequence space, pLDM performs in the word probability latent space formed by pretraining the pLM on the available protein sequence dataset (i.e., UniRef50 dataset for ESM-2 pLM adopted here), which is much larger than the datasets on protein structures or mechanical properties. For various downstream prediction tasks, the adoption of the pretrained pLM as a base model has proven to be beneficial (*33*). For example, the ESM-2 model adopted here has been used for protein folding prediction in ESMFold (*26, 87*) and achieved comparable accuracy with the folding tool using MSA as the input (e.g., AlphaFold2) but at a lower computational cost. At the same time, for inverse generation, the integration of a large language model and diffusion model has proven to be beneficial in conditioned generation tasks for images (*48, 91*) and texts (*92, 93*) [e.g., Diffusion-LM can achieve fine-grained controls on syntactic structure for text generation (*93*)]. However, the applications of pretrained pLM in the generative tasks of protein design remain rare (*94*) and the potential benefits are to be explored. Here, the mechanical unfolding response of the conditioned protein design task and our results provide a concrete example to study this aspect of the process.

To compare with pLDM, two alternative models (AMs) based on pDMs were constructed. The first one, AM1, uses one pDM and designs the protein sequence for a given pulling force vector in one shot, similar to the pLDM developed above. The second AM, AM2, consists of two pDMs as the protein designer and protein predictor separately. The designer is the same as the first pDM model, AM1, while the predictor is trained to predict pulling force vectors based on the given protein sequences. At deployment, the protein designer iteratively generates sequence candidates and the protein predictor then evaluates them to pick the most accurate one among five attempts, forming an on-the-fly iterative design scheme (fig. S1B). Both models are trained on the same dataset discussed earlier and tested with the same generation task based on existing mechanical unfolding responses from the test set. A summary of the performance of the two pDM-based models together with those of the pLDM is listed in Table 3. We use various metrics, R^2 and relative errors, to evaluate the design accuracy of the models on the whole test set, considering not only the overall mechanical properties (i.e., toughness and strength) but also the detailed mechanical unfolding responses in terms of the pulling force vectors. For most of the comparisons (9 out the 11 rows), pLDM achieves the best performance (e.g., the highest R^2 or the lowest error). This result demonstrates that, by

Table 3. Performance comparison between the current pLDM with one-shot prediction (last column) and the protein diffusion model with one-shot prediction (third to the last column) as well as iterative predictions (second to the last column). Tested with the existing unfolding responses from the test set, the pLDM shows an overall better performance in fulfilling the design target by achieving the best results (indicated with the underline, the minimum for errors and the maximum for R^2) in most rows considering mechanical properties as well as the detailed pulling force responses.

Performance on the test set				AM1: pDM with one-shot generation	AM2: iterative prediction with pDM-based protein designer and predictor	The current model: pLDM with one-shot prediction
Toughness (i.e., unfolding energy)		R^2		0.86	0.87	0.93
		Relative L_1 error	Mean	0.151	0.150	0.147
			Median	0.121	0.123	0.102
Strength (i.e., F_{max})		R^2		0.09	0.17	0.41
		Relative L_1 error	Mean	0.188	0.164	0.188
			Median	0.151	0.113	0.149
Pulling force vectors	As vectors	R^2	Mean	0.427	0.418	0.452
			Median	0.526	0.522	0.563
		Relative L_2 error	Mean	0.399	0.402	0.398
			Median	0.377	0.382	0.362
	As components	R^2		0.476	0.499	0.537

integrating the pretrained pLM, pLDM achieves enhancement in producing more accurate designs even when being trained only on a relatively small dataset. This advantage of pLDM with respect to pDM can be helpful for other property-to-sequence design tasks for proteins, given that usually data on protein mechanical or other properties are more limited or costly to collect.

DISCUSSION

Generating de novo proteins based on their mechanical unfolding responses presents unique challenges in property-targeted protein design. For rational design strategies, it is hard to grasp the complex relationships between sequences, structures, and properties. For data-driven methods, the labeled data on the mechanical properties are often costly to collect and limited in number, especially given the enormous possibilities in protein sequence space.

Here, we have developed a pLDM as an effective tool to tackle these challenges and generate de novo proteins that meet the mechanical properties' design objectives in an end-to-end manner. The pLDM developed combines the pretrained pLM and diffusion model as key components and leverages the strength of both. The pLM part is pretrained on the abundant protein sequence data and thus provides an effective representation of protein sequences in its latent space. The diffusion model part only operates in this latent space and learns the map between detailed pulling force responses and the sequence representation using only a relatively small set of data curated by performing full-atom simulations. By examining the unfolding details of individual designs and the statistics of the mechanical properties of many cases, we demonstrate that the proteins designed by our model meet the targeted overall mechanical properties, including toughness and strength, as well as the detailed unfolding force vectors with reasonably good accuracy. Moreover, the sequences generated are mostly de novo, sharing very limited similarity with existing/known proteins. Given the mechanical unfolding responses from known PDB proteins as the design input, our model still shows a strong tendency in discovering de novo proteins as alternatives. Constructing de novo unfolding responses as the input via a mixing scheme further boosts the probability of generating de novo designs. Finally, through controlled comparisons, we show that the pLDM outperforms the vanilla pDM with or without an iterative design scheme in achieving better design accuracy, thus clearly demonstrating the benefits of combining pretrained pLM and diffusion model in the pLDM developed here. A short summary of these key aspects about the pLDM is listed in Table 4.

As the initial steps of developing property-to-sequence generative models for de novo protein design, here we adopt the force-separation curves collected from MD simulations, in the hope of achieving consistency and relevance, avoiding bias from simplified models, and curating sufficient data points for DL model training. A few clarifications on the mechanical unfolding response data are included in the following.

First, there exist differences as well as commonalities between MD results and experimental measurements of the force-separation curves of protein unfolding that deserve more nuanced discussion. The mechanical unfolding process of proteins often involves entropic elasticity, a transition to energetic elasticity, and bond breaking (*95*). Thus, the force response often shows strong rate dependence. Limited by computational power, the MD simulations are performed at a pulling speed several orders faster than that in the experimental tests. Therefore, the corresponding unfolding mechanisms (e.g., sequential or simultaneous rupture of several hydrogen bonds) can be different (*96*), and a direct comparison of the force records is often challenging. It should be pointed out that our current model, trained with all-atom MD data, is not intended for designing proteins that directly meet the given pulling force response at a different pulling speed. Instead, we use the MD data under the fixed pulling speed as a consistent representation of mechanical properties of protein. While the absolute values of strength and toughness measured by MD may change with the pulling speed, the relative rank of the mechanical properties of the proteins often remain robust. At the same time, there do exist methods and procedures to bridge MD and experimental results (*97, 98*). For example, built upon the steered MD trajectory calculated here, further MD simulations can be performed to calculate the mean force potential during unfolding using statistical sampling methods. Unfolding force distribution in experimentally relevant regimes can be predicted based on the mean force potential via transition-state theory and Monte Carlo simulations (*97*). Therefore, the design goal of our current model can be connected to the response under experimental relevant pulling speed and force level with extra calculations and sampling efforts.

Second, our MD data include fundamental atomistic details and avoid bias from bottom-up coarse-grained (CG) or theoretical models with specific assumptions. By tracking atomic motions during unfolding, the MD results require little predefined assumptions like

Table 4. A short summary of the performance of protein language diffusion model developed in the present work and other models discussed.

Model name	Tested input conditions	Design accuracy	Design novelty
The developed model: protein language diffusion model using one-shot design	Mechanical unfolding responses from naturally existing proteins	Good agreement with the designed pulling force responses as well as the strength and toughness in trend and values	Tend to generate de novo ones, but can also rediscover ones that show some similarity to existing proteins
	De novo mechanical unfolding responses	Similar to the above	More probable to discover de novo sequences
AM1: Protein diffusion model using one-shot design	Mechanical unfolding responses from existing proteins	Slightly weaker than AM2	–
AM2: Protein diffusion model using multi-shot iterative design	Mechanical unfolding responses from existing proteins	Weaker than the developed model	–

those in CG models (*99*) or theoretical worm-like chain (WLC) models (*100*). At the same time, the information collected from the full-atom MD simulation can be used to fit parameters in CG and WLC models. The diffusion model was directly trained on these force patterns and learned to pick relevant features via the attention mechanisms embedded, avoiding any human intervention or preknowledge on the subject.

Third, similar models can be developed when sufficient data on mechanical unfolding force under other testing protocols become available. As an initial step to prove this framework, our current model is trained and validated with the freshly curated MD data and the consistent MD test protocol. At the same time, similar models can be straightforwardly trained with other sufficient databases. In particular, when a large number of force-separation curves measured from a standard experimental protocol become available, models with similar architectures can be developed via direct training or transfer learning with them. With such models, consistent validation will require synthesis and test of the designed proteins in the wet lab. This requires well-planned in-depth design for specific goals while the current work is focused on model development and has provided self-consistent validation. We will save the experimental studies for future study.

Finally, the specific choice of pulling force direction in our MD protocol is clarified here. It has been demonstrated that the mechanical unfolding responses of proteins can be affected by the detailed folding geometry and unfolding pathway (*101*). To effectively record the force history that corresponds to the deformation and uncoiling events of protein internal structures, we designed the test protocol to apply the mechanical force along the direction that connects the two ends of a protein chain after relaxation. When the mechanical pull is applied in another direction, the monomer is likely to first undergo a rotational motion to align along with the current pull direction before meaningful unfolding events happen given that in a quasi-static loading process rotation usually requires a smaller load than unfolding events like breaking hydrogen bonds. Once the protein monomer aligns along with the pulling direction, for the unfolding process that follows, we expect that a force record pattern similar to ours could be collected and our protocol and results remain relevant and robust. At the same time, some extreme cases could exist. For protein monomers mainly with weak internal folding structures (e.g., unstructured random coils), the load to rotate the monomer can be comparable to or larger than that of deformation and unfolding. We expect the detailed force pattern could be affected more strongly by the choice of pull direction in those cases.

The freshly curated MD dataset and the pLDM developed here offer a unique and powerful means to investigate the underlying sequence-structure-property relationship and explore the enormous protein sequence spaces with molecular mechanical properties as guidance, to meet specific mechanobiology properties. With the available dataset, one can conduct a detailed survey on the internal structural features (e.g., secondary structures) of the proteins and their correlation with the unfolding force pattern to see whether there is any uneven distribution centered on certain secondary structure patterns (e.g., α helix and β sheet) among PDB proteins and our training set for certain unfolding force patterns. Applying our model, future studies can start with a systematic study on the relationships between sequence-structure-mechanical properties in proteins.

For example, as demonstrated in Fig. 6, one can construct de novo mechanical unfolding responses by mixing those existing PDB proteins at different ratios and our model can generate protein candidates that meet those mechanical unfolding responses. With such generating capability, one can systematically classify the patterns of various unfolding responses of the existing PDB proteins as shown in Fig. 2C, construct de novo force-separation responses transferring between those different patterns, and generate the corresponding proteins, thus studying how the patterns of the unfolding force responses and their transition affect the protein sequence mutations and internal structure variations. At the same time, as the designed de novo proteins increase in number, their sequences and mechanical unfolding responses can be used as growing data to gradually increase the protein dataset on mechanical properties and our model can be further trained on this growing set. With the more powerful model, one can further study some challenging topics, such as designing proteins with optimal mechanical properties or even their combination in various engineering and biological applications.

While we have developed the pLDM that takes the mechanical unfolding responses as the design conditions here, we expect that similar pLDM frameworks can be generalized for other property-to-sequence design tasks in proteins. The enhancement brought by merging pretrained pLMs can be inspiring for other design tasks, especially where only small datasets on the property of interest are available or affordable at the beginning. At the same time, going beyond only one type of condition as the design target, our pLDM can also be generalized for design tasks under multiple objectives, given the flexibility of the diffusion model in incorporating these conditions (Fig. 3B). Combining the previous work using a pDM (*56*), one example can be taking both secondary structure and unfolding forces as the design target. Also, during the generation process, techniques like inpainting through selective masking or biasing certain amino acids (*102*) are straightforward to implement. Combining these under the pLDM framework, we envision a comprehensive generative model that moves towards designing proteins at all levels, including sequence, structure, and properties in harmony.

MATERIALS AND METHODS

Protein mechanical unfolding simulations by MD

We use Nanoscale Molecular Dynamics (NAMD) to perform full-atom MD simulations. The interaction between protein atoms is described by the CHARMM force field (*84*). We adopt a generalized Born implicit solvent model (*85*) for the effect of solvent on proteins. Compared with simulations with an explicit solvent model, our setup balances the accuracy and the computational costs. We develop a parallel workflow to simulate the mechanical unfolding process of about 7026 proteins of various sequence lengths.

Dataset

We curate the dataset based on the MD results. Key information for each protein case includes PDB ID, protein sequence, sequence length, pulling force vector, strength, and toughness. See Fig. 2 for details on their distributions. We use 85% of the dataset for training and keep 15% for testing.

Design of the neural network architectures and training

The pLDM developed here consists of a pretrained pLM and a diffusion model. Only the latter is trainable. For the pretrained pLM, we use a variant with 150 million parameters from the ESM-2 series of models

(*26*, *88*). During training, we first propagate a mini-batch of *B* tokenized sequences with a length smaller than or equal to N in terms of a B × 1 × N tensor to the last hidden layer of the pretrained pLM to compute the logits, then normalize them into a B × M × N tensor, where the M components in the second dimension represent the probability of that position being each of the M words in the pLM model. The diffusion model only operates in this word probability space introduced by the pLM. At deployment, the output of the diffusion model is translated back to the tokenized sequences by sampling the word with the highest probability.

Protein folding

We adopt OmegaFold (*23*) for rapid prediction of protein structures from the sequences. OmegaFold offers a rapid alternative as it does not require MSA, yet produces results of similar accuracy as AlphaFold2 (*20*) and trRosetta (*103*) (and similar, related state of the art methods).

Design accuracy evaluation

We use various metrics to compare the measured mechanical unfolding responses and mechanical properties with the input design conditions for individual designs as well as predictions for the whole test set.

For vectors, including the unfolding pulling force vector for one protein and toughness or strength for proteins in the test set, the R^2 and relative L_2 error are defined as the following

$$R^2[\vec{x},\vec{y}] = 1 - \frac{\sum_i (x_i - y_i)^2}{\sum_i (x_i - \bar{x})^2} \tag{4}$$

$$L_2^{rela}[\vec{x},\vec{y}] = \frac{\|\vec{x} - \vec{y}\|}{\|\vec{x}\|} = \frac{\sqrt{\sum_i (x_i - y_i)^2}}{\sqrt{\sum_i (x_i)^2}} \tag{5}$$

where $\vec{x}$ is the ground truth or input vector and $\vec{y}$ is the measured one from the predictions, x_i and y_i are their components, and $\bar{x}$ is the mean of the components x_i.

For scalars, including the toughness and strength for one protein, the relative L_1 error is defined as the following

$$L_1^{rela}[x,y] = \frac{|y - x|}{|x|} \tag{6}$$

where x is the ground truth or input value and y is the measured value based on the prediction.

BLAST analysis

The BLAST analysis (*89*) for the various cases is conducted using the blastp (protein-protein BLAST) algorithm and the nonredundant protein sequences (nr) database.

Visualization

We use Visual Molecular Dynamics (*104*) for visualization of the protein structures.

Software versions and hardware

We use Python 3.9.16, PyTorch 1.12.1 + cu13 (*105*) with CUDA (CUDA version 12.0), and an NVIDIA Tesla V100 with 32 GB VRAM for training and inference.

Supplementary Materials

This PDF file includes:

Fig. S1
Table S1
Legends for movies S1 to S8
Legends for data S1 to S4

Other Supplementary Material for this manuscript includes the following:

Movies S1 to S8
Data S1 to S4

REFERENCES AND NOTES

1. G. A. Petsko, D. Ringe, *Protein Structure and Function* (New Science Press, 2004).
2. D. López Barreiro, J. Yeo, A. Tarakanova, F. J. Martin-Martinez, M. J. Buehler, Multiscale modeling of silk and silk-based biomaterials—A review. *Macromol. Biosci.* **19**, e1800253 (2019).
3. G. Gronau, S. T. Krishnaji, M. E. Kinahan, T. Giesa, J. Y. Wong, D. L. Kaplan, M. J. Buehler, A review of combined experimental and computational procedures for assessing biopolymer structure–process–property relationships. *Biomaterials* **33**, 8240–8255 (2012).
4. C. Vepari, D. L. Kaplan, Silk as a biomaterial. *Prog. Polym. Sci.* **32**, 991–1007 (2007).
5. S. Ling, D. L. Kaplan, M. J. Buehler, Nanofibrils in nature and materials engineering. *Nat. Rev. Mater.* **3**, 18016 (2018).
6. K. A. Jansen, D. M. Donato, H. E. Balcioglu, T. Schmidt, E. H. J. Danen, G. H. Koenderink, A guide to mechanobiology: Where biology and physics meet. *Biochim. Biophys. Acta* **1853**, 3043–3052 (2015).
7. A. E. M. Beedle, S. Garcia-Manyes, The role of single-protein elasticity in mechanobiology. *Nat. Rev. Mater.* **8**, 10–24 (2023).
8. J. L. Balestrini, J. K. Skorinko, A. Hera, G. R. Gaudette, K. L. Billiar, Applying controlled non-uniform deformation for in vitro studies of cell mechanobiology. *Biomech. Model. Mechanobiol.* **9**, 329–344 (2010).
9. Z. Qin, M. J. Buehler, L. Kreplak, A multi-scale approach to understand the mechanobiology of intermediate filaments. *J. Biomech.* **43**, 15–22 (2010).
10. P.-S. Huang, S. E. Boyken, D. Baker, The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
11. U. G. K. Wegst, H. Bai, E. Saiz, A. P. Tomsia, R. O. Ritchie, Bioinspired structural materials. *Nat. Mater.* **14**, 23–36 (2015).
12. G. X. Gu, M. Takaffoli, M. J. Buehler, Hierarchically enhanced impact resistance of bioinspired composites. *Adv. Mater.* **29**, 1700060 (2017).
13. F. Barthelat, Z. Yin, M. J. Buehler, Structure and mechanics of interfaces in biological materials. *Nat. Rev. Mater.* **1**, 1–16 (2016).
14. W. Huang, A. Tarakanova, N. Dinjaski, Q. Wang, X. Xia, Y. Chen, J. Y. Wong, M. J. Buehler, D. L. Kaplan, Design of multistimuli responsive hydrogels using integrated modeling and genetically engineered silk–elastin-like proteins. *Adv. Funct. Mater.* **26**, 4113–4123 (2016).
15. S. T. Krishnaji, G. Bratzel, M. E. Kinahan, J. A. Kluge, C. Staii, J. Y. Wong, M. J. Buehler, D. L. Kaplan, Sequence–structure–property relationships of recombinant spider silk proteins: Integration of biopolymer design, processing, and modeling. *Adv. Funct. Mater.* **23**, 241–253 (2013).
16. M. J. Buehler, Generative pretrained autoregressive transformer graph neural network applied to the analysis and discovery of novel proteins. *J. Appl. Phys.* **134**, 084902 (2023).
17. A. Paladino, F. Marchetti, S. Rinaldi, G. Colombo, Protein design: From computer models to artificial intelligence. *WIREs Comput. Mol. Sci.* **7**, e1318 (2017).
18. Z. Qin, L. Wu, H. Sun, S. Huo, T. Ma, E. Lim, P. Y. Chen, B. Marelli, M. J. Buehler, Artificial intelligence method to design and fold alpha-helical structural proteins from the primary amino acid sequence. *Extreme Mech. Lett.* **36**, 100652 (2020).
19. J. Wang, H. Cao, J. Z. H. Zhang, Y. Qi, Computational protein design with deep learning neural networks. *Sci. Rep.* **8**, 6349 (2018).
20. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
21. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. Dustin Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. Van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. Christopher Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
22. M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Zídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli,

G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
23. R. Wu, F. Ding, R. Wang, R. Shen, X. Zhang, S. Luo, C. Su, Z. Wu, Q. Xie, B. Berger, J. Ma, J. Peng, High-resolution de novo structure prediction from primary sequence. bioRxiv 500999 [Preprint] (2022). https://doi.org/10.1101/2022.07.21.500999.
24. R. Chowdhury, N. Bouatta, S. Biswas, C. Rochereau, G. M. Church, P. K. Sorger, M. Alquraishi, Single-sequence protein structure prediction using language models from deep learning. bioRxiv 454840 [Preprint] (2021). https://doi.org/10.1101/2021.08.02.454840.
25. X. Fang, F. Wang, L. Liu, J. He, D. Lin, Y. Xiang, K. Zhu, X. Zhang, H. Wu, H. Li, L. Song, A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. *Nat. Mach. Intell.* **5**, 1087–1096 (2023).
26. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
27. X. Fang, F. Wang, L. Liu, J. He, D. Lin, Y. Xiang, X. Zhang, H. Wu, H. Li, L. Song, HelixFold-Single: MSA-free protein structure prediction by using protein language model as an alternative. arXiv:2207.13921 [q-bio.BM] (2022).
28. M. H. Høie, E. N. Kiehl, B. Petersen, M. Nielsen, O. Winther, H. Nielsen, J. Hallgren, P. Marcatili, NetSurfP-3.0: Accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Res.* **50**, W510–W515 (2022).
29. B. Zhang, J. Li, Q. Lü, Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC Bioinformatics* **19**, 293 (2018).
30. C. H. Yu, W. Chen, Y. H. Chiang, K. Guo, Z. Martin Moldes, D. L. Kaplan, M. J. Buehler, End-to-end deep learning model to predict and design secondary structure content of structural proteins. *ACS Biomater Sci. Eng.* **8**, 1156–1165 (2022).
31. G. Pollastri, A. McLysaght, Porter: A new, accurate server for protein secondary structure prediction. *Bioinformatics* **21**, 1719–1720 (2005).
32. C. Mirabello, G. Pollastri, Porter, PaleAle 4.0: High-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics* **29**, 2056–2058 (2013).
33. A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
34. J. Tubiana, D. Schneidman-Duhovny, H. J. Wolfson, ScanNet: An interpretable geometric deep learning model for structure-based protein binding site prediction. *Nat. Methods* **19**, 730–739 (2022).
35. F. Sverrisson, J. Feydy, B. E. Correia, M. M. Bronstein, Fast end-to-end learning on protein surfaces. bioRxiv 424589 [Preprint] (2020). https://doi.org/10.1101/2020.12.28.424589.
36. V. Thumuluri, H. M. Martiny, J. J. Almagro Armenteros, J. Salomon, H. Nielsen, A. R. Johansen, NetSolP: Predicting protein solubility in *Escherichia coli* using language models. *Bioinformatics* **38**, 941–946 (2022).
37. M. J. Buehler, MeLM, a generative pretrained language modeling framework that solves forward and inverse mechanics problems. *J. Mech. Phys. Solids* **181**, 105454 (2023).
38. E. Khare, C. Gonzalez-Obeso, D. L. Kaplan, M. J. Buehler, CollagenTransformer: End-to-end transformer model to predict thermal stability of collagen triple helices using an nlp approach. *ACS Biomater Sci. Eng.* **8**, 4301–4310 (2022).
39. Y. Hu, M. J. Buehler, End-to-end protein normal mode frequency predictions using language and graph models and application to sonification. *ACS Nano* **16**, 20656–20670 (2022).
40. K. Guo, M. J. Buehler, Rapid prediction of protein natural frequencies using graph neural networks. *Digit. Discov.* **1**, 277–285 (2022).
41. F. Y. C. Liu, B. Ni, M. J. Buehler, PRESTO: Rapid protein mechanical strength prediction with an end-to-end deep learning model. *Extreme Mech. Lett.* **55**, 101803 (2022).
42. A. J. Lew, M. J. Buehler, A deep learning augmented genetic algorithm approach to polycrystalline 2D material fracture discovery and design. *Appl. Phys. Rev.* **8**, 041414 (2021).
43. E. Khare, C. H. Yu, C. Gonzalez Obeso, M. Milazzo, D. L. Kaplan, M. J. Buehler, Discovering design principles of collagen molecular stability using a genetic algorithm, deep learning, and experimental validation. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2209524119 (2022).
44. G. Dong, G. Liao, H. Liu, G. Kuang, A review of the autoencoder and its variants: A comparative perspective from target recognition in synthetic-aperture radar images. *IEEE Geosci. Remote Sens. Mag.* **6**, 44–68 (2018).
45. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks. *Commun ACM* **63**, 139–144 (2020).
46. J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst* **33**, 6840–6851 (2020).
47. G. Marcus, E. Davis, S. Aaronson, A very preliminary analysis of DALL-E 2. arXiv:2204.13807 [cs.CV] (2022).
48. C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, M. Norouzi, Photorealistic text-to-image diffusion models with deep language understanding. arXiv:2205.11487 [cs.CV] (2022).
49. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2022-June**, 10674–10685 (2021).
50. M. Z. Makoś, N. Verma, E. C. Larson, M. Freindorf, E. Kraka, Generative adversarial networks for transition state geometry prediction. *J. Chem. Phys.* **155**, 024116 (2021).
51. T. Lebese, B. Mellado, X. Ruan, The use of Generative Adversarial Networks to characterise new physics in multi-lepton final states at the LHC. arXiv:2105.14933 [hep-ph] (2021).
52. Z. Yang, C. H. Yu, K. Guo, M. J. Buehler, End-to-end deep learning method to predict complete strain and stress tensors for complex hierarchical composite microstructures. *J. Mech. Phys. Solids* **154**, 104506 (2021).
53. Z. Yang, C. H. Yu, M. J. Buehler, Deep learning model to predict complex stress and strain fields in hierarchical composites. *Sci. Adv.* **7**, eabd7416 (2021).
54. M. J. Buehler, FieldPerceiver: Domain agnostic transformer model to predict multiscale physical fields and nonlinear material properties through neural ologs. *Mater. Today* **57**, 9–25 (2022).
55. B. Ni, H. Gao, A deep learning approach to the inverse problem of modulus identification in elasticity. *MRS Bull.* **46**, 19–25 (2021).
56. B. Ni, D. L. Kaplan, M. J. Buehler, Generative design of de novo proteins based on secondary-structure constraints using an attention-based diffusion model. *Chem* **9**, 1828–1849 (2023).
57. Z. Lin, T. Sercu, Y. LeCun, A. Rives, Deep generative models create new and diverse protein structures, in *Machine Learning in Structural Biology Workshop at the 35th Conference on Neural Information Processing Systems* (MLSB, 2021).
58. N. Anand, T. Achim, Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. arXiv:2205.15019 [q-bio.QM] (2022).
59. B. L. Trippe, J. Yim, D. Tischer, D. Baker, T. Broderick, R. Barzilay, T. Jaakkola, Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. arXiv:2206.04119 [q-bio.BM] (2022).
60. J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, D. Baker, De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
61. A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, N. Naik, Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
62. M. Mora, A. Stannard, S. Garcia-Manyes, The nanomechanics of individual proteins. *Chem. Soc. Rev.* **49**, 6816–6832 (2020).
63. J. Alegre-Cebollada, Protein nanomechanics in biological context. *Biophys. Rev.* **13**, 435–454 (2021).
64. M. J. Buehler, S. Keten, T. Ackbarow, Theoretical and computational hierarchical nanomechanics of protein materials: Deformation and fracture. *Prog. Mater. Sci.* **53**, 1101–1241 (2008).
65. H. Lu, B. Isralewitz, A. Krammer, V. Vogel, K. Schulten, Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophys. J.* **75**, 662–671 (1998).
66. K. C. Neuman, A. Nagy, Single-molecule force spectroscopy: Optical tweezers, magnetic tweezers and atomic force microscopy. *Nat. Methods* **5**, 491–505 (2008).
67. E. M. Puchner, H. E. Gaub, Force and function: Probing proteins with AFM-based force spectroscopy. *Curr. Opin. Struct. Biol.* **19**, 605–614 (2009).
68. M. L. Hughes, L. Dougan, The physics of pulling polyproteins: A review of single molecule force spectroscopy using the AFM to study protein unfolding. *Rep. Prog. Phys.* **79**, 076601 (2016).
69. X. Zhang, L. Ma, Y. Zhang, High-resolution optical tweezers for single-molecule manipulation. *Yale J. Biol. Med.* **86**, 367–383 (2013).
70. D. B. Ritchie, M. T. Woodside, Probing the structural dynamics of proteins and nucleic acids with optical tweezers. *Curr. Opin. Struct. Biol.* **34**, 43–51 (2015).
71. R. Tapia-Rojo, E. C. Eckels, J. M. Fernández, Ephemeral states in protein folding under force captured with a magnetic tweezers design. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 7873–7878 (2019).
72. X. Zhao, X. Zeng, C. Lu, J. Yan, Studying the mechanical responses of proteins using magnetic tweezers. *Nanotechnology* **28**, 414002 (2017).
73. J. Wu, P. Li, C. Dong, H. Jiang, B. Xue, X. Gao, M. Qin, W. Wang, B. Chen, Y. Cao, Rationally designed synthetic protein hydrogels with predictable mechanical properties. *Nat. Commun.* **9**, 620 (2018).
74. M. Sikora, J. I. Sulkowska, B. S. Witkowski, M. Cieplak, BSDB: The biomolecule stretching database. *Nucleic Acids Res.* **39**, D443–D450 (2011).
75. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

76. UniProt Consortium, UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
77. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. *Adv Neural Inf Process Syst* 2017-December, 5999–6009 (2017).
78. J. Ribeiro, C. Ríos-Vera, F. Melo, A. Schüller, Calculation of accurate interatomic contact surface areas for the quantitative analysis of non-bonded molecular interactions. *Bioinformatics* **35**, 3499–3501 (2019).
79. M. Hebditch, M. A. Carballo-Amador, S. Charonis, R. Curtis, J. Warwicker, Protein–Sol: A web tool for predicting protein solubility from sequence. *Bioinformatics* **33**, 3098–3100 (2017).
80. P. B. Dennis, E. L. Onderko, J. M. Slocik, L. J. Bird, D. A. Phillips, W. J. Crookes-Goodson, S. M. Glaven, Proteins for bioinspired optical and electronic materials. *MRS Bull.* **45**, 1027–1033 (2020).
81. T. Wang, D. He, H. Yao, X. Guo, B. Sun, G. Wang, Development of proteins for high-performance energy storage devices: Opportunities, challenges, and strategies. *Adv. Energy Mater.* **12**, 2202568 (2022).
82. G. Qin, P. B. Dennis, Y. Zhang, X. Hu, J. E. Bressner, Z. Sun, W. J. Crookes-Goodson, R. R. Naik, F. G. Omenetto, D. L. Kaplan, Recombinant reflectin-based optical materials. *J. Polym. Sci. B Polym. Phys.* **51**, 254–264 (2013).
83. J. Ren, Y. Wang, Y. Yao, Y. Wang, X. Fei, P. Qi, S. Lin, D. L. Kaplan, M. J. Buehler, S. Ling, Biological material interfaces as inspiration for mechanical and optical material designs. *Chem. Rev.* **119**, 12279–12336 (2019).
84. K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, A. D. Mackerell Jr., CHARMM general force field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **31**, 671–690 (2010).
85. A. V. Onufriev, D. A. Case, Generalized born implicit solvent models for biomolecules. *Annu. Rev. Biophys.* **48**, 275–296 (2019).
86. S. R. K. Ainavarapu, J. Brujić, H. H. Huang, A. P. Wiita, H. Lu, L. Li, K. A. Walther, M. Carrion-Vazquez, H. Li, J. M. Fernandez, Contour length and refolding rate of a small protein controlled by engineered disulfide bonds. *Biophys. J.* **92**, 225–233 (2007).
87. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, A. Dos, S. Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, M. Ai, Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv 500902 [Preprint] (2022). https://doi.org/10.1101/2022.07.20.500902.
88. facebookresearch/esm: Evolutionary Scale Modeling (esm): Pretrained language models for proteins. https://github.com/facebookresearch/esm.
89. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
90. T. Niwa, B. W. Ying, K. Saito, W. Jin, S. Takada, T. Ueda, H. Taguchi, Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 4201–4206 (2009).
91. A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with CLIP latents. arXiv:2204.06125 [cs.CV] (2022).
92. H. Zou, Z. M. Kim, D. Kang, A survey of diffusion models in natural language processing. arXiv:2305.14671 [cs.CL] (2023).
93. X. Lisa Li, J. Thickstun, I. Gulrajani, P. Liang, T. B. Hashimoto, Diffusion-LM improves controllable text generation. [Preprint] (2022). https://github.com/XiangLi1999/Diffusion-LM.git.
94. Z. Gao, C. Tan, S. Z. Li, DiffSDS: A language diffusion model for protein backbone inpainting under geometric conditions and constraints. arXiv:2301.09642 [q-bio.QM] (2023).
95. M. J. Buehler, S. Y. Wong, Entropic elasticity controls nanomechanics of single tropocollagen molecules. *Biophys. J.* **93**, 37–43 (2007).
96. T. Ackbarow, X. Chen, S. Keten, M. J. Buehler, Hierarchies, multiple energy barriers, and robustness govern the fracture mechanics of α-helical and β-sheet protein domains. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 16410–16415 (2007).
97. P.-C. Li, D. E. Makarov, Theoretical studies of the mechanical unfolding of the muscle protein titin: Bridging the time-scale gap between simulation and experiment. *J. Chem. Phys.* **119**, 9260–9268 (2003).
98. P. Cao, G. Yoon, W. Tao, K. Eom, H. S. Park, The role of binding site on the mechanical unfolding mechanism of ubiquitin. *Sci. Rep.* **5**, 8757 (2015).
99. S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, A. Kolinski, Coarse-grained protein models and their applications. *Chem. Rev.* **116**, 7898–7936 (2016).
100. S. Keten, M. J. Buehler, Strength limit of entropic elasticity in beta-sheet protein domains. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **78**, 061913 (2008).
101. R. A. Nome, J. M. Zhao, W. D. Hoff, N. F. Scherer, Axis-dependent anisotropy in protein unfolding from integrated nonequilibrium single-molecule experiments, analysis, and simulation. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 20799–20804 (2007).
102. J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
103. Z. Du, H. Su, W. Wang, L. Ye, H. Wei, Z. Peng, I. Anishchenko, D. Baker, J. Yang, The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.* **16**, 5634–5651 (2021).
104. W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
105. A. Paszke, S. Gross, F. Massa, A. Lerer, J. B. Google, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. K. Xamla, E. Yang, Z. Devito, M. R. Nabla, A. Tejani, S. Chilamkurthy, Q. Ai, B. Steiner, L. F. Facebook, J. B. Facebook, S. Chintala, PyTorch: An imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst* **32**, (2019).

Acknowledgments

Funding: We acknowledge support from USDA (2021-69012-35978), DOE-SERDP (WP22-S1-3475), ARO (79058LSCSB, W911NF-22-2-0213, and W911NF2120130), and the MIT-IBM Watson AI Lab and MIT's Generative AI Initiative. Additional support from NIH (U01EB014976 and R01AR077793) and ONR (N00014-19-1-2375 and N00014-20-1-2189) is acknowledged.
Author contributions: Conceptualization: M.J.B. and D.L.K. Investigation: B.N. and M.J.B. Methodology: B.N., M.J.B., and D.L.K. Resources: M.J.B. and D.L.K. Funding acquisition: M.J.B. and D.L.K. Data curation: B.N. and M.J.B. Validation: M.J.B. and B.N. Supervision: M.J.B. and D.L.K. Formal analysis: B.N. and M.J.B. Software: B.N. and M.J.B. Project administration: M.J.B. and D.L.K. Visualization: B.N., M.J.B., and D.L.K. Writing—original draft: B.N., M.J.B., and D.L.K. Writing—review and editing: B.N., M.J.B., and D.L.K. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 16 October 2023
Accepted 8 January 2024
Published 7 February 2024
10.1126/sciadv.adl4000