nature portfolio

Peer Review File

An end-to-end framework for the prediction of protein structure and fitness from single sequence

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This manuscript presents "structural prediction based on inter-residue relative displacement" (SPIRED), an algorithm for predicting protein structure from sequence without a multiple sequence alignment (MSA), as well as two extensions to the model for predicting protein fitness and stability: SPIRED-Fitness and SPIRED-Stab. At a high level, the structure prediction model bears some resemblance to ESMFold and includes the ESM protein language model representation of the input sequence as an input feature. However, the core innovation SPIRED makes is in the structure prediction objective function. Whereas prior methods like ESMFold train using what is called frame-aligned point error (FAPE) loss introduced by AlphaFold2, which requires rotating and aligning the true and predicted 3D protein structures to calculate the loss, SPIRED instead introduces the relative displacement (RD) loss. The RD loss is based on relative distances of the 3D coordinates and avoids the computationally expensive alignments. This is a major factor in making SPIRED considerably more efficient than ESMFold and comparable models. SPIRED-Fitness and SPIRED-Stab extend the SPIRED model and ESM sequence representation with a type of graph neural network that predicts either protein fitness or stability, respectively. The efficient training of SPIRED enables the SPIRED parameters to be fine tuned when training SPIRED-Fitness.

The manuscript makes many noteworthy contributions. The RD loss function achieves the goals of speeding up model training, and SPIRED is also much faster for inferring new protein structures, especially for longer protein sequences. SPIRED is one of the only models to integrate structure, fitness, and stability prediction (both stability and thermostability) into a single model and capitalizes on the multiple tasks by using the structure prediction task to initialize fitness prediction and the fitness prediction task to initialize stability prediction. With respect to protein fitness, SPIRED-Fitness is one of the only methods that can predict protein fitness scores for new proteins in a zero shot manner, that is, without observing sequence-function training data for the new protein. Other models make zero shot predictions about protein evolutionary conservation scores or stability that may be correlated with fitness or function, but they do not actually predict functional scores. Finally, the datasets and code presented follow best practices overall with models and

results archived on Zenodo.

Major comments:

1) It is fair to focus most of the structure prediction evaluation on other methods that do not use MSAs. However, showing readers the state of the art in structure prediction, even if that does require an MSA and structure templates, will show the tradeoffs between sequence-only and MSA-assisted prediction. Adding the top structure prediction methods from CASP15 (e.g. Oda 2023 doi:10.1002/prot.26551) or the standard AlphaFold2 would establish this tradeoff.

2) SPIRED-Fitness only supports making fitness predictions for single and double mutations. This is a major limitation. In machine learning directed protein engineering, the goal is to use a model to introduce many mutations far from the initial wild type sequence to optimize the protein function. See Biswas 2021 doi:10.1038/s41592-021-01100-y, Bryant 2021 doi:10.1038/s41587-020-00793-4, Fahlberg 2023 doi:10.1101/2023.11.08.566287, Gelman 2024 doi:10.1101/2024.03.15.585128, Thomas 2024 doi:10.1101/2024.03.21.585615 for arbitrary recent examples.

3) The proteolysis datasets used for training and evaluating SPIRED-Fitness far outnumber the MaveDB and DeepSequence datasets. These proteolysis datasets also represent only a narrow type of protein fitness, essentially an approximation of stability, rather than the diverse types of functions captured in the MaveDB and DeepSequence datasets. Therefore, it is important to separate the fitness prediction performance by proteolysis and non-proteolysis data, potentially as a supplementary analysis, to see whether the correlations in Table 3 hold for all protein functions or primarily only the proteins in the proteolysis dataset.

4) The protein fitness evaluations do not reflect how fitness prediction models are used or best practices for benchmarking these models. As seen in the ProteinGym benchmarking paper (Notin 2023 https://papers.nips.cc/paper_files/paper/2023/hash/cac723e5ff29f65e3fcbb0739ae91bee-Abstract-Datasets_and_Benchmarks.html) there are two main settings when predicting

protein function or fitness, and SPIRED-Fitness is applicable in both. The first is the zero shot setting where there is no labeled function data for the target protein. Here, protein language models, MSA-based models, and others included in ProteinGym are strong baselines. Critically, in contrast to what the manuscript claims many of these models can be run on multiple mutations (ProteinGym does it). GVP-MSA (Chen 2023 doi:10.1016/j.cels.2023.07.003) is one of the only other models that supports zero shot function prediction and therefore is important to cite and possibly benchmark. The second setting is a supervised setting where there is sequence-function data for the target protein. In this setting, ProteinGym again provides baselines. The recent ProteinNPT (Notin 2023 doi:2023.12.06.570473) appears to have a claim for being state of the art. Augmented linear regression (Hsu 2022 doi:10.1038/s41587-021-01146-5) is a simple and very powerful baseline and is worthwhile to include even if ProteinNPT is not run. Finally, ProteinGym includes the NDCG and recall metrics for protein fitness evaluation because these prioritize methods that can identify the most beneficial sequence variants, which is what is needed for protein engineering.

5) The SPIRED-Stab evaluations include a large number of sequence-based and structure-based models. However, the selection of these models is not well-justified and omits some of the most recent methods like ThermoMPNN (Dieckhaus 2024 doi:10.1073/pnas.2314853121), RaSP (Blaabjerg 2023 doi:10.7554/eLife.82593), PROSTATA (Umerenkov 2022 doi:10.1101/2022.12.25.521875), and Pythia (Sun 2023 doi:10.1101/2023.08.09.552725). Recent related work should be discussed and cited and potentially compared against.

Minor comments:

6) The poor performance of some of the baseline fitness prediction methods in Figure S9 like MSA Transformer and ESM-2 is surprising given their much stronger performance in Protein Gym. Is this related to the dominance of the proteolysis datasets or something else?

7) The text on page 13 starting with "Clearly, ESMFold shows better performances..." states that ESMFold outperforms SPIRED because it has more model parameters and trains on AlphaFold2 predictions. However, there are multiple differences between ESMFold and

SPIRED, and the manuscript does not contain specific ablation studies to isolate these specific differences as the cause of the performance gap between the models. It should add specific support for these claims or remove them.

8) Many results are shown with only a mean (or median, it is not clear) without any measure of variance. For example, the Figure 2 bar graphs, Table 3, and Table 4.

9) Several model design choices seem arbitrary. They are explained well, but the process used to arrive at that design is not intuitive because there does not appear to be a formal process of evaluating multiple alternative designs with a validation dataset. For example, there are multiple training stages and there is weight sharing in the first 4 modules folding modules but not the last 2.

10) Prior methods from these authors like GeoFitness are used as baselines and important for understanding this manuscript. However, those methods are not introduced here, and the differences between the prior models and the new models is not explained. Briefly explaining the prior methods would make this manuscript self-contained.

11) Many terms, methods, and acronyms are used in the figures and text before they are fully introduced, which makes the manuscript challenging to read in order. Examples include global and local coordinate spaces, GDFold2, FAPE loss, Instance/Row/Column Normalization, ESM-1v logits, and GeoDTm.

Typos:
Pg 10 "was was"
Pg 16 "binnized"
Fig 5 "SPTRED"
Pg 18 "descripted"
Supplement title "authos"
Supplement pg 44 "paremeter"

Reviewer #1 (Remarks on code availability):

Overall the software follows computational best practices. The Code Ocean capsule was runnable and appeared to reproduce the stored results. The pretrained SPIRED-Fitness model from Zenodo could be run following the instructions in the GitHub repository. The fitness web server provided a convenient and accessible way for less computationally-savvy users to take advantage of these models with appealing graphics and files available to download for local analysis. However, there are ways the software can be further improved:

1) The web server should use secure https.

2) When tested with the wild type avGFP sequence (SKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLSYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK), a protein family on which SPIRED performs well according to the manuscript, the double mutation predictions were poor. Comparing the top 1,624 double mutation fitness effects from SPIRED-Fitness model with the 12,777 experimental functional effects from deep mutational scanning (Sarkisyan 2016 doi:10.1038/nature17995) that included two mutations showed no overlap. The top five experimental double mutations V161A,S173R; I126T,K156E; K154E,Q182L; Y37H,N133S; Y37N,N103I were not in the top 1,624 SPIRED-Fitness predictions. None of the top 1,624 experimental double mutations intersected with the top 1,624 SPIRED-Fitness predictions. Even though not all avGFP double mutations were tested experimentally, the lack of overlap is concerning and shows how the Spearman correlation metrics in the manuscript may not be meaningful for practical protein engineering tasks.

3) The GitHub repository does not appear to support training the model from scratch or fine tuning a SPIRED model for fitness or stability prediction. The manuscript presents the efficient training of SPIRED as one of its main advantages, but that strength is negated if the code does not support it.

4) The inference scripts only support fitness or stability prediction. A script that supports standalone structure prediction from a sequence would also be useful.

5) It was unclear in the GitHub readme that "install GDFold2" means follow all of the instructions in the GDFold2 repository, including setting up another conda environment.

6) The versions of the required Python packages should be specified, as they are in the Code

Ocean Dockerfile. Ideally the conda environment could be created with pinned package versions from a single .yml file.

7) https://github.com/Gonglab-THU/SPIRED-Fitness/blob/aa1f08317ccb48c74316a4cead298c8ccb50f3d2/run_spired_fitness.sh#L67 assumes cuda:6 as the device, which is not the typical device for most users.

8) The code uses esm2_t36_3B_UR50D and esm1v_t33_650M_UR90S models, but the manuscript only discusses the 650M parameter ESM model. How is the 3B parameter model used?

9) The Code Ocean metadata.yml file may have extra line breaks in the name and description strings.

10 The Code Ocean Dockerfile would be generally valuable outside of Code Ocean as well.

Reviewer #2 (Remarks to the Author):

This paper presents an integrative approach to protein structure as well as missense mutation fitness, ΔΔG, and ΔTm prediction. The method, termed SPIRED, is referred to as an "end-to-end framework", because it takes a single protein sequence as input and carries out the entire process of predicting both structure and its fitness without needing additional information. The paper is compelling for several reasons: 1) The authors demonstrate an improved speed and reduced computational cost of training, which is essential for others to use and build on the method. 2) The authors demonstrate an increased speed of inference, which is tremendous advantage for end-users running the method. 3) A novel loss function called "relative displacement loss" is introduced, which is more efficient than AlphaFold's original FAPE loss. And 4) the authors made available the code (including pseudocode) and offer a server for users to run their predictions, which seems to be working well.

However, there are potentially two main issues with the manuscript; one concerning the validation, which I think should be more comprehensive (see comments below). The other concerns the fact that, despite the authors highlighting the utility of structure and fitness prediction for protein engineering, the reality is that a considerable fraction of the community will use it for clinical interpretation of pathogenic missense variants, as it happened to similar tools such as DeepSequence or ESM-1v. Therefore, the authors should

consider introducing a validation analysis to demonstrate performance related to this topic (for example, ClinVar vs gnomAD), and ideally provide pre-computed scores for single amino acid variants of human proteins.

Comments

1. It is unclear from the text whether the CAMEO test set was filtered to below a sequence identity threshold (e.g., <30% or TM < 0.5) (Figure 2a)? This is necessary to ensure that the comparison is not confounded by homology. The authors should include a list of identifiers in the supplement for both CAMEO and CASP15 proteins/domains (e.g., gene name, UniProt ID, or Pfam ID).

2. The same principle should apply to SCOPe and CATH validation sets (Figure 3 and Figures S3/S7). For example, the SCOPe database has the ASTRAL sub-dataset, which contains domain sequences with less than 40% pairwise identity (SCOPe 2.08: Structural Classification of Proteins — extended (berkeley.edu)) – if the authors used this dataset, it should be stated.

3. In the "Training and test sets for protein fitness prediction" section of the Methods, the authors have claim to have used 51 proteins from MaveDB and 22 proteins from the DeepSequence dataset – is there any overlap between the two datasets? The authors should include in the supplement a list of what the MaveDB proteins were and which DMS score values were used in their analysis.

4. In the section "Protein fitness prediction by SPIRED-Fitness", including Table3, the results should be discussed and shown separately for the 3 sources of functional data (cDNA proteolysis dataset of Tsuboyama et al., MaveDB, and DeepSequence Dataset – if the latter two are different). This is necessary because the datasets are different both in terms of size and in the type and quality of the functional assay readout.

5. Additionally, single and double mutations in Table 3 should be shown separately. This is necessary because in practice single mutations are a lot more common and are therefore clinically more important.

6. Table 3 would also benefit from adding the language model-based (ESM-2, ESM-1b, and ESM-1v) and empirical (ECNet, RFjoint,MSA Transformer, but also DeepSequence – if the authors can run it) methods to the comparison, subset by test data source. Although these do not necessarily output values for double mutation, single mutations are more important and comparing SPIRED-Fitness (Stage 2) to these more broadly used methods will give a better idea about the performance than the current aggregated results.

Server: The SPIRED-Fitness server (http://structpred.life.tsinghua.edu.cn/server_spired_fitness.html) predicts everything, including double mutants, for an input sequence but SPIRED-Stab (http://structpred.life.tsinghua.edu.cn/server_spired_stab.html) only predicts one single amino acid substitution at a time. Given that SPIRED is an “end-to-end” framework, could the authors add an option to predict ΔΔG and ΔTm of all missense substitutions for an input sequence?

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks on code availability):

Authors included version-controlled code along with pseudocode.
Installation is relatively simple, I was able to run their test examples.
Server works as intended.

Reviewer #4 (Remarks to the Author):

Authors present "SPIRED-Fitness", an innovative framework for predicting protein structure and fitness from a single sequence. This model, termed SPIRED (Structural Prediction based on Inter-Residue Relative Displacement), reduces computational costs significantly compared to state-of-the-art methods like ESMFold and OmegaFold, boasting about a 5-fold

acceleration in inference speed and a substantial reduction in training resource consumption. SPIRED-Fitness, combining SPIRED with downstream neural networks, offers rapid and accurate predictions for protein structure and mutational effects on protein fitness and stability, achieving state-of-the-art performance in these tasks. One of the notable achievements of SPIRED is its ability to predict protein structures quickly and with comparable accuracy to existing methods, which is particularly useful for high-throughput analyses. The use of a novel loss function, the Relative Displacement Loss, contributes to this efficiency by avoiding computationally expensive operations required by other methods. However, I have some concerns.

（1） For results shown in Figure 2, why does the accuracy decrease after refinement by GDFold2? Is it because GDFold2 was not specifically optimized for the results from SPIRED? If so, I suggest that the parameters of GDFold2 could be specifically optimized for the results from SPIRED to enhance its performance.

(2) You said “Without recycling (Cycle = 1), SPIRED performs well on the CAMEO set (average TM-score = 0.786), slightly surpassing OmegaFold (average TM-score = 0.784), as shown in Figure 2a.” However , according to the results from Figure 2, the average TM-score of OmegaFold is 0.778. Please carefully review the manuscript to avoid such errors.

（3） On the section evaluation of structure prediction at the level of protein folds, you use SPIRED, ESMFold and OmegaFold to predict the structures for 34,021 SCOPe domains belonging to a total of 1,231 folds.

（4） Why is the prediction accuracy of the proposed method slightly lower than ESM2 and OmegaFold on the CAMEO test proteins, yet higher on the SCOPe domains? Please provide a detailed analysis and discussion. If possible, analyze which types of proteins SPIRED performs well on and which it does poorly on. Given the multitude of protein structure prediction methods available today, it would be beneficial to tell users in which scenarios SPIRED is recommended for use.

（5） I recommend that the authors make some comparisons with EMBER3D, as it is also renowned for its fast inference capabilities. Additionally, analyze and compare some of the differences between them.

（6） The authors should conduct a more detailed comparison and analysis of their proposed method with other state-of-the-art protein structure prediction techniques. In addition to numerical comparisons between different methods, a detailed analysis and discussion of the reasons behind the differences in results are equally important.

（7） I would like to understand the differences in performance between your method and AlphaFold2, in both its MSA-using version and its non-MSA version. I recognize that the underlying principles of your method and AlphaFold2 are entirely different. However, displaying such results would be helpful in revealing the gap between the latest single-sequence-based structure prediction methods and AlphaFold2. I've noticed that your test protein set is quite large; if this presents a challenge in the short term, opting for a subset of them would also be acceptable.

Letter to reviewers

SPIRED-Fitness: an end-to-end framework for the prediction of protein structure and fitness from single sequence

1 Responses to Reviewer #1

1.1 Major Comments

1. It is fair to focus most of the structure prediction evaluation on other methods that do not use MSAs. However, showing readers the state of the art in structure prediction, even if that does require an MSA and structure templates, will show the tradeoffs between sequence-only and MSA-assisted prediction. Adding the top structure prediction methods from CASP15 (e.g. Oda 2023 doi:10.1002/prot.26551) or the standard AlphaFold2 would establish this tradeoff.

Answer:
Thank you for this suggestion. Indeed, almost all top structure prediction methods in CASP15 still used the AlphaFold2 model but focused on MSA searching and manipulation. Considering this fact, we took the standard AlphaFold2 as control to show the general performance of nowadays state-of-the-art single-sequence-based protein structure prediction methods. As shown in the updated **Figure 2** and **Figure S2** of the revised manuscript, the single-sequence-based structure predictors mainly tested in this work, including SPIRED, OmegaFold and ESMFold, are still less powerful than the MSA-based AlphaFold2 prediction but outperform the AlphaFold2 version that takes the single sequence as input.

2. The original SPIRED-Fitness only supports making fitness predictions for single and double mutations. This is a major limitation. In machine learning directed protein engineering, the goal is to use a model to introduce many mutations far from the initial wild type sequence to optimize the protein function. See Biswas 2021 doi:10.1038/s41592-021-01100-y, Bryant 2021 doi:10.1038/s41587-020-00793-4, Fahlberg 2023 doi:10.1101/2023.11.08.566287, Gelman 2024 doi:10.1101/2024.03.15.585128, Thomas 2024 doi:10.1101/2024.03.21.585615 for arbitrary recent examples.

Answer:
SPIRED-Fitness only supports prediction on single and double mutations just because single and double mutational data could be simply applied to the nodes and edges of the graph-based model as loss functions, which naturally allows the end-to-end training. Our current method could be slightly modified to allow the inference or prediction of multiple mutations.

First, in the zero-shot prediction mode of SPIRED-Fitness, we proposed an iterative approach to enable the inference of multiple mutational effects. As shown in the new **Figure S9**, after the first running of SPIRED-Fitness, we could instantly derive the predicted effects of all possible single mutations, among which we could pick up the site that has the largest magnitude in fitness change. Subsequently, the sequence carrying this single mutation will be fed into SPIRED-Fitness and we could instantly obtain the effect of the second mutation following exactly the same principle. The sequence carrying the first two mutation sites could be fed to SPIRED-Fitness for the inference of further-on mutations, a process that could proceed repeatedly to infer the effect of an arbitrary number of mutations, and summation of scores in all iterations will produce the prediction result for the target mutant. Considering the fast running speed of SPIRED-Fitness, this iterative approach is still time efficient. We have added a new **subsection 2.1.4** entitled "Zero-shot setting" in the Supplementary Materials to concretely describe this inference process.

Next, we modified the network architecture of SPIRED-Fitness to allow supervised learning of arbitrary mutational effects. As shown in the new **Figure S10**, resembling the work of Notin *et al.*, this architecture uses the one hot

encoding (OHE) to extract the zero-shot prediction results and then concatenate with SPIRED embeddings (after convolution operation and average pooling) for predictions. Notably, this model supports supervised training over known data of mutant sequence carrying an arbitrary number of mutations as well as the corresponding fitness change, and the trained model can also provide prediction for any proposed variant without the prior limitation on single and double mutations. We have added a new **subsection 2.4.2** entitled "Supervised setting" in the Supplementary Materials to concretely describe this modification on the network architecture.

The above mentioned modifications of SPIRED-Fitness could be validated on the zero-shot setting and supervised setting of ProteinGym, respectively. Following your point #4, we benchmarked our SPIRED-Fitness method on ProteinGym in these two settings. For both settings, we tested not only the single mutational effects but also double, triple and quadruple mutational effects. Benchmark testing results indicate that our method could provide acceptable prediction for the multiple mutational effects. Please refer to our response to your point #4 for details.

3. The proteolysis datasets used for training and evaluating SPIRED-Fitness far outnumber the MaveDB and DeepSequence datasets. These proteolysis datasets also represent only a narrow type of protein fitness, essentially an approximation of stability, rather than the diverse types of functions captured in the MaveDB and DeepSequence datasets. Therefore, it is important to separate the fitness prediction performance by proteolysis and non-proteolysis data, potentially as a supplementary analysis, to see whether the correlations in Table 3 hold for all protein functions or primarily only the proteins in the proteolysis dataset.

Answer:
This is a nice suggestion. We also noticed the imbalance between the proteolytic and non-proteolytic (*i.e.* MaveDB/DeepSequence) types of data. Following your suggestion, we evaluated the mean values and standard deviations of the Spearman correlation coefficients among tested proteins on the proteolytic and non-proteolytic data, respectively, for the methods listed in Table 3. As shown in the new **Table S2**, the correlation within all mutants (*i.e.* single + double) shows comparable values for SPIRED-Fitness, GeoFitness v2 and ECNet. Considering that the ECNet models optimized using data from individual proteins confront no data imbalance problem, this observation indicates that our model training process is not significantly perturbed by the imbalance between proteolytic and non-proteolytic types of data. Furthermore, we finetuned the SPIRED-Fitness model purely using non-proteolytic data and found only negligible changes in the final performance. This experiment further corroborates that the model training for SPIRED-Fitness in this work is not biased by the overwhelming proteolytic data.

Notably, the apparently lower correlations within the double mutants in SPIRED-Fitness and GeoFitness v2 are likely caused by the objective function used for model training, which enforces the general soft Spearman correlation within all mutants (*i.e.* single + double) rather than focusing on the ranking within individual single mutation or double mutation categories, because ranking among all available mutants are of higher importance in practical protein engineering.

In the revised manuscript, we have added a paragraph in the subsection entitled "Protein fitness prediction by SPIRED-Fitness" of the Results section to describe the above extra experiments.

4. The protein fitness evaluations do not reflect how fitness prediction models are used or best practices for benchmarking these models. As seen in the ProteinGym benchmarking paper (Notin 2023 https://papers.nips.cc/paper_files/paper/2023/hash/cac723e5ff29f65e3fcbb0739ae91bee-Abstract-Datasets_and_Benchmarks.html**) there are two main settings when predicting protein function or fitness, and SPIRED-Fitness is applicable in both. The first is the zero shot setting where there is no labeled function data for**

the target protein. Here, protein language models, MSA-based models, and others included in ProteinGym are strong baselines. Critically, in contrast to what the manuscript claims many of these models can be run on multiple mutations (ProteinGym does it). GVP-MSA (Chen 2023 doi:10.1016/j.cels.2023.07.003) is one of the only other models that supports zero shot function prediction and therefore is important to cite and possibly benchmark. The second setting is a supervised setting where there is sequence-function data for the target protein. In this setting, ProteinGym again provides baselines. The recent ProteinNPT (Notin 2023 doi:2023.12.06.570473) appears to have a claim for being state of the art. Augmented linear regression (Hsu 2022 doi:10.1038/s41587-021-01146-5) is a simple and very powerful baseline and is worthwhile to include even if ProteinNPT is not run. Finally, ProteinGym includes the NDCG and recall metrics for protein fitness evaluation because these prioritize methods that can identify the most beneficial sequence variants, which is what is needed for protein engineering.

Answer:

This is a nice suggestion. Following your suggestion, we benchmarked SPIRED-Fitness in both the zero-shot setting and the supervised setting on ProteinGym. As shown in our response to your major point #2, we modified the inference protocol and/or the network architecture to allow the prediction of effects caused by not only single and double mutations but also triple and higher-order ones for the two ProteinGym settings.

In the zero-shot setting, after removing proteins redundant with the training set of the original SPIRED-Fitness model, we only retained 50 assays for performance evaluation, where the proteins have sequence identity < 50% with the training set of the original SPIRED-Fitness model and length < 900 residues (see Table S11). Following your suggestion, we evaluated SPIRED-Fitness against GVP-MSA and strong ProteinGym baselines. Specifically, we clustered these ProteinGym assays into 10 groups based on protein sequence similarity (inter-group sequence identity < 50%), and then iteratively finetuned the original SPIRED-Fitness (Stage 2) model on 9 groups of assays for the inference of assays in the remaining 1 group. In the new **Table 4** and **Table S6**, we evaluated SPIRED-Fitness against ProteinGym baselines as well as GVP-MSA on the metrics including the Spearman correlation coefficient, NDCG, top 10% recall, AUC and MCC, as you suggested. All methods are classified into three categories: MSA-based, structure-based and single-sequence-based. SPIRED-Fitness definitely belongs to the last category, since it relies on the amino acid sequence rather than MSA or structure information for prediction. Based on the evaluation results, SPIRED-Fitness is among the best predictors within the single-sequence-based methods, in terms of the Spearman correlation coefficient for both single mutation and multiple mutation data.

In the supervised setting, we appended the zero-shot SPIRED-Fitness models by 2 additional layers (see Figure S10), whose parameters were optimized by supervised training. We evaluated SPIRED-Fitness against the two state-of-the-art methods as you suggested, ProteinNPT and Augmented linear regression, as well as other ProteinGym baselines. Again, all methods are classified into two categories: MSA-based and single-sequence-based. Clearly, SPIRED-Fitness belongs to the second category since it does not use any MSA information for prediction. Here, the comparison was only conducted on a subset of testing assays, considering the enormous computational costs for supervised training and 5-fold cross validation in three different data splitting strategies (*Random*, *Modulo* and *Contiguous*) on every individual assay. Specifically, we clustered the 50 zero-shot testing assays into 20 groups based on protein sequence similarity and chose 2 representatives from each group to compose the subset of testing assays (see the items highlighted in bold in Table S11). Based on the evaluation results in the new **Table S7**, where the *Random* mode counts all single + multiple mutations, SPIRED-Fitness attains satisfactory prediction results, ranking the best among the single-sequence-based methods. Particularly, SPIRED-Fitness outperforms Augmented linear regression, albeit markedly behind ProteinNPT due to the lack of using MSA information.

We have added a new subsection entitled "Benchmark of SPIRED-Fitness on ProteinGym" in the Results section of the revised manuscript to concretely describe these extra experiments and the evaluation results. In addition, we have added the definitions of new metrics in the "Evaluation metrics" subsection of the Methods section of the revised manuscript.

5. The SPIRED-Stab evaluations include a large number of sequence-based and structure-based models. However, the selection of these models is not well-justified and omits some of the most recent methods like ThermoMPNN (Dieckhaus 2024 doi:10.1073/pnas.2314853121), RaSP (Blaabjerg 2023 doi:10.7554/eLife.82593), PROSTATA (Umerenkov 2022 doi:10.1101/2022.12.25.521875), and Pythia (Sun 2023 doi:10.1101/2023.08.09.552725). Recent related work should be discussed and cited and potentially compared against.

Answer:
Following your suggestion, we have evaluated the performance of ThermoMPNN, RaSP, PROSTATA and Pythia on the S669/S461 test sets for the prediction of $\Delta\Delta G$ upon mutations. Results of these methods have been replenished into **Table 5**, **Table S8** and **Table S9**. In brief, the performance of ThermoMPNN, RaSP, PROSTATA and Pythia slightly exceeds traditional predictors like PremPS and ACDC-NN, consistent with the results reported by their original papers. However, our main conclusion holds: our SPIRED-Stab outperforms these newly proposed methods as well as traditional methods.

We have also slightly modified the words in the subsection entitled "Prediction of the mutational effects on protein stability by SPIRED-Stab" in the Results section of the revised manuscript.

1.2 Minor Comments

1. The poor performance of some of the baseline fitness prediction methods in Figure S9 like MSA Transformer and ESM-2 is surprising given their much stronger performance in Protein Gym. Is this related to the dominance of the proteolysis datasets or something else?

Answer:
In this work, MSA Transformer and ESM-2 are tested as purely zero-shot predictors, which means that the raw output by these models given the wild-type sequence are directly used for the fitness inference. In a new **Table S4** in the Supplementary Materials of the revised manuscript, we have provided extra results of models tested in our current work, including MSA Transformer and ESM-2, on the proteolytic and non-proteolytic data as well as on all data in combination. From the results, you could find balanced behaviors of these types of models. Hence, the poor performance is not related with the dominance of the proteolysis dataset.

2. The text on page 13 starting with "Clearly, ESMFold shows better performances..." states that ESMFold outperforms SPIRED because it has more model parameters and trains on AlphaFold2 predictions. However, there are multiple differences between ESMFold and SPIRED, and the manuscript does not contain specific ablation studies to isolate these specific differences as the cause of the performance gap between the models. It should add specific support for these claims or remove them.

Answer:
According to the ablation study evaluated on CAMEO targets in the original paper of ESMFold (Figure S3 of the ESMFold paper), removal of AlphaFold2-predicted structures from the training set deteriorates the performance of ESMFold by ~3 percentage points in terms of the mean lDDT, while reducing the number of Evoformer blocks further weakens the prediction power. Moreover, according to Table S1 of the same paper, the performance of ESM-2 in inter-residue contact prediction is monotonically related with the amount of model parameters. Hence, the above observations support that both the parameter amount and the use of AlphaFold-predicted structures are essential for maintaining the high prediction accuracy in ESMFold.

In the revised manuscript, we have slightly modified the corresponding words as:

"*Clearly, ESMFold shows better performances than SPIRED and OmegaFold on both CAMEO and CASP15 sets. This is, however, not unexpected considering that the model parameters of ESMFold outnumber those of SPIRED and OmegaFold by approximately five times, and that ESMFold engages a large amount of AlphaFold2-predicted protein structures for model training (see Table 2), both factors reported as essential for achieving the high prediction accuracy in the ESMFold paper.*"

3. Many results are shown with only a mean (or median, it is not clear) without any measure of variance. For example, the Figure 2 bar graphs, Table 3, and Table 4.

Answer:
In the revised manuscript, we have replaced the bar plots in **Figure 2**, **Figure S2**, **Figure 3** and **Figure S6** by boxplots that contain distribution information and have added standard deviations to evaluation metrics in all tables in the main text (*e.g.*, **Tables 3-6**) as well as the corresponding tables in the Supplementary Materials (*e.g.*, **Tables S1-S2**, **Table S4**, **Tables S6-S8**). Thank you.

4. Several model design choices seem arbitrary. They are explained well, but the process used to arrive at that design is not intuitive because there does not appear to be a formal process of evaluating multiple alternative designs with a validation dataset. For example, there are multiple training stages and there is weight sharing in the first 4 modules folding modules but not the last 2.

Answer:
Actually, the current network architecture was gradually formed during the full course of our method development, which cost at least two years. Alternative designs were generally tested based on different subsets of protein structures during the early stage of our method development. Because the full training process of the final version of SPIRED model is very time consuming (about 85 GPU days), it is unlikely to perform thorough ablation studies on the various aspects of network architecture design.

However, we could provide intuitive explanations for most design choices. As for the weight sharing in Folding Units4 as you proposed, the first 4 modules are constrained by the type 1 RD loss (see Algorithm 11), while the last 2 modules are constrained by the type 2 RD loss (see Algorithm 12). Based on this consideration, we enforced weight sharing within the two groups of modules, respectively. Notably, the two types of RD loss functions are designed for slightly different purposes. The type 1 RD loss focuses on constraining the overall topology by the initial prediction, whereas the type 2 RD loss is more appropriate for constraining the later-on corrections on the coordinates. Hence, these two types of RD loss functions are also engaged over the first four and last two structure modules, respectively.

As for the multiple training stages, it was specifically designed for enhancing the training efficiency. For example, in the first stage, we only engaged a small subset of high-quality proteins (sample number = 22K), in order to allow the model to quickly learn the sequence-structure mapping. Next, in the second stage, more proteins are gradually enriched into the training set (sample number = 63K). In the third stage, we included the all protein structures (sample number = 138K). In the last stage, we further increase the cropping size from 256 to 350 and eventually to 420. Despite the complexity, the full training process of our SPIRED model is highly reproducible.

We have slightly modified the words in the Methods section of the revised manuscript to clarify the above points.

5. Prior methods from these authors like GeoFitness are used as baselines and important for understanding this manuscript. However, those methods are not introduced here, and the differences between the prior models and the new models is not explained. Briefly explaining the prior methods would make this manuscript self-contained.

Answer:
This is a good suggestion. Thank you. We have added a detailed description of GeoStab-suite (GeoFitness, GeoDDG and GeoDTm) as well as the difference between the older version and the newer version introduced in this work in the Supplementary Materials. Please see the new **subsection 2.6.4** entitled "Brief introduction and training details of GeoFitness v2 and GeoStab v2".

6. Many terms, methods, and acronyms are used in the figures and text before they are fully introduced, which makes the manuscript challenging to read in order. Examples include global and local coordinate spaces, GDFold2, FAPE loss, Instance/Row/Column Normalization, ESM-1v logits, and GeoDTm.

Answer:
Thanks for your advice. We are sorry for the confusion.

The global coordinate system refers to the laboratory coordinate system in which a protein structure is determined experimentally (*i.e.* coordinate system of the PDB file), whereas the local coordinate system of a residue means the Cartesian coordinate system located on a residue by placing the C_α atom at the origin, the C atom on the x-axis and the N atom within the xy-plane. Here, our definition of the local coordinate system is identical to that in the AlphaFold2 paper. ESM-1v logits refer to the logits before the Softmax operation in the last output layer of the ESM-1v model. We have added annotation to these terms in the Methods section of the revised manuscript (see the "Network architecture of SPIRED-Fitness and SPIRED-Stab" subsection).

As for GeoDDG/GeoDTm, they are two models for $\Delta\Delta G/\Delta T_m$ prediction in our prior work of GeoStab. We have added a new **subsection 2.6.4** entitled "Brief introduction and training details of GeoFitness v2 and GeoStab v2" in the Supplementary Materials to provide a thorough explanation.

GDFold2 is an in-house protein folding environment, developed specifically for generating all-atom coordinates based on the prediction results of deep learning models like SPIRED. The paper has been published on bioRxiv with DOI of https://doi.org/10.1101/2024.03.13.584741. FAPE loss and Instance/Row/Column Normalization are two specific techniques. They are hard to explain in the manuscript. Therefore, we have added citations of the corresponding papers for GDFold2, FAPE loss and Instance/Row/Column Normalization in the revised manuscript.

7. Typos: Pg 10 "was was", Pg 16 "binnized", Fig 5 "SPTRED", Pg 18 "descripted", Supplement title "authos", Supplement pg 44 "paremeter"

Answer:
These typo issues have been fixed. Thank you.

1.3 Remarks on code availability

1. The web server should use secure https.

Answer:
Our server is hosted on the Tsinghua University campus. According to Tsinghua University's regulations, using secure HTTPS on campus servers requires a separate application process, which is quite complicated. Therefore, we are currently maintaining HTTP access only. We plan to upgrade our server to use secure HTTPS in the future. Thank you for your suggestion.

2. When tested with the wild type avGFP sequence (SKGEELFTGVVPILVELDGDVNGHKFSVSGEGEG-DATYGKLTLKFICTTGKLPVPWPTLVTTLSYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDD GNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNI EDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFV-TAAGITHGMDELYK), a protein family on which SPIRED performs well according to the manuscript, the double mutation predictions were poor. Comparing the top 1,624 double mutation fitness effects from SPIRED-Fitness model with the 12,777 experimental functional effects from deep mutational scanning (Sarkisyan 2016 doi:10.1038/nature17995) that included two mutations showed no overlap. The top five experimental double mutations V161A,S173R; I126T,K156E; K154E,Q182L; Y37H,N133S; Y37N,N103I were not in the top 1,624 SPIRED-Fitness predictions. None of the top 1,624 experimental double mutations intersected with the top 1,624 SPIRED-Fitness predictions. Even though not all avGFP double mutations were tested experimentally, the lack of overlap is concerning and shows how the Spearman correlation metrics in the manuscript may not be meaningful for practical protein engineering tasks.

Answer:
In our manuscript, we only stated that SPIRED performs well on the structure prediction of GFP. We also evaluated the data set you mentioned (Sarkisyan 2016 doi:10.1038/nature1799), which contains a large number of double mutations. We have to admit that the zero-shot prediction by SPIRED-Fitness, which corresponds to the mode of our server version, is indeed poor. The Spearman correlation coefficients for single, double and single + double mutations are 0.51, 0.40 and 0.42, respectively. Our reevaluation is consistent with your testing in general, suggesting that the Spearman correlation coefficient may still be an appropriate metric. The reason of poor performance is that single mutation data outnumber double ones substantially in our training set, which hinders our model in effectively learning the double mutational effects.

Despite the relatively poor prediction power of our method on this protein in the zero-shot mode, SPIRED-Fitness could provide more accurate predictions on avGFP in the supervised mode. Considering the abundant data in the avGFP assay, we slightly modified the original SPIRED-Fitness model in the similar network architecture to SPIRED-Stab (see Figure 1e), where the wild-type and mutant sequences are fed into pre-trained SPIRED-Fitness

module with shared weights and their embeddings are processed by MLP layers for the prediction of mutational effects caused by an arbitrary number of mutations. After supervised training on the avGFP data, this modified SPIRED-Fitness model achieves the Spearman correlation coefficient of 0.80, 0.75 and 0.75 for single, double and single + double mutations, respectively. Hence, at least for the avGFP case you mentioned, supervised training over the ample data guarantees rather accurate prediction results.

3. The GitHub repository does not appear to support training the model from scratch or fine tuning a SPIRED model for fitness or stability prediction. The manuscript presents the efficient training of SPIRED as one of its main advantages, but that strength is negated if the code does not support it.

Answer:
This is a good point. Following your suggestion, we have provided the full training scripts for SPIRED, SPIRED-Fitness, and SPIRED-Stab on the GitHub repository.

4. The inference scripts only support fitness or stability prediction. A script that supports standalone structure prediction from a sequence would also be useful.

Answer:
SPIRED-Fitness model can provide both the structure predicted by SPIRED and the fitness landscape. In case that there is also a need of predicted protein structure without fitness results, we have provided a standalone SPIRED model on the GitHub repository. Please note that SPIRED and SPIRED-Fitness provide the coordinates of the backbone C_α atoms, while GDFold2 is required to generate the coordinates of all atoms. Detailed instructions on how to use SPIRED and GDFold2 to predict protein structures are provided on GitHub.

5. It was unclear in the GitHub readme that "install GDFold2" means follow all of the instructions in the GDFold2 repository, including setting up another conda environment.

Answer:
We have made a clarification on how to install and use GDFold2 to generate the coordinates of all atoms based on the C_α coordinates predicted by SPIRED. The instruction is available on GitHub. Thank you.

6. The versions of the required Python packages should be specified, as they are in the Code Ocean Dockerfile. Ideally the conda environment could be created with pinned package versions from a single .yml file.

Answer:
The spired_fitness.yml file has been provided on GitHub. The versions of the required Python packages have also been specified. Thank you for your advice.

7. https://github.com/Gonglab-THU/SPIRED-Fitness/blob/aa1f08317ccb48c74316a4cead298c8ccb50f3d2/run_spired_fitness.sh**#L67 assumes cuda:6 as the device, which is not the typical device for most users.**

Answer:

It's a typo issue and we have fixed it. Thank you.

8. The code uses esm2_t36_3B_UR50D and esm1v_t33_650M_UR90S models, but the manuscript only discusses the 650M parameter ESM model. How is the 3B parameter model used?

Answer:
650M parameter ESM-1v model is used to generate logits for SPIRED-Fitness prediction in the output layer. The detailed description can be found in Algorithm 16 (line 10). 650M parameter ESM-2 model is used to generate input embedding for SPIRED-Fitness and SPIRED-Stab. The detailed description can be found in Algorithm 17 (line 1). 3B parameter ESM-2 model is used to generate input embedding for SPIRED. The detailed description can be found in Algorithm 1 (line 1).

We are sorry for the confusion and have added an annotation to **Algorithm 1** to emphasize the use of ESM-2 3B model there in the revised manuscript.

9. The Code Ocean metadata.yml file may have extra line breaks in the name and description strings.

Answer:
We have fixed this issue. Thank you.

10. The Code Ocean Dockerfile would be generally valuable outside of Code Ocean as well.

Answer:
We have provided the dockerfile in our GitHub repository. Thank you.

2 Responses to Reviewer #2

2.1 General description

However, there are potentially two main issues with the manuscript; one concerning the validation, which I think should be more comprehensive (see comments below). The other concerns the fact that, despite the authors highlighting the utility of structure and fitness prediction for protein engineering, the reality is that a considerable fraction of the community will use it for clinical interpretation of pathogenic missense variants, as it happened to similar tools such as DeepSequence or ESM-1v. Therefore, the authors should consider introducing a validation analysis to demonstrate performance related to this topic (for example, ClinVar vs gnomAD), and ideally provide pre-computed scores for single amino acid variants of human proteins.

Answer: This is a good suggestion. However, unlike the real-valued protein fitness data on which SPIRED-Fitness was originally trained, the clinical data are categorical, typically falling in binary classification. The different characteristics of data need specially designed loss function for model optimization. For instance, the cross entropy is clearly more suitable for learning the clinical data than the soft Spearman ranking loss engaged in the current model training. Consequently, the current SPIRED-Fitness model is not expected to achieve high performance on clinical data, where the soft-Spearman-based pre-training is supposed to introduce negative effects. Here, I would show you some trial

validation on the clinical data available on the ProteinGym official website (https://proteingym.org/). In the following table (**Table R1**), we list the ROC_AUC (*i.e.* area under the ROC curve) and PR_AUC (*i.e.* area under the Precision-Recall curve) values for the current SPIRED-Fitness model as well as a few strong ProteinGym baselines. Specifically, we selected proteins with length $\leq$ 1,000 residues for testing, resulting in 1,113 proteins for zero-shot pathogenic prediction. According to the results, SPIRED-Fitness clearly does not exhibit strong performance in this prediction task, especially when compared with the protein language model ESM-1b.

Table R1. Evaluation on the clinical data from ProteinGym

Model	ROC_AUC	PR_AUC
TranceptEVE_L	0.92	0.95
GEMME	0.92	0.94
EVE	0.92	0.94
ESM-1b	0.90	0.93
MutPred	0.89	0.93
SIFT4G	0.89	0.92
SIFT	0.89	0.91
Provean	0.89	0.92
MutationAssessor	0.88	0.92
LIST-S2	0.85	0.90
SPIRED-Fitness	0.83	0.89
PrimateAI	0.83	0.88
LRT	0.79	0.83

I fully understand that the prediction of clinical effects of single mutations are of high importance. However, in order to achieve this object, we have to modify the loss function, the output layer as well as the training protocol for further performance improvement. Obviously, this task needs tremendous investigation and is not supposed to be finished within the three months assigned by the editor. We will definitely explore in this direction, but that will belong to our future work. Nevertheless, thank you for this suggestion.

2.2 Comments

1. It is unclear from the text whether the CAMEO test set was filtered to below a sequence identity threshold (e.g., <30% or TM < 0.5) (Figure 2a)? This is necessary to ensure that the comparison is not confounded by homology. The authors should include a list of identifiers in the supplement for both CAMEO and CASP15 proteins/domains (e.g., gene name, UniProt ID, or Pfam ID).

Answer:
CAMEO and CASP are both double-blind competitions in the field of protein structure prediction. During the competition, the organizers collect newly solved protein structures that have not been deposited into the PDB database and distribute their sequences to attendants for prediction. As a convention, participants could use any known protein structures determined before the starting time of the double-blind competition for their model training. In this work, we used the CAMEO and CASP15 test sets to evaluate the performance of our model. Considering that all proteins included in our training set are solved before the corresponding starting time points of CAMEO and CASP15 sets, the performance evaluation is fair and objective.

I do not understand the meaning of "filtering" you mentioned in this question, particularly on whether the pairwise identity threshold refers to comparing with proteins in the training set or self-comparison within the test set. If you meant the former, the comparison is unnecessary, since a large proportion of proteins in the test sets have sort of homology with the PDB database. After the filtering by threshold like sequence identity < 30% or TM-score < 0.5, most of the proteins would be eliminated from the testing set. Indeed, using the whole double-blind testing set for performance evaluation is conventional in this field. You could refer to the papers of AlphaFold2, RoseTTAFold2, ESMFold and OmegaFold. In all of these works, the authors evaluated their methods either on the CASP sets or on the CAMEO sets.

Nevertheless, to address your concern, we conducted a trial experiment to filter out test targets in the CAMEO set by the criterion of sequence identity < 30%. As shown in the following table (**Table R2**), only 212 of the 680 CAMEO targets are retained after filtering against the training sets of SPIRED and OmegaFold. Evaluation on these remaining proteins still suggests the comparable performance of SPIRED and OmegaFold. Comparison with ESMFold is impossible, since the training set of ESMFold contains over 10 million AlphaFold2-predicted structures, filtering against which will eliminate nearly all testing targets.

Table R2. Evaluation on CAMEO samples with sequence identity < 30% to the training set (Cycle = 1)

Model	Sample_Number	TM-score	RMSD
SPIRED	212	0.671	8.10
OmegaFold	212	0.673	9.29

If you meant the latter possibility, concerning that the evaluation within the test set may be biased by a proportion of highly similar proteins, I believe that you could find a solution in our subsequent evaluation on the SCOPe and CATH fold/topology types. When we investigated this problem, we also suspected whether the small number of CAMEO and CASP proteins could fully represent the space of natural proteins. The answer is clearly negative, according to our evaluation results. You are suggested to refer to our response to your point #2 for this topic.

Finally, following your request, we have provided the full lists of proteins in the CAMEO set and in the CASP15 set at the end of the Supplementary Materials (see the new **Table S12** and **Table S13**). Following the convention in the field of protein structure prediction, CAMEO proteins are shown by their PDB IDs and chain names, while CASP protein domains are shown by their CASP target IDs and domain names.

2. The same principle should apply to SCOPe and CATH validation sets (Figure 3 and Figures S3/S7). For example, the SCOPe database has the ASTRAL sub-dataset, which contains domain sequences with less than 40% pairwise identity (SCOPe 2.08: Structural Classification of Proteins — extended (berkeley.edu)) – if the authors used this dataset, it should be stated.

Answer:
In this work, we engaged the SCOPe and CATH datasets to simply assess the performance differences of the state-of-the-art single-sequence-based structure prediction methods on different fold types. To be honest, all proteins in the SCOPe or CATH databases have already been included in the training sets of state-of-the-art structure prediction methods like SPIRED, OmegaFold and ESMFold, which utilize the full PDB database for model training. In principle, proteins in the training set should have successful inference. Our evaluation is, however, quite surprising, since a non-negligible proportion of protein folds could not be reliably predicted by nowadays state-of-the-art

single-sequence-based structure prediction methods, although SPIRED exhibits a more balanced performance among the fold types than OmegaFold and ESMFold.

Let me take the SCOPe analysis as an example. We actually used the SCOPe v2.08 S95 database, within which all proteins have pairwise sequence identity < 95%. The name and version of database have been clearly stated in the **caption of Figure 3**. Following your suggestion, we also mentioned this information in the main text of the Results section of the revised manuscript. In this test, the threshold for pairwise sequence identity, 95% as we used or 40% as you proposed, is unimportant, because all proteins have been grouped into fold types based on the backbone topology in the SCOPe database. In our evaluation, for a specific fold type, we computed the mean TM-score, averaged over all proteins belonging to this fold, as the evaluation indicator. Therefore, each point in Figure 3a,b refers to a fold type and its evaluation indicators by different methods are compared there. Similarly, the probability density function and the final average TM-score are also evaluated upon all fold types in Figure 3c,d. Hence, evaluation results on the SCOPe database are not confounded by homology.

As for the evaluation on CATH topologies, we used the CATH v4.2 S35 database, which means that the proteins have pairwise sequence identity < 35%. The name and version of the CATH database have been clearly stated in the **caption of Figure S6** as well as the text following this figure in the Supplementary Materials.

3. In the "Training and test sets for protein fitness prediction" section of the Methods, the authors have claim to have used 51 proteins from MaveDB and 22 proteins from the DeepSequence dataset – is there any overlap between the two datasets? The authors should include in the supplement a list of what the MaveDB proteins were and which DMS score values were used in their analysis.

Answer:
This is a good question. Thank you. In fact, there were overlaps between the original MaveDB database and DeepSequence datasets. As a solution, we filtered out the redundant portion from the DeepSequence datasets and finally retained 22 protein assays. Hence, there is no longer any overlap within the combined MaveDB/DeepSequence datasets used in this work.

We have slightly modified description of the DeepSequence dataset in the Methods section of the revised manuscript as:

"***DeepSequence Dataset*** *collects fitness data of mutated proteins from DMS experiments. After filtering out data that are redundant with the MaveDB database, we finally retained 22 proteins from this dataset for the subsequent fitness training and testing purposes.*"

The full list of the overall 73 proteins (51 in MaveDB + 22 in DeepSequence) has been provided in our prior work of GeoFitness. Following your suggestion, we have also appended this table (see the new **Table S10**) at the end of the Supplementary Materials.

4. In the section "Protein fitness prediction by SPIRED-Fitness", including Table3, the results should be discussed and shown separately for the 3 sources of functional data (cDNA proteolysis dataset of Tsuboyama et al., MaveDB, and DeepSequence Dataset – if the latter two are different). This is necessary because the datasets are different both in terms of size and in the type and quality of the functional assay readout.

Answer:
This is a nice suggestion and it is coincident with reviewer #1's major point #3.

We agree with you that the cDNA proteolysis dataset and MaveDB/DeepSequence datasets have highly distinct characteristics. Therefore, we analyzed the performance of methods in Table 3 in the two types of data, proteolytic and non-proteolytic, and summarized the results in a new Table S2. As shown in the Methods section, the original MaveDB database and DeepSequence datasets are quite similar to each other and have overlapping parts, so there is no need to compare them separately.

Based on results in **Table S2**, our SPIRED-Fitness model exhibits a generally similar behavior to ECNet in the Spearman correlation within all mutants for the proteolytic and non-proteolytic (*i.e.* MaveDB/DeepSequence) types of data. Considering that ECNet uses the data of each individual target protein to train a specific model for inference, it is insensitive to the characteristics of data from different sources. Thus, the similar performance of SPIRED-Fitness and ECNet in proteolytic and non-proteolytic data suggest that the training process of our SPIRED-Fitness model is not perturbed significantly by nature and size of training data of the two different sources. The detailed discussion on model training and data imbalance could be found in our response to reviewer #1's major point #3.

Besides, we have inserted a paragraph in the subsection entitled "Protein fitness prediction by SPIRED-Fitness" of the Results section in the revised manuscript to describe and discuss these results.

5. Additionally, single and double mutations in Table 3 should be shown separately. This is necessary because in practice single mutations are a lot more common and are therefore clinically more important.

Answer:
Following our response to your preceding point, in the new **Table S2**, we have listed the model performance for single and double mutations separately. Generally, SPIRED-Fitness achieves comparable performance to ECNet in terms of single mutations and single+double mutations. SPIRED-Fitness and GeoFitness v2 show lower correlation within the double mutations. This apparent disadvantage, however, results from the fact that the objective function (*i.e.* soft Spearman ranking loss) is enforced to optimize the ranking within all mutants rather than the double mutants alone in the training process, since the former is more important in practical protein engineering. Besides, as you proposed, single mutations are a lot more common (see the new Table S3). Therefore, results separately displayed in the two data types support the good prediction performance of SPIRED-Fitness.

6. Table 3 would also benefit from adding the language model-based (ESM-2, ESM-1b, and ESM-1v) and empirical (ECNet, RFjoint, MSA Transformer, but also DeepSequence – if the authors can run it) methods to the comparison, subset by test data source. Although these do not necessarily output values for double mutation, single mutations are more important and comparing SPIRED-Fitness (Stage 2) to these more broadly used methods will give a better idea about the performance than the current aggregated results.

Answer:
This is a good suggestion. In the manuscript, considering that some methods like RF_{joint} are unable to provide predictions for multi-mutation effects, we retrained SPIRED-Fitness using the single mutations alone and compared with GeoFitness v2, MSA Transformer, RF_{joint}, ESM-2, ESM-1v, ESM-1b, ECNet and SESNet for single mutants in **Figure S8**. Following your suggestion, we have replenished the comparison with DeepSequence in this figure. Moreover, we have summarized these results in a new **Table S4** in the Supplementary Materials. Notably, the results

have been presented as the mean values and standard deviations of the Spearman correlation coefficients calculated over all tested proteins as well as over the proteolytic and non-proteolytic types of data, respectively, as you suggested.

7. Server: The SPIRED-Fitness server (http://structpred.life.tsinghua.edu.cn/server_spired_fitness.html) predicts everything, including double mutants, for an input sequence but SPIRED-Stab (http://structpred.life.tsinghua.edu.cn/server_spired_stab.html) only predicts one single amino acid substitution at a time. Given that SPIRED is an "end-to-end" framework, could the authors add an option to predict $\Delta\Delta G$ and ΔT_m of all missense substitutions for an input sequence?

Answer:
This is a good suggestion. As a response, we have modified the pipeline to allow users to infer the multi-mutational effects on protein stability through the SPIRED-Stab online server. You are welcome to have a test.

3 Responses to Reviewer #3

3.1 Remarks on code availability

1. Authors included version-controlled code along with pseudocode. Installation is relatively simple, I was able to run their test examples. Server works as intended.

Answer:
Thank you for the testing. It's good to know that our model can be easily used from the user side.

4 Responses to Reviewer #4

4.1 Comments

1. For results shown in Figure 2, why does the accuracy decrease after refinement by GDFold2? Is it because GDFold2 was not specifically optimized for the results from SPIRED? If so, I suggest that the parameters of GDFold2 could be specifically optimized for the results from SPIRED to enhance its performance.

Answer:
GDFold2 is a self-developed framework for folding proteins parallelly based on the geometric constraints predicted by deep learning models like SPIRED and trRosetta/RoseTTAFold. The paper of GDFold2 has been published on the bioRxiv with DOI of https://doi.org/10.1101/2024.03.13.584741. In the current GDFold2, one structural optimization mode, namely the "model 2", is specifically designed for folding proteins (*i.e.* generating all-atom coordinates) based on the multiple sets of C_α coordinates predicted by SPIRED. You could find implementation details and folding examples in the paper.

To answer your question, the average C_α TM-score usually will decrease slightly after the GDFold2 optimization, just because GDFold2 tries to establish all-atom coordinates that agree with both the geometric constraints predicted by SPIRED and the statistical constraints summarized from the high-resolution protein structures in the PDB database. Despite the high accuracy of SPIRED in general, the C_α coordinates output by SPIRED might slightly deviate from

the statistical rules, *e.g.*, the statistical distributions for the distance between two adjacent C_α atoms, scalar angles between three consecutive C_α atoms and torsion angles between four consecutive C_α atoms. GDFold2 aims for providing a satisfactory solution for this inconsistency. Despite the slight drop of C_α TM-score or rise of global RMSD, after GDFold2 structural generation, the local structural quality is dramatically improved, including the distribution of backbone torsion angles, hydrogen bond networks, secondary structure contents, side-chain rotamers, *etc*. For instance, in the following table (**Table R3**), you could find that lDDT, the metric that reflects the local structural quality, is uniformly improved after the GDFold2 coordinate optimization for both CAMEO and CASP15 targets.

Table R3. Comparison of global and local structural quality between SPIRED and SPIRED+GDFold2

Model	Test Set	TM-score	lDDT	RMSD
SPIRED(Cycle=1)	CAMEO	0.786	0.753	5.80
SPIRED(Cycle=1)+GDFold2	CAMEO	0.784	0.756	5.84
SPIRED(Cycle=4)	CAMEO	0.787	0.757	5.89
SPIRED(Cycle=4)+GDFold2	CAMEO	0.785	0.759	5.93
SPIRED(Cycle=1)	CASP15	0.610	0.656	12.09
SPIRED(Cycle=1)+GDFold2	CASP15	0.608	0.668	12.17
SPIRED(Cycle=4)	CASP15	0.614	0.675	12.62
SPIRED(Cycle=4)+GDFold2	CASP15	0.612	0.684	12.67

2. You said "Without recycling (Cycle = 1), SPIRED performs well on the CAMEO set (average TM-score = 0.786), slightly surpassing OmegaFold (average TM-score = 0.784), as shown in Figure 2a." However , according to the results from Figure 2, the average TM-score of OmegaFold is 0.778. Please carefully review the manuscript to avoid such errors.

Answer:
Thank you for the kind remind. We have fixed this typo issue.

3. On the section evaluation of structure prediction at the level of protein folds, you use SPIRED, ESMFold and OmegaFold to predict the structures for 34,021 SCOPe domains belonging to a total of 1,231 folds. Why is the prediction accuracy of the proposed method slightly lower than ESM2 and OmegaFold on the CAMEO test proteins, yet higher on the SCOPe domains? Please provide a detailed analysis and discussion. If possible, analyze which types of proteins SPIRED performs well on and which it does poorly on. Given the multitude of protein structure prediction methods available today, it would be beneficial to tell users in which scenarios SPIRED is recommended for use.

Answer:
As a double blind competition, CAMEO tries to collect newly determined protein structures and then distributes them to the attendants for prediction. During the past years, statistics over the SCOPe database suggest that the number of protein folds have nearly reached a convergent number, meaning that almost all natural protein folds have been solved experimentally at least once. Indeed, the newly determined proteins only comprise a very small proportion in the topological space of all natural proteins. For instance, searching by Foldseek identifies 578 out of the 680 CAMEO targets as intersecting with the overall 1,231 SCOPe fold types (see the following table, **Table R4**). However, these 578 targets only cover 253 SCOPe fold types (coverage ratio = 20.6%), and more importantly, some specific fold

types are over-represented by CAMEO targets. Hence, even for these 578 targets, the apparent advantage of ESMFold over SPIRED and OmegaFold by the raw averaging almost diminishes when the averaging is performed on the fold level instead.

Table R4. Evaluation on CAMEO samples with similar Folds in SCOPe

Averaging strategy	Number	SPIRED	OmegaFold	ESMFold
Raw average	578	0.823	0.815	0.875
Fold-level average	253	0.878	0.866	0.880

Given the above results, it is unfair to evaluate the performance of protein structure prediction methods solely using the raw average score of the CAMEO or CASP targets. Consequently, in this work, we tried to analyze the prediction power of existing methods at the level of protein folds, which to our best knowledge has not been done before. Although all proteins in the SCOPe database have been included in the training set of nowadays state-of-the-art protein structural prediction methods, ESMFold and OmegaFold, two popular methods that are generally believed as highly accurate predictors, are unable to provide good prediction for a number of protein folds, a phenomenon that also surprised us.

We have provided an analysis and speculation at the end of the "Supplementary Results for Structure Prediction" section in the Supplementary Materials (see **subsection 1.6** entitled "Detailed comparison and discussion on single-sequence-based structure prediction methods"). Here, I would just give a brief introduction. ESMFold enriches more than 10 million AlphaFold2-predicted structures into its training set, which would disrupt the protein distribution in the topological space of natural proteins. OmagaFold throws away a number of proteins difficult for loss function optimization, which also leads to the same effect. In contrast, SPIRED simply trains on all natural proteins without artificial perturbation. This may be the reason of its balanced performance among protein folds.

Besides, following your suggestion, we have provided lists of protein folds at which SPIRED, OmegaFold and ESMFold have poor prediction powers (*i.e.* mean TM-score < 0.5) at the end of the Supplementary Materials (see the new **Table S14**, **Table S15** and **Table S16**), respectively, so that users could choose the appropriate predictor upon their practical needs.

4. I recommend that the authors make some comparisons with EMBER3D, as it is also renowned for its fast inference capabilities. Additionally, analyze and compare some of the differences between them.

Answer:

Following your suggestion, we evaluated the performance of EMBER3D (UR), the unrelaxed version, on the CAMEO and CASP15 test sets. Despite the ultrafast inference speed (even faster than SPIRED), the performance of EMBER3D is poor, in comparison with state-of-the-art single-sequence-based methods like SPIRED, OmegaFold and ESMFold. For example, EMBER3D achieves an average TM-score of **0.539** in the CAMEO test set, markedly lower than the values of SPIRED, OmegaFold and ESMFold, with the gap of at least 0.2. The similar trend is also observed on the CASP15 set, where EMBER3D exhibits an inferiority of at least 0.15 in the average TM-score.

According to the EMBER3D paper, it relies on Rosetta to further relax the protein structure generated by the unrelaxed version. However, the improvement by Rosetta is usually marginal, unlikely to fill the gap of at least 0.15 in the average TM-score. Moreover, the Rosetta relaxation is computationally prohibitive, particularly for large proteins. Considering

the large number of proteins in our testing sets, among which some proteins have more than 1,000 residues (see Table S12), we can not afford testing the relaxed version of EMBER3D.

Nevertheless, in the revised manuscript, we have included the performance of EMBER3D (US) in **Figure 2**, together with standard AlphaFold2 versions taking MSA and single sequence as input, respectively, to comprehensively show the general performance of nowadays state-of-the-art single-sequence-based protein structure prediction methods.

5. The authors should conduct a more detailed comparison and analysis of their proposed method with other state-of-the-art protein structure prediction techniques. In addition to numerical comparisons between different methods, a detailed analysis and discussion of the reasons behind the differences in results are equally important.

Answer:
Although I agree with your point, analyzing the reasons behind the performance differences between various protein structure prediction methods far exceeds our capability, since rigorous exploration needs a series of ablation studies, in which every model should be trained with distinct designs. Remember that every state-of-the-art model costs heavy computational resources for training. For instance, it takes around 85 GPU days (three months on a single GPU like in my laboratory) for a single full training process even for SPIRED, a model design that can significantly reduce the training consumption. The computational costs for OmegaFold and ESMFold are at least 10 times larger.

Due to this difficulty, we can only provide our speculation. As shown in our response to your point #3, we have added a new **subsection 1.6** entitled "Detailed comparison and discussion on single-sequence-based structure prediction methods" in the Supplementary Materials to roughly discuss this topic. Detailed analysis over the difference between methods is also shown in the new **Table S14**, **Table S15** and **Table S16**. Nevertheless, systematic investigation on this topic still awaits diversified explorations by many independent research groups including ours in the future.

6. I would like to understand the differences in performance between your method and AlphaFold2, in both its MSA-using version and its non-MSA version. I recognize that the underlying principles of your method and AlphaFold2 are entirely different. However, displaying such results would be helpful in revealing the gap between the latest single-sequence-based structure prediction methods and AlphaFold2. I've noticed that your test protein set is quite large; if this presents a challenge in the short term, opting for a subset of them would also be acceptable.

Answer:
Thank you for your suggestion. We have updated **Figure 2** and **Figure S2**, by including the results of the standard AlphaFold2 predictions with MSA and single sequence taken as inputs, respectively. Please also refer to our response to the major point #1 of reviewer #1. In brief, the results suggest that the single-sequence-based structure predictors mainly tested in this work, including SPIRED, OmegaFold and ESMFold, are still less powerful than the MSA-based AlphaFold2 prediction but outperform the AlphaFold2 version that takes the single sequence as input.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The revised manuscript contains extensive new benchmarking analyses that thoroughly addressed my major comments. The new evaluations with AlphaFold2, ProteinGym, GVP-MSA, ThermoMPNN, RaSP, PROSTATA and Pythia combined with the existing evaluations present a comprehensive overview of SPIRED's performance in the context of other modern protein structure, fitness, and stability prediction tools. In nearly all cases, SPIRED's performance is quite strong. Overall, the results support that SPIRED is a valuable addition to the protein modeling landscape through its fast single-sequence-based structure prediction and downstream fitness and stability prediction.

The remaining comments primarily concern how the new results are presented.

Minor comments:

1) In the summary of Table 3, the text "SPIRED Fitness slightly outperforms both ECNet and GeoFitness v2, in terms of the average Spearman correlation coefficient" is a slight overstatement. Now that the standard deviation has been added, these methods are effectively tied.

2) The discussion of Table S3 in the main text gives a misleading overview of double mutations and their ranking for practical protein engineering. There are fewer double mutations in these combined datasets because many of the individual datasets are screens that only test single mutations. For screens that do include double mutations, the number of double or higher-order mutations vastly outnumbers the single mutations. Predicting the effects of and ranking double mutations is also extremely important for protein engineering because a useful model will help extrapolate from single mutations to higher order mutations. This extrapolation setting is specifically studied in other machine learning papers (e.g. ECNet Figure 4).

3) In the discussion of Table 4, the text states "SPIRED-Fitness outperforms the other single-sequence-based methods" but the margin between it and the other best models is small for Spearman correlation and they are tied or effectively tied on metrics that are relevant for protein fitness: NDCG and recall.

4) Likewise, the main text description of Table S7 is factually correct but does not acknowledge that the MSA-based methods have better performance. It is also unclear why Table S7 includes the average column because the random, modulo, and contiguous evaluation settings are not directly comparable.
5) The response letter and revised manuscript still do not explain why the other methods in Figure S8 perform so poorly. Most of these same methods are compared to SPIRED in Table 4 in the ProteinGym comparison, where the overall correlations are similar. The datasets used in Figure S8 are mostly all in ProteinGym as well, though the exact overlap is not reported. What factors drive the large performance differences, both the absolute correlations and relative performance between SPIRED and the other strong baselines?
6) The text states the models have similar performance on proteolytic and non-proteolytic data based on Table S2, but the actual gap is substantial for all models.
7) Figure 3d is still a bar plot. Only the caption was changed.
8) The definition of the Pearson correlation coefficient correlation was incorrectly changed to use the standard error instead of the standard deviation.
9) Section 2.5.1 in the supplement is blank.

Concerns of Reviewer #4:
I was also asked to comment on how the revised manuscript addressed the concerns of the original Reviewer #4. Overall, those concerns have also been satisfied. The revised manuscript now includes benchmarking against EMBER3D, AlphaFold2 with an MSA, and AlphaFold2 without an MSA in the main structure prediction evaluation (Figure 2). There is additional emphasis on the evaluation by protein fold, including new supplementary tables showing which folds lead to poor performance for SPIRED, OmegaFold, and ESMFold. The supplement also includes new discussion of the differences between these models and their training strategies that could be driving some of the fold-specific performance differences.

A minor comment is that the response letter explanation regarding the effects of GDFold2 is quite good, but these are still not obvious in the manuscript itself. When discussing the results of Figure 2, briefly summarizing the tradeoffs between C_alpha coordinates and TM-score versus overall structure quality that are presented in the response would provide readers who lack this background important context.

Reviewer #1 (Remarks on code availability):

I reviewed the updates to the SPIRED-Fitness GitHub repository, and they addressed all of the comments raised in my initial review. The manuscript now has a new Zenodo repository https://zenodo.org/doi/10.5281/zenodo.12560925, but the manuscript, web server, and GitHub readme still reference the old Zenodo URL.

Reviewer #3 (Remarks to the Author):

The authors have thoroughly addressed my and the other reviewers' questions. Major changes include results from software that were not included in the previous manuscript but overall make the revised manuscript stronger; better presentation of the results section, with supplemental tables showing the single and double mutation datasets separately; improvements to the server that will benefit the user; and an overall improved language in the manuscript that makes this work more accessible to the broad readership of the journal.

Reviewer #3 (Remarks on code availability):

SPIRED-Fitness was successfully installed and run in the first round of the review. The server runs without issues.

Letter to reviewers

An end-to-end framework for the prediction of protein structure and fitness from single sequence

1 Responses to Reviewer #1

1.1 Remarks to the Author

1. The revised manuscript contains extensive new benchmarking analyses that thoroughly addressed my major comments. The new evaluations with AlphaFold2, ProteinGym, GVP-MSA, ThermoMPNN, RaSP, PROSTATA and Pythia combined with the existing evaluations present a comprehensive overview of SPIRED's performance in the context of other modern protein structure, fitness, and stability prediction tools. In nearly all cases, SPIRED's performance is quite strong. Overall, the results support that SPIRED is a valuable addition to the protein modeling landscape through its fast single-sequence-based structure prediction and downstream fitness and stability prediction.

Answer:
Thank you.

1.2 Minor comments

1. In the summary of Table 3, the text "SPIRED Fitness slightly outperforms both ECNet and GeoFitness v2, in terms of the average Spearman correlation coefficient" is a slight overstatement. Now that the standard deviation has been added, these methods are effectively tied.

Answer:
I agree with your point. In the revised manuscript, we have changed the corresponding statement to:

"As shown in Table 1, when tested on all single and double mutational data, SPIRED-Fitness exhibits a comparable performance to both ECNet and GeoFitness v2 in terms of the average Spearman correlation coefficient between predicted and experimental values (0.85 vs. 0.84 and 0.83), implying that the precision of SPIRED structure prediction can fulfill the requirement of fitness prediction."

Notably, the original Table 3 has been renumbered as Table 1 in the revised manuscript. Thank you.

2. The discussion of Table S3 in the main text gives a misleading overview of double mutations and their ranking for practical protein engineering. There are fewer double mutations in these combined datasets because many of the individual datasets are screens that only test single mutations. For screens that do include double mutations, the number of double or higher-order mutations vastly outnumbers the single mutations. Predicting the effects of and ranking double mutations is also extremely important for protein engineering because a useful model will help extrapolate from single mutations to higher order mutations. This extrapolation setting is specifically studied in other machine learning papers (e.g. ECNet Figure 4).

Answer:
This is a good point. Thank you. In the revised manuscript, we have changed the statement to:

"The apparently weaker Spearman correlation of SPIRED-Fitness within the double mutants is suspected to arise from the design of our training loss function, which is proposed to constrain on the ranking of all mutants within each individual protein without distinguishing the single and double ones. Considering the imbalanced distribution

of single and double mutational data among DMS assays (Supplementary Table S4), further model finetuning over assays with sufficient amount of double mutational data may correct this deficit."

Notably, the original Table S3 has been renumbered as Supplementary Table S4 in the revised manuscript.

3. In the discussion of Table 4, the text states "SPIRED-Fitness outperforms the other single-sequence-based methods" but the margin between it and the other best models is small for Spearman correlation and they are tied or effectively tied on metrics that are relevant for protein fitness: NDCG and recall.

Answer:
Although I agree with your point here, I have to emphasize that the seemingly marginal advantage of our method over the other single-sequenced-based methods is still meaningful. In the traditional evaluation of this field, nearly all previous works only calculate the mean values of metrics and rank methods without considering the standard deviations. Thanks to your prior suggestion, in this work, we provide both the mean value and the standard deviation among all tested proteins for each evaluation metric. Given our results in Table 2 (*i.e.* the original Table 4), most of the mainstream methods show only marginal performance difference when statistical factors like the standard deviation are taken into account. For instance, even the best-performing GEMME and TranceptEVE L do not show significant differences compared to other methods. Therefore, in evaluating various methods, we still adhere to the traditional approach in the field, that is, ranking the results based on the mean.

Nevertheless, we acknowledge that your suggestion is valid and have appropriately toned down the text in the revised manuscript:

"As shown in Table 2 for the prediction of single mutational effects, at least in terms of the Spearman correlation coefficient, AUC and MCC, SPIRED-Fitness is only slightly inferior to two state-of-the-art MSA-based predictors, TranceptEVE L and GEMME, and outperforms the other single-sequence-based methods, although the advantage diminishes in the NDCG and the top 10% recall."

Notably, the original Table 4 has been renumbered as Table 2 in the revised manuscript.

4. Likewise, the main text description of Table S7 is factually correct but does not acknowledge that the MSA-based methods have better performance. It is also unclear why Table S7 includes the average column because the random, modulo, and contiguous evaluation settings are not directly comparable.

Answer:
In this round of revision, we developed a new model training/testing protocol for the supervised setting of ProteinGym. For traditional ProteinGym baselines as well as our choice in the last round of revision, a specific model should be independently trained and cross-validated for each individual assay. Hence, the number of repeated model training and/or hyper-parameter selection amounts to 15 for each assay: 3 data splitting strategies and 5 fold of cross validations. As a consequence, in the last round of revision, we could only provide evaluation results for the supervised setting on 20 assays due to the heavy computational costs. However, we realized recently that our model allowed the training of a unified model using mixed-sourced ProteinGym data due to the unique design of our loss function (*i.e.* the Soft Spearman Loss). Hence, in this round of revision, for each data splitting strategies (*i.e. Random*, *Modulo* and *Contiguous*) and each fold of cross validation experiment, we combined data from the overall 201 ProteinGym assays to train a universal SPIRED-Fitness model. This unified model was then evaluated within each individual assay following the traditional ProteinGym evaluation process. On the one hand, our new protocol greatly

simplified the overall evaluation process and thus allowed a comprehensive and efficient performance evaluation over all ProteinGym assays, considering that the number of repeated model training was reduced from 3,015 to 15 for the overall 201 assays. On the other hand, the unified SPIRED-Fitness model was capable of learning general mutational effects from multiple DMS assays, which elicited further performance enhancement in the supervised setting of ProteinGym. According to our new results in Supplementary Table S8 (*i.e.* the original Table S7), the new version of the supervised SPIRED-Fitness performs only behind ProteinNPT but is better than the other MSA-based methods and all single-sequence-based methods. We have also changed the words in the section entitled "Benchmark of SPIRED-Fitness on ProteinGym" in the revised manuscript to describe these new results.

Regarding the issue of the Average column in Supplementary Table S8 (*i.e.* the original Table S7), we followed the standard proposed by Notin *et al*. Please see the articles of ProteinGym (doi:10.1101/2023.12.07.570727) and ProteinNPT (doi:10.1101/2023.12.06.570473). Moreover, this presentation is consistent with the ways by which results are displayed on the ProteinGym official website. Although we also believe that simple averaging over different data splitting strategies may not be entirely appropriate, we have to retain the Average column in order to align with the mainstream evaluation practices in the field.

5. The response letter and revised manuscript still do not explain why the other methods in Figure S8 perform so poorly. Most of these same methods are compared to SPIRED in Table 4 in the ProteinGym comparison, where the overall correlations are similar. The datasets used in Figure S8 are mostly all in ProteinGym as well, though the exact overlap is not reported. What factors drive the large performance differences, both the absolute correlations and relative performance between SPIRED and the other strong baselines?

Answer:
This is a good question. We think that the discrepancies between Table 2 (*i.e.* the original Table 4) and Supplementary Figure S8 may be attributed to the following possible reasons.

(1) Different Datasets: The dataset used in Supplementary Figure S8 includes some data not present in ProteinGym. For instance, only parts of the cDNA proteolytic data have been integrated into ProteinGym. Additionally, ProteinGym evaluation in Table 2 (*i.e.* the original Table 4) was attained by the rigorous 5-fold cross validation, whereas that in Supplementary Figure S8 was achieved by testing once on the randomly selected 20% of the overall data, a strategy that might introduce fluctuation in the results.

(2) Comparison Methods: In Supplementary Figure S8 and Table S5 (*i.e.* the original Table S4), all models were evaluated as supervised methods in a unified manner without treating the zero-shot and supervised methods separately. We understand that this approach might introduce an artificial advantage for the supervised methods. However, we conducted a fair and more rigorous comparison in the following section entitled "Benchmark of SPIRED-Fitness on ProteinGym" as a remedy.

(3) Difference in MSA searching methods: Despite the variations in the specific data values, the overall trends in Table 2 (*i.e.* the original Table 4) and Supplementary Table S5/Figure S8 are consistent. For example, the results of single-sequence-based methods like ESM-1v and ESM-1b demonstrate that Table 2 (*i.e.* the original Table 4) and the non-proteolytic column of Supplementary Table S5 are largely congruent, with only tiny distinctions. The final results of MSA-based methods are, however, influenced by the MSA searching approach additionally. In Supplementary Figure S8 and Table S5, we used only the UniRef30 (doi:10.1186/s12859-019-3019-7) and BFD (doi:10.1038/s41592-019-0437-4) databases, which may result in slightly shallower MSA depths in comparison to the ProteinGym pipeline that typically recruits more sequence data for MSA search. In this work, we did not replicate the

entire ProteinGym evaluation process, but downloaded the evaluation results from the ProteinGym official website. This may lead to some discrepancies in the results of MSA-based methods.

6. The text states the models have similar performance on proteolytic and non-proteolytic data based on Table S2, but the actual gap is substantial for all models.

Answer:
I am sorry that this is a misunderstanding. In the original manuscript, we meant that different models (*e.g.*, SPIRED-Fitness and ECNet) exhibited similar trends across the two categories of data, thereby ruling out the possibility of training bias in SPIRED-Fitness, since ECNet training is unaffected by the mixing of multi-sourced data. As for the comparison between proteolytic and non-proteolytic data, of course, their results have significant difference for all models.

In the revised manuscript, we have slightly modified the text to avoid the potential confusion:

"As shown in Supplementary Table S3, the generally similar level of performance between SPIRED-Fitness and ECNet within each of the two categories of data (in terms of single + double mutations) negates the presence of significant bias elicited by data imbalance during model training, considering that each ECNet model is optimized for one individual protein using one specific set of DMS data. "

Notably, the original Table S2 has been renumbered as Supplementary Table S3 in the revised manuscript.

7. Figure 3d is still a bar plot. Only the caption was changed.

Answer:
I am sorry. But we are quite sure that Figure 3d had been changed to a boxplot in the last version of manuscript. Nevertheless, to ensure that the correct boxplot is included in this figure, we have double-checked the manuscript. Please see this figure in the updated version of the manuscript.

8. The definition of the Pearson correlation coefficient correlation was incorrectly changed to use the standard error instead of the standard deviation.

Answer:
Following your suggestion, we have changed "standard error" to "standard deviation" in the revised manuscript. Thank you.

9. Section 2.5.1 in the supplement is blank.

Answer:
Section 2.5.1 is not blank actually, because its content is Supplementary Table S10.

10. I was also asked to comment on how the revised manuscript addressed the concerns of the original Reviewer #4. Overall, those concerns have also been satisfied. The revised manuscript now includes benchmarking against EMBER3D, AlphaFold2 with an MSA, and AlphaFold2 without an MSA in the main structure

prediction evaluation (Figure 2). There is additional emphasis on the evaluation by protein fold, including new supplementary tables showing which folds lead to poor performance for SPIRED, OmegaFold, and ESMFold. The supplement also includes new discussion of the differences between these models and their training strategies that could be driving some of the fold-specific performance differences. A minor comment is that the response letter explanation regarding the effects of GDFold2 is quite good, but these are still not obvious in the manuscript itself. When discussing the results of Figure 2, briefly summarizing the tradeoffs between C_α coordinates and TM-score versus overall structure quality that are presented in the response would provide readers who lack this background important context.

Answer:
This is a nice suggestion. Thank you. We have added a brief description of the tradeoff between global and local structural quality by GDFold2 optimization in the revised manuscript and have included a detailed explanation and discussion in a new **section 1.7** of the Supplementary Materials, entitled "GDFold2 optimization and the tradeoff between global and local structural quality", where we have also inserted a new Supplementary Table S1 to enable numerical comparison on the global and local structural quality.

1.3 Remarks on code availability

1. I reviewed the updates to the SPIRED-Fitness GitHub repository, and they addressed all of the comments raised in my initial review. The manuscript now has a new Zenodo repository https://zenodo.org/doi/10.5281/zenodo.12560925, but the manuscript, web server, and GitHub readme still reference the old Zenodo URL.

Answer:
We have updated the corresponding links in the manuscript, web server and GitHub readme file. Thank you.

2 Responses to Reviewer #3

2.1 Remarks to the Author

1. The authors have thoroughly addressed my and the other reviewers' questions. Major changes include results from software that were not included in the previous manuscript but overall make the revised manuscript stronger; better presentation of the results section, with supplemental tables showing the single and double mutation datasets separately; improvements to the server that will benefit the user; and an overall improved language in the manuscript that makes this work more accessible to the broad readership of the journal.

Answer:
Thank you.

2.2 Remarks on code availability

1. SPIRED-Fitness was successfully installed and run in the first round of the review. The server runs without issues.

Answer:
Thank you.