nature research

Peer Review Information

Journal: Nature Computational Science
Manuscript Title: Persistent spectral theory-guided protein engineering
Corresponding author name(s): Guo-Wei Wei

Editorial Notes:

Redactions – published data	Parts of this Peer Review File have been redacted as indicated to remove third-party material.

Reviewer Comments & Decisions:

Decision Letter, initial version:

Dear Professor Wei,

Your manuscript "TopFit: topology-offered protein fitness" has now been seen by 3 referees, whose comments are appended below. You will see that while they find your work of interest, they have raised points that need to be addressed before we can make a decision on publication.

The referees' reports seem to be quite clear. Naturally, we will need you to address *all* of the points raised.

While we ask you to address all of the points raised, the following points need to be substantially worked on:

* Please make sure to better explain and clarify the persistent homology and Laplacian spectral analysis aspects of the work, which were the focus of Referee #2, an expert in computational topology.
* Please test the method's generalizability to unseen protein sites.
* Please better clarify the performance improvements as suggested by Referee #1.

Please use the following link to submit your revised manuscript and a point-by-point response to the referees' comments (which should be in a separate document to any cover letter):

[REDACTED]

** This url links to your confidential homepage and associated information about manuscripts you

may have submitted or be reviewing for us. If you wish to forward this e-mail to co-authors, please delete this link to your homepage first. **

To aid in the review process, we would appreciate it if you could also provide a copy of your manuscript files that indicates your revisions by making of use of Track Changes or similar mark-up tools. Please also ensure that all correspondence is marked with your Nature Computational Science reference number in the subject line.

In addition, please make sure to upload a Word Document or LaTeX version of your text, to assist us in the editorial stage.

To improve transparency in authorship, we request that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We hope to receive your revised paper within three weeks. If you cannot send it within this time, please let us know.

We look forward to hearing from you soon.

Best,
Fernando

--
Fernando Chirigati, PhD
Chief Editor, Nature Computational Science
Nature Portfolio

Reviewers comments:

Reviewer #1 (Remarks to the Author):

This manuscript proposed TopFit, a novel machine learning (ML) method to learn the sequence-function relationship for protein engineering. In contrast to previous ML methods that mainly leverage sequence-based models, including protein language models or generative models built on multiple sequence alignment, TopFit further incorporates 3D protein structure data through a topology technique called persistent spectral theory (PST), previously developed by the authors that generalized persistent homology and combinatorial Laplacians.

Comprehensive evaluations were performed to assess TopFit's accuracy in protein fitness prediction. Using several ablation studies, the authors first demonstrated that the PST embedding was a better protein representation for fitness prediction than existing sequence embeddings. Then the authors combined PST embeddings with two other evolutionary scores to build their model TopFit. Evaluated on ~30 deep mutagenesis datasets, TopFit achieved clear improvements over existing methods in those tests.

Overall, this manuscript proposed a novel topology-based ML model for protein engineering. The authors are aware of the literature on this topic and have designed comprehensive experiments to evaluate key aspects of the model. The evaluation performances achieved by the proposed method were strong.

I have a few comments to help strengthen the manuscript, as detailed below.

Major comments:
- Train/test split. The training and test sets were randomly split in this work. In protein engineering, the random split may create a "too easy" prediction task. For example, some sites in the protein are very intolerable to mutations, and as long as a few mutations of this position were sampled in the training set, the ML model would easily learn this position prior and correctly predict the mutation effect for other mutations on the same position. I suggest the authors split the train/test sets by positions in the sequence, i.e., testing the model's generalizability to unseen sites.

- Ensemble ablation. In Fig 3, the ensemble regression was used to predict fitness from embeddings, which includes three predictive models, namely ANN, kernels, and tree models. I am curious when those embeddings are combined with only one of the three predictive models individually, do we still have a performance trend and ranking as in Fig 3a? This ablation analysis will help us better understand which predictive model is better suited for a specific embedding.

- Explaining the improvements. If Fig 4-5, clearly TopFit, by incorporating structure data, has improved a lot over existing sequence-based approaches. It can also be noted that TopFit outperformed other methods by different margins on different proteins. So a natural question to ask is on which proteins did TopFit obtain larger improvements? The story of this work would be more convincing if the larger improvements are really due to the integration of structure data.

- Comparison to the state-of-the-art method. The following recent paper has demonstrated state-of-the-art prediction accuracy on ML for protein engineering. The authors can consider discussing and/or comparing with that method.

Notin, Pascal, et al. "Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval." International Conference on Machine Learning (ICML). PMLR, 2022. (https://proceedings.mlr.press/v162/notin22a/notin22a.pdf)

Minor point:
Page 11, "EcNet" -> "ECNet"

Reviewer #2 (Remarks to the Author):

This paper presents a substantial analysis of protein structure data using various mathematical summaries in machine learning models to conduct deep mutational screening.

My review focusses on the persistent homology and Laplacian spectral analysis aspects of the work. The authors combine the topological and geometric information provided by these quantities with standard amino-acid sequence data to achieve highly successful predictions of relative fitness of protein mutations

I find the presentation of persistent homology and Laplacian a little brief, and some of the claims made need greater justifications. First, the claim made in the abstract and on p.5, that `persistent homology fails to capture homotopic shape changes' rather depends on what objects are being compared. The example given in Figure 2, is the dimension 0 and 1 persistence barcodes for the filtration of a single fixed object. The illustration points out that the topology doesn't change between the two central complexes, although more simplices are added. This is true, but if we were to compare the persistence barcodes of two different objects that are homotopic but different shapes, then the barcodes of these two shapes will be different. i.e., the end-points of each bar will differ, and possibly also the number of bars if the geometric change is significant. For example consider points uniformly spaced around a circle, an ellipse, and the outline of an hour-glass. The circle and ellipse will each have a single bar in PH_1, with different end points, while the hour-glass will have two bars. There are standard methods in persistent homology that allows us to compare different barcodes, called the bottleneck and Wasserstein distances, the authors have not used or referred to these.

Second, the claim that persistent spectral theory `recovers the full topological persistence' is a significant one, and explained only by reference to a previous paper [34]. That paper does not contain a formal proof, but it does define a p-persistent combinatorial Laplacian (in dimension q) for a filtration $\{K_t\}$, and states that the nullity of this matrix is the same as the rank of the p-persistent q-dimensional homology for K_t, i.e., a type of betti number. This seems reasonable, but it is not the same as `full topological persistence'. Incidentally, most of the examples in [34] then set $p=0$ which means no actual persistence information is used at all.

Further details about the filtrations used and information derived from persistent homology and persistent spectra are given in the Methods section. The section `Persistent spectral theory' on p.20--21 greatly simplifies the material in [34]. In particular, the definition of `the persistent k-combinatorial Laplacian' is in fact the 0-persistent version of the definition in [34]. Equation (19) defines the `persistent Betti number', $\beta_k^{t}$, but in fact this is just the regular Betti number in dimension $k$ of the simplicial complex, $K_t$. For something to be \emph{persistent}, information about how $K_s$ maps inside $K_t$ must be used. In short, everything called `persistent Laplacian' or `persistent Betti number' in this paper is really a \emph{parametrised} Laplacian and Betti number, where the parameter is the length-scale t. It is well known that the quantity β_k^t is highly unstable with respect to small changes in the complex K_t, whereas the bottleneck distance for persistence barcodes of alpha-complexes are stable with respect to small adjustments to the positions of points.

The text on pp.22-23 covers the specific quantities (feature vectors) used as input to the regression models. The PST feature vector treats 0-dimensional and 1-, 2-dimensional cases differently.
\begin{itemize}
\item In 0-dimension, the filtration is a Rips complex with D_{mod} distance, meaning K_t is a graph with vertices a, b joined when $a \in \mathcal{A}_1$ and $b \in \mathcal{A}_2$ and $||a - b||_2 \leq 2t$.
The eigenvalues of the graph Laplacian are computed at 10 values of t and summarised at each t by `the number of harmonic spectra' (i.e. the nullity or multiplicity of the e-value 0), and the minimum, maximum, mean, standard deviation and sum of the `non-harmonic spectra' (i.e. non-zero eigenvalues).

\item For dimensions 1 and 2, the alpha shape complex is used, and the persistence barcodes are found using the software package GUDHI. The vectorisation of the barcodes is made using seven summary statistics. This is a very crude method to summarise and `vectorise' persistent homology. The dominant methods used in the topological data analysis community are \emph{persistence

landscapes} and \emph{persistence images} both of which are known to be stable representations of persistent homology.
\end{itemize}

The PH feature vector is described as using the identical filtration construction to the PST case. The dimension 1, 2, case contains exactly the same information as for the PST feature vectors. The dimension-0 case `counts the number of bars' at particular length-scales. This is identical information to the `number of harmonic spectra' found for the PST feature vector.

So, from my reading of the information provided, the PH feature vector is a strict subset of the PST feature vector. Small wonder then that using the PST feature vector gives better results than the PH feature vector in the regression analysis. The author's statement on p.14 that `PST significantly outperforms persistent homology in all datasets' exaggerates the distinction between the two methods as it implies that they contain different information. In the current paper, the PST features simply add extra information derived from a graph Laplacian.

Nonetheless, the authors have clearly achieved interesting results and demonstrated that geometric and topological information at multiple length-scales enhances the capacity for predicting protein mutational fitness.

Reviewer #3 (Remarks to the Author):

``TopFit: topology-offered protein fitness'' by Qiu and Wei
Protein engineering has significantly changed the landscape of food production and also becomes an important approach for drug discovery and enzyme catalysis. Protein can now be designed on a computer and then, realized through directed evolution in the laboratory. However, most computer-aided design of protein engineering is based on protein sequences due to the availability of advanced natural language processing (NLP) algorithms and numerous protein sequences carrying evolutionary fitness information. In contrast, structure-based approaches were rarely used in computer-aided protein engineering.
This work seems to fill this gap by using an advanced mathematical tool, the persistent spectral graph (PSG) proposed by the Wei team (Ref. [34]). PSG is a new combinatorial graph tool in topological data analysis (TDA), which has emerged as a popular new field in data science in the past decade. The Wei team's PSG can recover the topological invariants as those from persistent homology and can also capture the homotopic shape evolution in the multiscale analysis. The Wei team is one of the leaders in the development of new advanced mathematical tools for solving biological problems.
The proposed TopFit was validated on 34 benchmark datasets to establish a new state-of-the-art. The proposed method was compared with a large number of existing methods in the literature. The authors' topology-based structure-based methods outperform sequence-based methods. The proposed TopFit complements structure-based methods and sequence-based methods to achieve high accuracy, reliability, and robustness. The proposed model explored tens of machine learning algorithms and an ensemble of regressors was constructed to enable the robustness against data size variation and quality diversity. With the sharing of the source code, TopFit could have a major impact on computer-aided protein engineering. This paper is well-written and deserves publication in Nature Computational Science. Nonetheless, after carefully going through the manuscript, I have found some minor problems that need to be fixed or clarified in the revision.
1) Is it necessary to use 18 regression models in the ensemble regressor? Why do not use the top 3 or 5 performing models as an ensemble?

2) The authors need to elaborate on why their method offers a new answer to the question "Can one hear the shape of a drum?".
3) In Figure 3. The authors need to give references to Hsu et al and Luo et al. The results from the literature in Figure 3 should be properly cited.
4) It is interesting to note that the NMR structure leads to better performance than AlphaFold structure does. Do the authors use the latest AlphaFold2? If so, this is an interesting issue as the community does not have a good idea about the relative reliability of Alphafold2 predicted structures. Is this a general phenomenon or just for the specific proteins used? The quality of AlphaFold2 structures may not be uniform.
5) A list of all datasets used should be given in the Supplementary Information and the original sources should be carefully cited.
6) In the Supplementary Information, involved datasets and literature results should be carefully cited so the reader can better trace the original work.
7) ``An exception is on the GB1 dataset'' ---> ``An exception was found on the GB1 dataset''.

Author Rebuttal to Initial comments

Reviewer #1 (Remarks to the Author):

This manuscript proposed TopFit, a novel machine learning (ML) method to learn the sequencefunction relationship for protein engineering. In contrast to previous ML methods that mainly leverage sequence-based models, including protein language models or generative models built on multiple sequence alignment, TopFit further incorporates 3D protein structure data through a topology technique called persistent spectral theory (PST), previously developed by the authors that generalized persistent homology and combinatorial Laplacians.

Comprehensive evaluations were performed to assess TopFit's accuracy in protein fitness prediction. Using several ablation studies, the authors first demonstrated that the PST embedding was a better protein representation for fitness prediction than existing sequence embeddings. Then the authors combined PST embeddings with two other evolutionary scores to build their model TopFit. Evaluated on ~30 deep mutagenesis datasets, TopFit achieved clear improvements over existing methods in those tests.

Overall, this manuscript proposed a novel topology-based ML model for protein engineering. The authors are aware of the literature on this topic and have designed comprehensive experiments to evaluate key aspects of the model. The evaluation performances achieved by the proposed method were strong.

I have a few comments to help strengthen the manuscript, as detailed below.

Response: We appreciate the reviewer's insightful comments

Major comments:

- Train/test split. The training and test sets were randomly split in this work. In protein engineering, the random split may create a "too easy" prediction task. For example, some sites in the protein are very intolerable to mutations, and as long as a few mutations of this position were sampled in the training set, the ML model would easily learn this position prior and correctly predict the mutation effect for other mutations on the same position. I suggest the authors split the train/test sets by positions in the sequence, i.e., testing the model's generalizability to unseen sites.

Response: Thank you for the suggestion. It is an important task for protein engineering to predict unseen mutational sites. In the revision, we added new computations for splitting train/test sets without overlap sites. The details are discussed in the **new Supplementary Note 4** along with 5 figures (**new Supplementary Figures 19-23**).

Our main findings and remarks are briefly discussed in the Discussion section:

Another extrapolation task is to predict mutations at unseen sites. Despite our PST remains to be ranked as the best embedding, all embeddings suffer from reduced accuracy comparing to the random split where they all largely underperform evolutionary scores. Our TopFit inherits advantages from all methods, particularly evolutionary scores, and it achieves small improvement over evolutionary scores (Supplementary Figure 19-Supplementary Figure 23; Supplementary Note 4). However, predicting out-of-distribution data may violate the nature of supervised models, and how to build a more accurate supervised model for this task is interesting. Nonetheless, the unsupervised evolutionary scores may be more effective for extrapolation at the early protein engineering stage with insufficient training data (i.e., low-N case).

- Ensemble ablation. In Fig 3, the ensemble regression was used to predict fitness from embeddings, which includes three predictive models, namely ANN, kernels, and tree models. I am curious when those embeddings are combined with only one of the three predictive models individually, do we still have a performance trend and ranking as in Fig 3a? This ablation analysis will help us better understand which predictive model is better suited for a specific embedding.

Response: Thank you for the good suggestion. In the original submission, we have compared the full ensemble of 18 models with ridge regression. In the revision, we added three ensemble models by using 3 ANNs, 10 kernel or 5 tree models for comparisons (original **Figure 3d** and **Supplementary Figure 10**). In addition, **new Supplementary Figure 8** was added to show the average performance of embeddings using different ensembles.

Among different ensemble models in the ablation analysis, the full ensemble of 18 models performs slightly better than the other three partial ensemble models. The behaviors of different embeddings are similar to each other by using different ensembles. The discussions were added in the second paragraph in **Section 2.6**.

- Explaining the improvements. If Fig 4-5, clearly TopFit, by incorporating structure data, has improved a lot over existing sequence-based approaches. It can also be noted that TopFit outperformed other methods by different margins on different proteins. So a natural question to ask is on which proteins did TopFit obtain larger improvements? The story of this work would be more convincing if the larger improvements are really due to the integration of structure data.

Response: This is a good point. In the revision, we first added TopFit performance to original **Supplementary Figures 1-3** which allows direct comparisons on individual datasets for the TopFit improvement. However, it is difficult to summarize when TopFit has larger improvement from individual datasets. Indeed, we further investigated how different factors affect TopFit improvement. The **original Figure 3c**, **original Supplementary Figure 9**, **new Supplementary Figures 6-7** and the **new Supplementary Note 3** addressed this question from many aspects.

Specifically, we performed analysis by classifying the 34 datasets based on **1)** the quality of structure data (original submission); **2)** type of protein classified by taxonomy (revision); and **3)** type of fitness (revision). The analysis enhances the understanding of the strength of TopFit. Details are given below.

1) In the original submission, we have shown PST embedding may achieve larger improvement over sequence-based methods if the 3D structure data has high quality. Particularly, B-factors from X-ray structure provides a direct quantity for the data quality. The percent of random coils in one structure implicitly affects the accuracy of structure data. We showed both quantities are correlated to the PST embedding performance (original Figure 3c and Supplementary Figure 8).

2) In the revision, we further examined embedding performance on datasets categorized by taxonomy for eukaryote, prokaryote, and human datasets (**new Supplementary Figure 6)**. TopFit has the largest improvement on human datasets with 16.6% and 25.1% improvement over ESM embedding and VAE score, respectively, for 240 training data.

3) In addition, we added **new Supplementary Figure 7** to compare average performance on datasets classified by the fitness type measured by experiments. Evolutionary scores, sequence-based embedding, and PST embedding behave differently on each class. TopFit combining them all shows improvement over all single strategy. It has largest improvement over VAE score on class of binding

(48.8%) and it has largest improvement over the ESM embedding on class of enzyme activity (29.3% improvement) when training data size is 240.

Details are discussed in the **new Supplementary Note 3.** We also added a few sentences for the results for datasets classified by taxonomy in the last paragraph in **Section 2.4**.

- Comparison to the state-of-the-art method. The following recent paper has demonstrated state-of-the-art prediction accuracy on ML for protein engineering. The authors can consider discussing and/or comparing with that method.
Notin, Pascal, et al. "Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval." International Conference on Machine Learning (ICML). PMLR,
2022. (https://proceedings.mlr.press/v162/notin22a/notin22a.pdf)

Response: Thank you for pointing out the reference for evolutionary scores. In the revision, we further included two top evolutionary scores reported by this paper: Tranception [1] and EVE [2] (**revised Figure 3a, Supplementary Figures 1-5**). One original Supplementary Figure is split into two (**new Supplementary Figures 16-17**) to test TopFit performance by integrating single score or multiple scores for further comparisons. The descriptions of these two scores were added in Introduction, Results, and Methods. We also added additional points in Discussions for the potential of TopFit in including state-of-art evolutionary scores. The details are described below.

1. **First, our comparisons added Tranception and EVE scores along with three scores discussed in the original submission** (**Figure 3a** and **Supplementary Figures 1-5** were revised accordingly). We found DeepSequence VAE (the one we mainly used in TopFit) is the best evolutionary score among the five scores averaged on the 34 datasets. As a result, we kept using DeepSequence VAE in TopFit for our main results. While TopFit using different evolutionary scores are also discussed (see **Point 2**).

   DeepSequence VAE achieved the highest average Spearman correlation 0.504 over 34 datasets. Tranception underperforms VAE with average Spearman correlation 0.477. EVE shows slight underperformance with average Spearman correlation 0.497 (1.4% lower Spearman). In addition, we found DeepSequence VAE has the highest frequency to be ranked as the best model measured by either Spearman correlation or NDCG (**Figure R1**).

Our benchmarks subsample the datasets to have comparisons with supervised models. To exclude the possibility that the underperformance is caused subsampling, we looked into the comparisons reported in Tranception (https://www.proteingym.org). With available reports for 31/34 datasets in Tranception work, we found DeepSequence outperforms Tranception which are consistent with our discovery (**Table R1**). And DeepSequence slightly underperforms EVE (1.4% lower Spearman), where the performance is similar to the level reported by us. Tranception has large margin over DeepSequence VAE on viral protein datasets. But 34 datasets discussed in our work have no viral protein involved (i.e. datasets from DeepSequence paper). This may be the reason that Tranception did not show the best performance in this work. But Tranception and EVE are also powerful with significantly better performance than ESM and eUniRep scores. Remarks on these points were added as described in **point 3** below.

[REDACTED]

2. **Second, we further included more evolutionary scores in TopFit, i.e., either individual score or multiple of individual ones.** Since VAE DeepSequence showed the best average performance on 34 datasets we tested, we kept it in TopFit for main results for a proof of principle (**Figures 3-5**). But our TopFit can in fact integrate with any evolutionary scores to have accurate performance.

   To further examine TopFit performance with different evolutionary scores, we performed additional computations to integrate them (**new Supplemenatary Figures 16-17**). Instead of using single score each time, we also tested the integration of multiple scores in one TopFit simulation. The integration of multiple scores allows better TopFit performance over the one with single score. The integration of two scores, Tranception and VAE, in TopFit achieves the best performance over the integration of multiple scores. The detailed discussions are given in **revised Supplementary Note 2**.

3. **Last, we added additional comments for Tranception and EVE for their descriptions, comparisons, and potentials in TopFit (Introduction, Section 2.7, Methods, and Discussion).**

   For example, we pointed out the advances of Tranception in Introduction:

   The Transformer-based Tranception not only score mutations via a global autoregressive inference, but also a local retrieval inference utilizing MSAs [1]. While the majority of methods only evaluates mutations from substitution, the combined Tranception score from two inferences can also predict mutations from insertions and deletions.

   In Discussions, we discussed TopFit ability in integrating state-of-art evolutionary scores:

TopFit provides a general framework to build supervised protein fitness models by combining PST structural features with sequence-based features. Arbitrary sequencebased embedding and evolutionary scores can be used**. TopFit can be continuously improved by the quickly evolving state-of-art sequence-based models.** In this work, ESM Transformer embedding was directly applied to generate sequence embedding. TopFit may be further improved using the fine-tune procedure in ESM [21**]. In this work, we mainly tested TopFit with single type of score, especially, DeepSequence VAE. Tranception score may be more powerful than VAE for datasets with low MSA depths or viral protein datasets [25]. The equipment of any evolutionary score can largely enhance model generalization and accuracy to training set with various sizes (Supplementary Figure 16). Interestingly, inclusion of multiple scores can further improve TopFit performance (Supplementary Note 2 and Supplementary Figure 17).** Even more, one can improve performance by combining multiple structure data for ensemble predictions (Supplementary Figure 12). **However, all these approaches grant additional computational costs. The combined features tested in this work provide the minimal models that combine models built on distinct data resources such as local homologous sequences, large-scale sequence data, and three-dimensional structure data.**

[1] Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval. Notin et al., PMLR, 2022

[2] Disease variant prediction with deep generative models of evolutionary data. Frazer et al., Nature, 2021.

Minor point: Page 11, “EcNet” -> “ECNet”

Response: Thanks for pointing this out. We have corrected it in the revision.

Reviewer #2 (Remarks to the Author):

This paper presents a substantial analysis of protein structure data using various mathematical summaries in machine learning models to conduct deep mutational screening.

My review focusses on the persistent homology and Laplacian spectral analysis aspects of the work. The authors combine the topological and geometric information provided by these quantities with standard amino-acid sequence data to achieve highly successful predictions of relative fitness of protein mutations

Response: Thank you for the nice summary from the reviewer.

I find the presentation of persistent homology and Laplacian a little brief, and some of the claims made need greater justifications. First, the claim made in the abstract and on p.5, that `persistent homology fails to capture homotopic shape changes' rather depends on what objects are being compared. The example given in Figure 2, is the dimension 0 and 1 persistence barcodes for the filtration of a single fixed object. The illustration points out that the topology doesn't change between the two central complexes, although more simplices are added. This is true, but if we were to compare the persistence barcodes of two different objects that are homotopic but different shapes, then the barcodes of these two shapes will be different. i.e., the end-points of each bar will differ, and possibly also the number of bars if the geometric change is significant. For example, consider points uniformly spaced around a circle, an ellipse, and the outline of an hour-glass. The circle and ellipse will each have a single bar in PH_1, with different end points, while the hour-glass will have two bars. There are standard methods in persistent homology that allows us to compare different barcodes, called the bottleneck and Wasserstein distances, the authors have not used or referred to these.

Response: Thanks for pointing this out. It is correct that persistent homology can distinguish two sets of point cloud that are homotopic with different shapes. In our manuscript, the **homotopic shape evolution** in the filtration is a different concept. It is specifically for the shape evolution of simplicial complexes in the filtration of a given point cloud. PST can fully capture the homotopic shape evolution during the filtration, while persistent homology can only capture changes in topological invariants during the filtration.

In the revision, we clarified this by explicitly pointing out the homotopic shape evolution during **the filtration process**. In Abstract we wrote:

Persistent homology, an established algebraic topology tool for protein structural complexity reduction, fails to capture the homotopic shape evolution **during the filtration of a given data**.

In Introduction, we clarified this:

Filtration of a given point cloud may induce both homotopic shape evolution and changes in topological invariants, which allow multiscale analysis from TDA. However, persistent homology only captures changes of topological invariants and is insensitive to the homotopic shape evolution.

Furthermore, we pointed this out in Section 2.2 (p5):

However, it is not sensitive to homotopic shape evolution of data given by filtration.

Second, the claim that persistent spectral theory `recovers the full topological persistence' is a significant one, and explained only by reference to a previous paper [34]. That paper does not contain a formal proof, but it does define a p-persistent combinatorial Laplacian (in dimension q) for a filtration $\{K_t\}$, and states that the nullity of this matrix is the same as the rank of the p-persistent q-dimensional homology for K_t, i.e., a type of betti number. This seems reasonable, but it is not the same as `full topological persistence'. Incidentally, most of the examples in [34] then set $p=0$ which means no actual persistence information is used at all.

Response: This is a good point. First, for 0-persistent, the nullity of combinatorial Laplacian reveals the Betti numbers according to the combinatorial Hodge Theorem. Similar to the nonpersistent case, the nullity of the q-persistent Laplacian equals the q-persistent Betti (For a proof, please see ``*Facundo Mémoli, Zhengchao Wan, and Yusu Wang, Persistent Laplacians: Properties, Algorithms and Implications, SIAM Journal on Mathematics of Data Science, 4, 858-884, 2022.*''). In that case, the full topological persistence can be revealed by the number of harmonic spectra of p-persistent combinatorial Laplacian.

In the revision, we refer to this reference when we talked about “topological persistence” revealed by PST.

Further details about the filtrations used and information derived from persistent homology and persistent spectra are given in the Methods section. The section `Persistent spectral theory' on p.20--21 greatly simplifies the material in [34]. In particular, the definition of `the persistent $k$combinatorial Laplacian' is in fact the 0-persistent version of the definition in [34]. Equation (19) defines the `persistent

Betti number', β_k^{t}, but in fact this is just the regular Betti number in dimension k of the simplicial complex, K_t. For something to be
\emph{persistent}, information about how K_s maps inside K_t must be used. In short, everything called `persistent Laplacian' or `persistent Betti number' in this paper is really a \emph{parametrised} Laplacian and Betti number, where the parameter is the length-scale t. It is well known that the quantity β_k^t is highly unstable with respect to small changes in the complex K_t, whereas the bottleneck distance for persistence barcodes of alphacomplexes are stable with respect to small adjustments to the positions of points.

Response: Thank you for pointing this out. The persistence is a critical component in PST.

In the revision, we included descriptions of p-persistent q-combinatorial Laplacian and $p$persistent spectra and Betti numbers in Methods.

In dimension 0, the “parametrized” and “persistent” are the same concepts since all vertices have birth at 0 in the present setting. Then, we only focus on “persistence” for dimension 1 and 2. In the revision, we added results using non-harmonic spectra of 0- or p-persistent

Laplacians for high dimension (n=1 and 2) (**new Supplementary Note 5 and Supplementary Figures 25-26**). The inclusion of the non-harmonic spectra in the machine learning model has in fact slightly reduced accuracy. Different values of p for persistence do not make difference of performance. **Details are provided in answering the next question**.

We have a different view about persistence barcode stability. In molecular interactions, small length bars are important in our element-specific topological representation. The small bar length differences correspond to the differences in interaction strengths (please see ``Characteristics of TDA in applications'' in the article:

https://en.wikipedia.org/wiki/Topological_data_analysis). Additionally, we have also carried out computations of biomolecules using the bottleneck distance and p-Wasserstein distance metrics in our early work (*Zixuan Cang, Elizabeth Munch, Guo-Wei Wei, Evolutionary homology on coupled dynamical systems with applications to protein flexibility analysis. Journal of Applied and Computational Topology (2020) 4:481-507*). Our experience is they do not offer better results.

The text on pp.22-23 covers the specific quantities (feature vectors) used as input to the regression models. The PST feature vector treats 0-dimensional and 1-, 2-dimensional cases differently.
\begin{itemize}

\item In 0-dimension, the filtration is a Rips complex with D_{mod} distance, meaning K_t is a graph with vertices a, b joined when $a \in \mathcal{A}_1$ and $b \in \mathcal{A}_2$ and $||a - b||_2 \leq 2t$.
The eigenvalues of the graph Laplacian are computed at 10 values of t and summarised at each t by `the number of harmonic spectra' (i.e. the nullity or multiplicity of the e-value 0), and the minimum, maximum, mean, standard deviation and sum of the `non-harmonic spectra' (i.e. non-zero eigenvalues).

\item For dimensions 1 and 2, the alpha shape complex is used, and the persistence barcodes are found using the software package GUDHI. The vectorisation of the barcodes is made using seven summary statistics. This is a very crude method to summarise and `vectorise' persistent homology. The dominant methods used in the topological data analysis community are \emph{persistence landscapes} and \emph{persistence images} both of which are known to be stable representations of persistent homology.
\end{itemize}

The PH feature vector is described as using the identical filtration construction to the PST case. The dimension 1, 2, case contains exactly the same information as for the PST feature vectors. The dimension-0 case `counts the number of bars' at particular length-scales. This is identical information to the `number of harmonic spectra' found for the PST feature vector.

Response: Yes, the reviewer is correct for all points here.

Proteins often have thousands of atoms. Rips complex can handle dimension-0 calculations of proteins. But it is too slow to do dimemsion-1 and 2. Alpha complex is used to speed up the calculations.

Our PST featurization mainly uses spectra at dimension 0, which provide the most important information with basic connectivity between vertices. The features at high dimension are simplified to retain only essential information. Our designs are out of consideration of our practical problem to have a reasonably small number of features to prevent the overfitting issue in fitness prediction, especially, for the small training data.

For high dimension, additional non-harmonic spectra or the popular persistent landscape or persistent image for vectorization can provide more information, but they lead to much larger number of features, which may cause more severe overfitting problems and more expensive computational cost. In particular, our site- and element-specific strategies are designed to extract biophysical information of the proteins, which will largely amplify the feature vector. For example, our “crude” treatment for dimension 1 and 2 creates small size of feature vector (N=7) for each point cloud, but the final feature

vector already has $60N=420$ dimension. In the revision, we further explored the possibility by including 1) persistent landscape vectorization **(new Supplementary Note 5; new Supplementary Figure 24)** or 2) nonharmonic spectra at dimension 1 and 2 **(new Supplementary Note 5; new Supplementary Figures 25-26)**. However, all of them render a lower accuracy than our original "simplified" vectorization. We added a comment in Discussions:

PST features for dimension 0 have relatively high dimension to provide the most critical information with basic connectivity between vertices. For dimensions 1 and 2, our PST features are relatively crude to retain essential information and keep the feature dimension low under the element- and site- specific strategies. The low-dimensional features can better accommodate with machine learning models in avoiding overfitting issues for small training data size, as well as reducing computational costs. Although the stable representations such as persistent landscape [56] and the persistent image [57] for persistent homology, and a potential informative representation of non-harmonic persistent spectra may provide more enrich information, they typically generate a large number of features (Supplementary Note 5). How to handle potential overfitting from these representations in biomolecular systems remains an interesting issue.

Here are some details for our exploration in **new Supplementary Note 5**:

1) **Persistent landscape**: We further compared our original featurization for dimension 1 and 2 with the featurization using persistent landscape. We used small resolution 10 and 3 landscapes. Then the feature vector is already as high as 1800 dimension using site- and element-specific strategies. By replacing our original feature by the landscape feature, slightly reduced performance was observed (**new Supplementary Figure 24**).

2) **Non-harmonic persistent spectra at dimension 1 and 2**: We computed the p-persistent Laplacian for dimensions 1 and 2, and use their non-harmonic spectra in featurization. We also explored 9 values of p=0.0,0.1,0.2,0.3,0.4,0.5,1.0,1.5, and 2.0. The inclusion of non-harmonic feature leads to slightly reduced performance (**new Supplementary Figures 25-26**).

So, from my reading of the information provided, the PH feature vector is a strict subset of the PST feature vector. Small wonder then that using the PST feature vector gives better results than the PH feature vector in the regression analysis. The author's statement on p.14 that `PST significantly outperforms persistent homology in all datasets' exaggerates the distinction between the two methods as it implies that they contain different information. In the current paper, the PST features simply add extra information derived from a graph Laplacian.

Response: We agree with this point. The outperformance of PST over PH is mainly due to the inclusion of non-harmonic spectra. In the revision, we revised the sentence on p.14 to:

With the inclusion of non-harmonic spectra, PST significantly outperforms persistent homology in all datasets.

Nonetheless, the authors have clearly achieved interesting results and demonstrated that geometric and topological information at multiple length-scales enhances the capacity for predicting protein mutational fitness.

Response: We thank the reviewer for the positive comment.

Reviewer #3 (Remarks to the Author):

``TopFit: topology-offered protein fitness'' by Qiu and Wei

Protein engineering has significantly changed the landscape of food production and also becomes an important approach for drug discovery and enzyme catalysis. Protein can now be designed on a computer and then, realized through directed evolution in the laboratory. However, most computer-aided design of protein engineering is based on protein sequences due to the availability of advanced natural language processing (NLP) algorithms and numerous protein sequences carrying evolutionary fitness information. In contrast, structure-based approaches were rarely used in computer-aided protein engineering.

This work seems to fill this gap by using an advanced mathematical tool, the persistent spectral graph (PSG) proposed by the Wei team (Ref. [34]). PSG is a new combinatorial graph tool in topological data analysis (TDA), which has emerged as a popular new field in data science in the past decade. The Wei team's PSG can recover the topological invariants as those from persistent homology and can also capture the homotopic shape evolution in the multiscale analysis. The Wei team is one of the leaders in the development of new advanced mathematical tools for solving biological problems.

The proposed TopFit was validated on 34 benchmark datasets to establish a new state-of-theart. The proposed method was compared with a large number of existing methods in the literature. The authors' topology-based structure-based methods outperform sequence-based methods. The proposed TopFit complements structure-based methods and sequence-based methods to achieve high accuracy, reliability, and robustness. The proposed model explored tens of machine learning algorithms and an ensemble of regressors was constructed to enable the robustness against data size variation and quality diversity. With the sharing of the source code, TopFit could have a major impact on computer-aided protein engineering. This paper is well-written and deserves publication in Nature Computational Science. Nonetheless, after carefully going through the manuscript, I have found some minor problems that need to be fixed or clarified in the revision.

Response: We thank the reviewer for the praises.

1) Is it necessary to use 18 regression models in the ensemble regressor? Why do not use the top 3 or 5 performing models as an ensemble?

Response: Sorry for the confusion. Yes, we have performed ensemble for top models in the ensemble strategy. Our ensemble will rank all 18 models by cross-validation measured in RMSEs. The top N regressors will be selected for the ensemble. In particular, we used N=3 and N=5 for small and large sizes of training data in this work. Our ensemble model can select top models which may vary for different datasets and different sizes of training data (**Figure 3e; Supplementary Figure 11**).

In the revision, we clarify this point in **revised Figure 1** and its legend. From original submission, this point was mentioned in 4th paragraph in section 2.1 (Overview of TopFit). In the revision, we pointed out the top model selection when we first mentioned the ensemble model in early 4th paragraph. We rephrased the sentence to:

The design of the ensemble regression can fulfill the task by averaging predictions from top N models selected and ranked from a pool of multiple regressors [41].

2) The authors need to elaborate on why their method offers a new answer to the question “Can one hear the shape of a drum?”.

Response: We thank the reviewer for pointing out this problem. In the revision, we clarified this sentence:

PST comprehensively characterizes the geometry of an object from a family of frequencies induced by the filtration of evolving shapes, which provides an answer to the famous question: “Can one hear the shape of a drum?”, raised by Mark Kac [48].

3) In Figure 3. The authors need to give references to Hsu et al and Luo et al. The results from the literature in Figure 3 should be properly cited.

Response: Thanks for the point. We have added references for embeddings, and evolutionary scores used in the legend of Figure 3.

4) It is interesting to note that the NMR structure leads to better performance than AlphaFold structure does. Do the authors use the latest AlphaFold2? If so, this is an interesting issue as the community does not have a good idea about the relative reliability of Alphafold2 predicted structures. Is this a general phenomenon or just for the specific proteins used? The quality of AlphaFold2 structures may not be uniform.

Response: Sorry for the confusion. For the three datasets we tested, the predictions from single NMR structure achieves similar or worse performance than AlphaFold (**Supplementary Figure 12**). But we can make predictions from NMR better than AlphaFold via an ensemble from multiple NMR structures. These results only demonstrate the ensemble strategy can improve performance. They are not sufficient to make remarks on the relative qualities of AlphaFold and NMR structures.

In the revision, we reclarify this point in Discussions:

The AlphaFold structure achieves similar accuracy with single NMR structure. While the ensemble techniques using multiple NMR structures provide the chance to improve the performance (Supplementary Figure 12).

5) A list of all datasets used should be given in the Supplementary Information and the original sources should be carefully cited.
6) In the Supplementary Information, involved datasets and literature results should be carefully cited so the reader can better trace the original work.

Response: We thank the reviewer for the suggestion. In the revision, we added original sources of the datasets in section “Data availability”. In original submission, the PMID for source of each dataset was included in Supplementary Data 1. In the revision, titles of dataset papers were added to Supplementary Data 1.

7) ``An exception is on the GB1 dataset'' ---> ``An exception was found on the GB1 dataset''.

Response: Thanks for pointing out. We have corrected it and checked the grammar carefully during the revision.

Decision Letter, first revision:

Dear Dr. Wei,

Thank you for submitting your revised manuscript "TopFit: topology-offered protein fitness" (NATCOMPUTSCI-22-0836A). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Computational Science, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week or so. Please do not upload the final materials and make any revisions until you receive this additional information from us.

TRANSPARENT PEER REVIEW
Nature Computational Science offers a transparent peer review option for original research manuscripts. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't.** Failure to state your preference will result in delays in accepting your manuscript for publication.
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

Thank you again for your interest in Nature Computational Science Please do not hesitate to contact me if you have any questions.

Best,
Fernando

--
Fernando Chirigati, PhD

Chief Editor, Nature Computational Science
Nature Portfolio

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance:
https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Reviewer #1 (Remarks to the Author):

The authors have addressed my comments in the revision, and the manuscript was much improved.

Reviewer #1 (Remarks on code availability):

Detailed instructions in the README file. The GitHub has a demo dataset as an example but includes a script to download the full dataset. Pre-computed features are also provided.

Reviewer #2 (Remarks to the Author):

I thank the authors for the extra work they've done in response to questions raised by the reviews.

I'm happy with the author's clarification of homotopic shape evolution as referring to the changes during the evolution of the filtration.

I'm also glad that the authors have cited an additional mathematical paper establishing properties of the persistent Laplacian [36, Memoli et al]. They have also expanded the description of persistent Laplacian in the Methods section (p.22).

On the issue of "Can one hear the shape of a drum?" (p.7) It is known that the answer - in a strict sense - is "no"; see Buser, Conway, Doyle, Semmler "Some planar isospectral domains" (1994). But, on the other hand, there is clearly much information about the topology and geometry of a domain encoded in the Laplacian. I would recommend rephrasing the sentence on p.7 to say that the PST "provides another illustration of the famous question", rather than claiming it "answers" this question.

In the discussion (p.14) there is now further explanation that most of the information content of the PST features is contained in the non-harmonic spectra of the dimension-0 Laplacian, and that the dimensions-1 and -2 features are deliberately kept brief to reduce overfitting.

In the rebuttal the authors state: "We have a different view about persistence barcode stability. In molecular interactions, small length bars are important in our element-specific topological representation. The small bar
length differences correspond to the differences in interaction strengths."

I whole-heartedly agree that small persistence does not always imply "insignificant" (this is a common

mis-interpretation of persistent homology). However, on p.23 in the methods section the authors state that for dimension-0 "the short scales below 2 angstroms are excluded" and for dimensions-1, -2, "bars with lengths lower than 0.1 angstroms are excluded". So I'm a little confused here about what the underlying principle is for interpreting small changes in persistence in their model?

In Supplementary Note 5, and Figures 24, 25, 26, the authors summarise results from extra numerical computations conducted in response to the review. These show that using persistence landscapes versus their elementary statistical summaries do not significantly change the performance of their models. Similarly, adding p-persistent Laplacian information does not change any of their conclusions.

Small points:

page 2 line (-2) (i.e., filtration) -> (e.g., filtration)

page 2 line (-1) workhouse -> workhorse

page 7 line 1 second use of the word "harmonic" should be non-harmonic.

Reviewer #3 (Remarks to the Author):

The authros have provided extensive responses and made corresponding revisions to the reviewers' questions and comments. I am generally satisfied with these revisions and responses.

Author Rebuttal, second revision:

Reviewer #1

Reviewer #1 (Remarks to the Author):
The authors have addressed my comments in the revision, and the manuscript was much improved.

Reviewer #1 (Remarks on code availability):
Detailed instructions in the README file. The GitHub has a demo dataset as an example but includes a script to download the full dataset. Pre-computed features are also provided.

Response: We sincerely thank the reviewer for valuable suggestions and comments. We are pleasure to find that the reviewer is satisfactory with our current revised manuscript.

Reviewer #2

I thank the authors for the extra work they've done in response to questions raised by the reviews.
I'm happy with the author's clarification of homotopic shape evolution as referring to the changes during the evolution of the filtration.

I'm also glad that the authors have cited an additional mathematical paper establishing properties of the persistent Laplacian [36, Memoli et al]. They have also expanded the description of persistent Laplacian in the Methods section (p.22).

Response: We thank for the reviewer for carefully going through the revised manuscript.

On the issue of "Can one hear the shape of a drum?" (p.7) It is known that the answer - in a strict sense - is "no"; see Buser, Conway, Doyle, Semmler "Some planar isospectral domains" (1994). But, on the other hand, there is clearly much information about the topology and geometry of a domain encoded in the Laplacian. I would recommend rephrasing the sentence on p.7 to say that the PST "provides another illustration of the famous question", rather than claiming it "answers" this question.

Response: We agree with this suggestion, and we have made the revision accordingly.

In the discussion (p.14) there is now further explanation that most of the information content of the PST features is contained in the non-harmonic spectra of the dimension-0 Laplacian, and that the dimensions-1 and -2 features are deliberately kept brief to reduce overfitting.

Response: We thank the reviewer for carefully going over the revised manuscript.

In the rebuttal the authors state: "We have a different view about persistence barcode stability. In molecular interactions, small length bars are important in our element-specific topological representation. The small bar length differences correspond to the differences in interaction strengths."

I whole-heartedly agree that small persistence does not always imply "insignificant" (this is a common mis-interpretation of persistent homology). However, on p.23 in the methods section the authors state that for dimension-0 "the short scales below 2 angstroms are excluded" and for dimensions-1, -2, "bars with lengths lower than 0.1 angstroms are excluded". So I'm a little confused here about what the underlying principle is for interpreting small changes in persistence in their model?

Response:

We thank the reviewer for picking up this seemingly discrepancy. A more precise statement is that *Betti-0 bars* with different lengths correspond to the differences in interaction strengths. The Betti-0 bars shorter than 2 angstroms are excluded since they are related to bond-breaking events which are biologically unrelated to the systems we studied. Small Betti-1 and Betti-2 bars can be used to detect the packing density differences of different materials. However, for the same type of biomolecules in the present system, the small Betti-1 and Betti-2 bars can be excluded since they do not have a clear physical interpretation.

In Supplementary Note 5, and Figures 24, 25, 26, the authors summarise results from extra numerical computations conducted in response to the review. These show that using persistence landscapes versus their elementary statistical summaries do not significantly change the performance of their models. Similarly, adding p-persistent Laplacian information does not change any of their conclusions.

Response: We thank the reviewer for carefully going over the revised manuscript.

Small points:

page 2 line (-2) (i.e., filtration) -> (e.g., filtration)

page 2 line (-1) workhouse -> workhorse

page 7 line 1 second use of the word "harmonic" should be non-harmonic.

Response: Thank you for pointing out. We have corrected them, and carefully proofread the manuscript.

Reviewer #3

The authors have provided extensive responses and made corresponding revisions to the reviewers' questions and comments. I am generally satisfied with these revisions and responses.

Response: We sincerely thank the valuable suggestions and comments from the reviewer. We are pleasure to find the reviewer is satisfactory with our current revised manuscript.

Final Decision Letter:

Dear Professor Wei,

We are pleased to inform you that your Article "Persistent spectral theory-guided protein engineering" has now been accepted for publication in Nature Computational Science.

Once your manuscript is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

Please note that Nature Computational Science is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see https://www.nature.com/natcomputsci/for-authors). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

Before your manuscript is typeset, we will edit the text to ensure it is intelligible to our wide readership and conforms to house style. We look particularly carefully at the titles of all papers to ensure that they are relatively brief and understandable.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

If you have queries at any point during the production process then please contact the production team at rjsproduction@springernature.com. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Content is published online weekly on Mondays and Thursdays, and the embargo is set at 16:00 London time (GMT)/11:00 am US Eastern time (EST) on the day of publication. If you need to know the exact publication date or when the news embargo will be lifted, please contact our press office after you have submitted your proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number NATCOMPUTSCI-22-0836B and the name of the journal, which they will need when they contact our office.

About one week before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Computational Science. Our Press Office will contact you closer to the time of publication, but if you or your Press Office have any inquiries in the meantime, please contact press@nature.com.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Computational Science as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). We also welcome suggestions for the Hero Image, which appears at the top of our home page; these should be 72 dpi at 1400 x 400 pixels in JPEG format. Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Computational Science logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

We look forward to publishing your paper.

Best,
Fernando

--
Fernando Chirigati, PhD
Chief Editor, Nature Computational Science
Nature Portfolio

P.S. Click on the following link if you would like to recommend Nature Computational Science to your librarian: https://www.springernature.com/gp/librarians/recommend-to-your-library

** Visit the Springer Nature Editorial and Publishing website at www.springernature.com/editorial-and-publishing-jobs for more information about our career opportunities. If you have any questions please click here.**