Peer Review File

Manuscript Title: A backbone-centered energy function of neural networks for protein design

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1

One of the key challenges in de novo protein design is building models of the protein backbone that are inherently designable and can serve as the input into sequence optimization protocols. In this study a low-resolution force field that is sequence agnostic is developed that can be used to refine protein backbones so that good hydrogen bonding is formed within secondary structural elements and places the secondary structural elements appropriate distances apart. The force field is a linear combination of statistical terms (modeled with neural networks) for modeling torsional preferences, backbone-backbone hydrogen bonding, and residue spacing. The authors show that the force field can be used in combination with dynamics simulations to refine unfavorable starting backbones and produce backbones that serve as good starting points for high resolution sequence design. Impressively, they use X-ray crystallography to show that the structures of six of their designs closely match the design models. Overall, I am impressed with the results and think this is a strong paper, but I don't think the methods presented here are that different from approaches previously used to design proteins de novo. Specific comments are provided below:

Strengths:

- De novo protein design continues to be challenging, and the high-resolution design of six separate proteins is an important achievement that indicates that the underlying methods are working well.

- One of the more novel aspects of this paper is to train neural networks (NN) to predict the favorability of observing different features in a protein, for instance the preference for seeing a particular backbone torsion angle. It is not uncommon for protein modeling packages to include terms of this nature, but typically this is done by placing features into bins and counting how favorable each bin is. The NN approach avoids the generation of explicit bins and may allow for a smoother more nuanced energy landscape. However, this paper does not provide a direct comparison with more traditional statistical approaches, so the possible advantages or disadvantages of the NN approach are not really explored.

Weaknesses:

- With the recent success of Alphafold in CASP there is a lot of excitement about the use of machine learning in protein folding and design. One of the most innovative aspects of Alphafold's approach in the latest CASP was the development of neural networks that work directly with protein coordinates and have been trained to energy minimize protein structures. What is conceptually exciting about this approach is that the Alphafold team did not need to break the energy function calculation up into different terms that are meant to represent different physical phenomena (such as torsional preferences, hydrogen bonding, van der Waals forces, solvation). This is an attractive approach because it may eliminate/minimize the double counting that effects most standard force fields, including the paper being reviewed here. As soon as protein energetics are separated into separate terms that are not truly independent (for instance torsional preferences depend on hydrogen bonding and van der Waals forces), then it is almost unavoidable that different terms in the energy function will be representing the same physical phenomena (and

therefore there will be double counting). When seeing the title for this paper that is being reviewed, I thought that perhaps it was going to use some of the new ML approaches being used by Alphafold to avoid the need for separate energy terms that are treated independently but are not really independent. This turns out to not be the case, and I think the approach developed here will have the same issues as other knowledge-based force fields that have been developed for protein folding. With that being said, that does not mean that the methods developed are not useful and effective, it just means that I do not think that they are as innovative as some of the other ML-based approaches that are currently being developed.

- It is standard in de novo design papers to use denaturation experiments to characterize the stability of the designed proteins and the cooperativity of protein unfolding. The paper would be improved by including data of this type.

Other comments:

- No details are provided about how the derivatives for the force field are calculated and transformed into the proper form for use in the Langevin equations. For the terms based on NNs are the derivatives based on backpropagation? In cases where the energy term is a function of dihedral angles or relative orientation, how are these converted into forces that push on individual atoms?

- Was anything done to prespecify the h-bond pattern in the beta-sheets? If it was not predefined, in the dynamics simulations did you see strands slide up and down to establish good h-bonds across the sheet? If this is happening it would be exciting to report this behavior as strand sliding can be challenging to achieve in simulations.

Referee #2

This manuscript proposes a new approach to protein design based on optimizing a continuous function dependent only on backbone atom coordinates to generate a putative protein structure whose precise sequence is then subsequently optimized using pre-existing approaches (namely SCUBA in this case). Unlike existing approaches which either use a parametric equation to describe a space of pre-defined helical structures or fragment-assembly-based approaches that rely on reusing known protein fragments, this approach in principle allows one to explore arbitrary backbone structures and then fills in the sequence. The paper is generally well-written (although some editing of the prose is necessary, particularly for the first parts of the manuscript) and the experimental validation is extensive and compelling (six solved crystal structures). Nonetheless, some issues need addressing, primarily in matters of clarity but also in terms of the results themselves.

1. The premise of this approach is that it allows one to design a much broader range of protein geometries than what is observed in nature. To more quantitatively assess this claim, it would be useful to know how far in sequence space the H2E4 and H4 designed proteins are far from natural proteins, to ensure that their model didn't simply recall a memorized data point from the training set. I did a quick check with designed protein H4A1R (PDB accession: 7DGU) using HHblits, and found a hit that essentially covers the whole of designed protein with 38% sequence identity (UniRef100_A0A1R1ID41). That is a fairly close hit. In particular, one can imagine using their sketching approach to simply find natural proteins matching that template, and then using SCUBA2 to fill in sequences with the same backbone geometry. How different/more useful is the new proposed approach relative to that baseline? That's presently unclear.

2. To be sure, the concern above is ameliorated by the authors' grafting example (EXTD-3), where they take a natural protein and incorporate a new fragment that makes the overall structure distinct from any naturally occurring one. This is a compelling demonstration of the value of their approach. Thus, point 1 above is more about quantifying the degree to which the H2E4 and H4 proteins are different from what's known and also about establishing stronger/more quantitative

baselines for assessing the proposed method.

3. A clear description of how their 'topological sketches' are made, particularly in the main text, is warranted (it is described in more detail in the supplement). What would be ideal is if the authors have some way of visualizing the sketching process, perhaps through a recorded screen capture of the sketching, so that the reader understands what is involved and to what degree of precision the sketch needs to be made, and how faithfully the minimized structures reproduce the sketch.

4. The authors refer to a neighbor counting-neural network (NC-NN) approach. This is not a broadly known method, certainly not to a general readership like that of Nature, and so its outlines need to be described in the main text. Right now it is literally only referred to by name. Given that this is fundamentally a (very interesting) method development paper, the method's description needs to be front and center. Very detailed mathematical aspects can be left to the supplement, but the core idea of the approach and how it works, conceptually, must go in the main text, otherwise the manuscript leaves the reader wondering the entire time what was exactly done.

5. On this point, I would like a clarification--the non-parametric kernels are already estimating the energies, with the NNs simply relearning what they've estimated. This raises the question of why even have the NN in the first place? Is it simply that the NN provides an analytic equation that can be minimized using stochastic dynamics? That is, there is no learning being done beyond what the statistical potential itself is encoding. The NN is only acting as an interpolator, is this correct? If so I would explain this directly, preferably in the main text and at a minimum in the supplement. Right now the reader is left guessing as to why the NN is really necessary.

6. Other, less central aspects, of the method are also vaguely described, for example: "Thus, SCUBA was extended to consider explicit sidechains in a manner that its outcomes remain insensitive to sidechain types. Specifically, we focused on a way to have arbitrary sequence variations cause only subtle changes to the optimized backbones." There is no further description of what this "way" is. (If this is the LVG approach described later, then a clearer connection needs to be made between this paragraph and the subsequent section, to explain that the "way" described above is indeed the LVG approach.)

7. Finally, I am very impressed that the authors chose to report the number of designs they tested and the fraction that succeeded (in the text and in Tables S5 and S6)--this degree of transparency is refreshing and makes the results far more informative for other researchers intent on utilizing this approach for their design. Kudos!

Referee #3

A Summary of the key results

The paper describes a method for computing an energy function for
protein backbones independent of their amino acid types.

The paper demonstrates that from a relatively simple sketch (e.g. derived from secondary structure packing description), optimization under the energy function in two stages can produce a plausible backbone that can then be sequenced using an existing design algorithm.

The SCUBA energy optimization is just one part of a complex process of designing proteins (building the sketch, closing the loops then optimizing under the energy function followed by sequencing, selection using Rosetta forward folding) but does result in moderate success in designed proteins.

B Originality and significance: if not novel, please include reference

The complete protocol is a novel way to achieve protein backbone structures based on a sketch,

but somewhat similar effects have been achieved using Rosetta in previous publications.

A bottom-up approach for the de novo design of functional proteins
C Yang, F Sesterhenn, J Bonet, E van Aalen, L Scheller, L Abriata, ... bioRxiv

Principles for designing ideal protein structures
N Koga, R Tatsumi-Koga, G Liu, R Xiao, TB Acton, GT Montelione, ... Nature 491, 222-227

C Data & methodology: validity of approach, quality of data, quality of presentation

Overall the method seems valid and justified.

I find it lacking in comparison and analysis of the performance of the novel part of the algorithm. It is asserted that the yield of the system is comparable to that of Rosetta, but this is not well-supported. It is hard to understand what contribution the energy function contributes to the overall success of the system.

The sampling of torsion angles seems to produce non-strands / distorted helices in the cartoon visualizations (Fig. S3 etc.), even though the sketch was known to be helix / strand, which exaggerates the benefit of using your energy function. It seems that using ideal strands / helices might be fairer? On the other
hand, it seems that the paper is claiming to be about the usefulness of this potential so it seems that these figures should be in the body of the paper, and there ought to be clearer metrics showing the improvement in the structures from using the potential vs the original sketch. For instance: packing metric, as well as a comparison with trying to just use ABACUS2 on the sketches without using SCUBA.

Computation requirements are not clear. How much compute is required for training the energy functions and how much computation is required to design the proteins (broken down by all the stages)?

In the introduction the definition of "high designability" is contradictory to that in the abstract and to my understanding. (It's defined here in terms of the "sequence", but AIUI is a property of a structure _independent_ of the sequence.)

Explain in more detail the reasoning for the LVG intermediate representation.

I would appreciate a table or figure (probably just SI) summarizing the numbers of decoys generated at each stage for each target. It was hard to keep track of in the text.

p. 29, "a majority": What does this mean?

l. 622: "parallel" not "in parallel".

E Overall I think the paper is interesting and demonstrates that a novel approach works, but I do have reservations about the impact of the work. It describes one piece of a complex system with little validation of the specific piece, and does not enable a new capability in protein design (function, scale, accuracy...).

F Suggested improvements

Experiments, data for possible revision

G References: appropriate credit to previous work?

N/A

H Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

The start of the second paragraph is very hard to make sense
of. Explain what you're doing clearly. "Conceptual framework" isn't a good start to a paragraph.

I found the explanation in 1.2 seemed unnecessarily hard to read.

Author Rebuttals to Initial Comments:

Point-by-point responses to reviewers' comments

We thank the expert reviewers for their insightful comments and constructive criticisms. We have revised our manuscript to address all their concerns.

Before providing our detailed responses point-by-point, we would like to mention that our revisions include new results of three *de novo* helical proteins to provide proper context and to strengthen our arguments. Each of these protein has been experimentally verified to fold into a novel architecture that, instead of being pre-defined, has been found through SCUBA-driven explorations of the structural space.

In our revised manuscript, all revisions from the previous version have been highlighted in red. In the following responses, sentences describing specific revisions have been underlined.

Referee #1 *"-One of the more novel aspects of this paper is to train neural networks (NN) to predict the favorability of observing different features in a protein, for instance the preference for seeing a particular backbone torsion angle. It is not uncommon for protein modeling packages to include terms of this nature, but typically this is done by placing features into bins and counting how favorable each bin is. The NN approach avoids the generation of explicit bins and may allow for a smoother more nuanced energy landscape. However, this paper does not provide a direct comparison with more traditional statistical approaches, so the possible advantages or disadvantages of the NN approach are not really explored*."

The reviewer was right that statistical terms were commonly used for protein modeling. However, the traditional statistical approaches based on placing features into bins suffer from a fundamental drawback: they work only for very low dimensional feature spaces. As the number of dimensions of the feature space increases, the number of bins increases exponentially unless one dramatically lowers the resolution of the bins. With more than a handful of dimensions, the majority of bins would become empty, and the remaining small fraction of non-empty bins would each contain only one or a few sample points (the well-known "curse of dimensionality" phenomenon).

The primary reason that we developed NC-NN is to be able to faithfully capture correlations in much higher dimensions than traditional statistical approaches. Smoother and more nuanced energy landscapes were also wanted (but this was only a secondary purpose). To obtain backbones that are useful for *de novo* protein sequence design, our backbone-centered statistical energy function needs to accurately retain correlations in high dimensional spaces.

Consider that, when using conventional binning approaches, the different dimensions must be considered separately, thereby losing inter-dimension correlations in the results. *Nota Bene*: our data obtained using the alternative NC-NN approach establish that these very correlations are indeed structurally informative.

In SCUBA, the most high-dimensional energy term is the through-space interaction between two backbone sites, **which is of 14 dimensions** (the six inter-site relative geometry variables plus eight backbone torsional angle variables surrounding the two sites). The only approaches capable of learning statistical energies in such high dimensional spaces at a useful resolution (i.e., to support atomistic backbone design) are those like NC-NN, which learn, directly, the joint distributions of all dimensions with high fidelity. In our revised paper, we added supplementary data (fig. S1) showing that the NC-NN site-pair energy indeed capture correlations in this high dimensional space in physically meaningful ways.

Of course, when considering low-dimensional energy terms, conventional binning approaches do work. But it bears mention that even in these straightforward use cases there is no disadvantage of our NC-NN approach compared to conventional approaches. Indeed, NC-NN can fully recover the results of traditional approaches while offering the additional advantages of i) being automatically continuous and ii) having analytical gradients.

In the revised manuscript, we have added sections to emphasize the important differences between NC-NN and conventional binning approaches, especially the capability of NC-NN to capture high-dimensional correlations.

Referee #1 *"With the recent success of Alphafold in CASP there is a lot of excitement about the use of machine learning in protein folding and design. One of the most innovative aspects of Alphafold's approach in the latest CASP was the development of neural networks that work directly with protein coordinates and have been trained to energy minimize protein structures. What is conceptually exciting about this approach is that the Alphafold team did not need to break the energy function calculation up into different terms that are meant to represent different physical phenomena (such as torsional preferences, hydrogen bonding, van der Waals forces, solvation). This is an attractive approach because it may eliminate/minimize the double counting that effects most standard force fields, including the paper being reviewed here. As soon as protein energetics are separated into separate terms that are not truly independent (for instance torsional preferences depend on hydrogen*

bonding and van der Waals forces), then it is almost unavoidable that different terms in the energy function will be representing the same physical phenomena (and therefore there will be double counting). When seeing the title for this paper that is being reviewed, I thought that perhaps it was going to use some of the new ML approaches being used by Alphafold to avoid the need for separate energy terms that are treated independently but are not really independent. This turns out to not be the case, and I think the approach developed here will have the same issues as other knowledge-based force fields that have been developed for protein folding. With that being said, that does not mean that the methods developed are not useful and effective, it just means that I do not think that they are as innovative as some of the other ML-based approaches that are currently being developed"

We would like to start by noting that up to now, ML-approaches like those used in AlphaFold have been far more successful in structure prediction than in protein design. Their effectiveness for protein design, especially for structure design, remain to be established with extensive experimental validations (such as those reported here based on our SCUBA+ABACUS2 approach).

In addition to the widespread media coverage for AlphaFold, perhaps one reason the reviewer focused on this comparison is because our originally submitted manuscript did not make it very clear that our NC-NN model has been developed to achieve exactly what was pointed out by the reviewer: treating correlated dimensions jointly as single terms as much as possible.

Briefly, in SCUBA the "phi-psi-5" energy terms take care of correlations between neighboring backbone conformations, the "site-pair" energy terms take care of correlations between inter-backbone packing and intra-backbone local conformations, etc. The situation with models like AlphaFold is very different: their pseudo-energy functions do not correspond to true physical interactions (nor do they purport to). Rather, they represent geometric constraints deduced from data such as multiple sequence alignments.

Admittedly and as pointed out by the reviewer, we are still using combinations of separate energy terms. So there could be some correlations missing between variables in different energy terms. However, this "limitation" would in principle exist in any type of model. The reason is that currently available protein structure data in the PDB are still very limited, at least when considered relative to the immensity of possible protein structures. Consequently, information contained in available data can likely only support quantification of correlations among a sufficiently localized set of variables. To model a system of a size beyond this localization size limit, combinations between separately treated parts become unavoidable—otherwise the model will be limited by overfitting of the incomplete data.

There are now ways to move beyond the current constraint caused by incomplete coverage of current PDB entries for structural spaces at larger protein sizes: we and others are designing and characterizing *de novo* proteins with novel architectures. Since submitting the previous version of our paper, we have applied our method to successfully design three

de novo proteins that fold into overall architectures unobserved in PDB. In these architectures, super-secondary structures are combined in new ways which have no observational evidence among current PDB structures. High resolution X-ray structures of the designed proteins have been determined and reported in our revised manuscript.

Please note that a hypothetical statistical model that extracts full correlations beyond the level of super-secondary structures combinations from current PDB would not be able to produce these novel proteins because of the overfitting limitations. From this example, we can see that for designing genuinely novel structures, treating the system as a combination of separately considered parts is not a bad thing: even though some (presumably unimportant) correlations would be missed, it can propose new combinations of parts as novel design results.

In the revised manuscript, we have added parts to present the three *de novo* proteins that fold into novel overall architectures, and discussed the implications of the new results for developing data-driven models of protein structures.

Referee #1 "*It is standard in de novo design papers to use denaturation experiments to characterize the stability of the designed proteins and the cooperativity of protein unfolding. The paper would be improved by including data of this type."*

Please kindly note that the *de novo* proteins successfully designed by ABACUS2 on SCUBA-optimized backbones usually exhibit very high thermal stability, which is similar to previous sequences selected with the ABACUS design program on given (natural) backbones or *de novo* proteins designed with other methods.

As examples, we present in the revised manuscript the temperature-dependent CD spectroscopies of two designed proteins (supplementary fig. S11), which indicate very small changes in secondary structure contents with the temperature increased from 20 to 95 ℃. As the current study focuses on the accuracy of the designed structures rather than on their folding/unfolding properties, more thorough characterization of the folding/unfolding behaviors of the designed proteins may constitute interesting future research topics.

Referee #1 *"No details are provided about how the derivatives for the force field are calculated and transformed into the proper form for use in the Langevin equations. For the terms based on NNs are the derivatives based on backpropagation? In cases where the energy term is a function of dihedral angles or relative orientation, how are these converted into forces that push on individual atoms?"*

We have described the calculation of derivatives in the revised manuscript. It is straightforward application of the chain rule for derivatives, just like back propagation, only that we need derivatives with respect to the network's input instead of with respect to the network's internal parameters. Eventually, derivatives with respect to angles and

orientations (which analytically depend on atomic Cartesian coordinates) are propagated on to the Cartesian coordinates also by using the chain rule.

Referee #1 *"Was anything done to prespecify the h-bond pattern in the beta-sheets? If it was not predefined, in the dynamics simulations did you see strands slide up and down to establish good h-bonds across the sheet? If this is happening it would be exciting to report this behavior as strand sliding can be challenging to achieve in simulations."*

Hydrogen bond restraints between beta-strands are not needed. That is, when backbones are sampled and optimized by SD at lower temperatures, the SCUBA energy terms (specifically the through-space backbone site pair interaction term) can automatically drive the formation and maintenance of inter-strand hydrogen bonds. At higher temperatures (and in simulated annealing), the inter-strand hydrogen bonds may break and reform (sometimes with shifted pairing), achieving the effects of strand sliding.

We added clarifications about this point in the revised manuscript.

Referee #2: *"1. The premise of this approach is that it allows one to design a much broader range of protein geometries than what is observed in nature. To more quantitatively assess this claim, it would be useful to know how far in sequence space the H2E4 and H4 designed proteins are far from natural proteins, to ensure that their model didn't simply recall a memorized data point from the training set. I did a quick check with designed protein H4A1R (PDB accession: 7DGU) using HHblits, and found a hit that essentially covers the whole of designed protein with 38% sequence identity (UniRef100_A0A1R1ID41). That is a fairly close hit. In particular, one can imagine using their sketching approach to simply find natural proteins matching that template, and then using SCUBA2 to fill in sequences with the same backbone geometry. How different/more useful is the new proposed approach relative to that baseline? That's presently unclear."*

To look at the differences between the H2E4 and H4 sequences and naturally occurring ones, we compared the sequences of these designed proteins with their structure neighbors in PDB. The sequence identities at structurally aligned positions ranged between 3.5% to 25%, with an average value of 14%. For comparisons, redesigned sequences on natural backbones using ABACUS2 have averaged sequence identities around 33% with native sequences. The latter value can serve as a baseline for comparing the sequences designed on SCUBA-optimized backbones with those on natural backbones.

Further, following the reviewer's suggestion, we have completed sequence neighbor searches based on HHblits to quantify the differences comparing our H2E4 and H4 sequences with natural sequences. We note that all hits were of insignificant E-values (from 0.42 to 72) despite the sometimes apparently high sequence identities. In addition, it bears

emphasis that none of the hits are contained in the training set of either SCUBA or ABACUS2. Thus simple recalling of memorized data points is in principle not possible.

We have added the database search results discussed above in the revised manuscript.

Referee #2: *"2. To be sure, the concern above is ameliorated by the authors' grafting example (EXTD-3), where they take a natural protein and incorporate a new fragment that makes the overall structure distinct from any naturally occurring one. This is a compelling demonstration of the value of their approach. Thus, point 1 above is more about quantifying the degree to which the H2E4 and H4 proteins are different from what's known and also about establishing stronger/more quantitative baselines for assessing the proposed method."*

As commented by the reviewer, the EXTD-3 example can already ameliorate the concern that our models might be simply recalling memorized data points. Please kindly note that in our revised manuscript, we have reported three additional *de novo* proteins of novel architectures. They represent even stronger evidence to bolster the conceptually central innovation of our study in designing broader protein geometries than what is observed in nature.

Referee #2: *"3. A clear description of how their 'topological sketches' are made, particularly in the main text, is warranted (it is described in more detail in the supplement). What would be ideal is if the authors have some way of visualizing the sketching process, perhaps through a recorded screen capture of the sketching, so that the reader understands what is involved and to what degree of precision the sketch needs to be made, and how faithfully the minimized structures reproduce the sketch."*

We have added supplementary fig. S5, which graphically illustrate the process of defining a sketch and constructing an initial structure according to it.

Referee #2: *"4. The authors refer to a neighbor counting-neural network (NC-NN) approach. This is not a broadly known method, certainly not to a general readership like that of Nature, and so its outlines need to be described in the main text. Right now it is literally only referred to by name. Given that this is fundamentally a (very interesting) method development paper, the method's description needs to be front and center. Very detailed mathematical aspects can be left to the supplement, but the core idea of the approach and how it works, conceptually, must go in the main text, otherwise the manuscript leaves the reader wondering the entire time what was exactly done."*

We have significantly extended our description of the NC-NN method in the main text following the reviewer's suggestion.

Referee #2: *"5. On this point, I would like a clarification--the non-parametric kernels are already estimating the energies, with the NNs simply relearning what they've estimated. This raises the question of why even have the NN in the first place? Is it simply that the NN provides an analytic equation that can be minimized using stochastic dynamics? That is, there is no learning being done beyond what the statistical potential itself is encoding. The NN is only acting as an interpolator, is this correct? If so I would explain this directly, preferably in the main text and at a minimum in the supplement. Right now the reader is left guessing as to why the NN is really necessary."*

The reviewer is right that the non-parametric kernels are already estimating the energies, and the main role of the NNs is to provide an analytical representation of the energy landscape. However, we want to point out that the kernel approach by itself has not actually finished the task of "encoding" the training data, because one always needs to go through the original training data to estimate the energy of a new configuration. **Only after the NN step, the task of "encoding" can be considered as finished**, because then energies can be efficiently estimated without referring to the original training data.

NNs are also **practically indispensable.** Without using the NN, at every step of conformation sampling/optimization, one would have to repeat the kernel estimations for thousands of energy terms, to estimate each term involving looking for neighbors in datasets containing up to millions of entries. This would be computationally too demanding to perform, let alone that the thus-obtained noisy energy values without gradients would almost certainly be of little value for guiding the sampling/optimization process.

In our revised description of NC-NN in the main text, we have made the above clearer.

Referee #2: *"6. Other, less central aspects, of the method are also vaguely described, for example: "Thus, SCUBA was extended to consider explicit sidechains in a manner that its outcomes remain insensitive to sidechain types. Specifically, we focused on a way to have arbitrary sequence variations cause only subtle changes to the optimized backbones." There is no further description of what this "way" is. (If this is the LVG approach described later, then a clearer connection needs to be made between this paragraph and the subsequent section, to explain that the "way" described above is indeed the LVG approach.)"*

We have revised in the main text the part discussing the sidechain type insensitivity of SCUBA. Related to this, we have also given the reasons for using the LVG sequences.

Referee #3: *"The SCUBA energy optimization is just one part of a complex process of designing proteins (building the sketch, closing the loops then optimizing under the energy function followed by sequencing, selection using Rosetta forward folding) but does result in moderate success in designed proteins."*

Please kindly note that in our computational process, all the structure changes, starting from the artificially built initial structures to the final structures which turned out to agree well with the experimental structures, have been driven by SCUBA. Based on this we can say that SCUBA is not merely a part, but rather the driving engine of our backbone design system.

Referee #3: *"The complete protocol is a novel way to achieve protein backbone structures based on a sketch, but somewhat similar effects have been achieved using Rosetta in previous publications."*

Please kindly note that a major conceptual innovation of our method relative to previous work is that our method does not require existing structures as building blocks (the references cited by Reviewer 3 also used fragment assembling). Thus, our method may enable extensive, unrestricted explorations of the backbone structure space.

In our revised manuscript, we presented additional designed proteins of novel architectures not as pre-defined sketches, but computationally found by using SCUBA to explore the packing space of small all-helical proteins. This illustrate the new capability of our method for backbone design with a concrete example.

Referee #3: *"I find it lacking in comparison and analysis of the performance of the novel part of the algorithm. It is asserted that the yield of the system is comparable to that of Rosetta, but this is not well-supported. It is hard to understand what contribution the energy function contributes to the overall success of the system.*

The sampling of torsion angles seems to produce non-strands / distorted helices in the cartoon visualizations (Fig. S3 etc.), even though the sketch was known to be helix / strand, which exaggerates the benefit of using your energy function. It seems that using ideal strands / helices might be fairer? On the other hand, it seems that the paper is claiming to be about the usefulness of this potential so it seems that these figures should be in the body of the paper, and there ought to be clearer metrics showing the improvement in the structures from using the potential vs the original sketch. For instance: packing metric, as well as a comparison with trying to just use ABACUS2 on the sketches without using SCUBA."

We would like to start by noting again that the structure changes from the artificially built initial structures to the final structures have been driven by SCUBA. We are certain that the initial structures before SCUBA optimization are not designable (for the three *de novo* proteins of novel helical architectures newly presented in our revised manuscript, the randomly placed initial helices are even more unphysical, having numerous inter-backbone steric clashes), and **using ABACUS2 to select sequences for these initial structures cannot lead to well-packed proteins**.

Just for completeness, we applied ABACUS2 on the initial structures of the H2E4 sketch. Simple computational analysis suggest that the resulting sequences cannot fold into the expected sketch: the per residue ABACUS2 energies were too high (0.82 ± 0.27 for the ABAUCS2 sequences selected on the initial structures, which should be compared with the values of -0.58 ± 0.08 for the backbones optimized under the LVG sequences, and of -1.00 ± 0.14 for the iteratively refined final backbones). For the same groups of protein models, the corresponding per-residue Rosetta energies calculated after structure relaxation are respectively -1.48 ± 0.24, -3.01 ± 0.1, and -3.30 ± 0.11. These energy values clearly indicate that sequences selected with the initial backbones cannot lead to well-packed folded structures, while those selected with the SCUBA-optimized backbones can.

We agree with the reviewer that idealized secondary structure elements could have been used in building the initial structures for H2E4 and H4. However, we specifically selected torsional angle sampling to exclude influences from predefining specific inter-strand hydrogen bonds. Our results demonstrate proof-of-concept that SCUBA can drive the automatic formation of regular secondary structure elements. This is a highly desired property for a backbone design energy function. The usefulness of this property can be found in our successful design of EXTD-3, in which the merging of two separate helices into the H3 helix of the final structure was not predefined, but rather automatically produced by SCUBA-driven energy optimization.

Finally, we would also like to point out that besides the automatic driving of the formations of some of the secondary structures, the sub-angstrom agreements between the SCUBA optimized and the experimentally determined loop structures represent strong evidence for the effectiveness and validity of SCUBA-driven optimization.

In our revised manuscript, we emphasized the role of SCUBA as a central engine to drive the structure change from undesignable to designable backbones. The above-mentioned energy comparisons between sequences designed on the initial backbones and those on the SCUBA-optimized backbones are presented in these discussions. We have also rewritten the questioned conclusion sentence summarizing the success rate of our designs relative to that previously reported by RosettaDesign.

Referee #3: *"Computation requirements are not clear. How much compute is required for training the energy functions and how much computation is required to design the proteins (broken down by all the stages)?"*

The computational costs for NC-NN training and SCUBA SD simulations are now reported in the revised manuscript. Although learning all the NC-NN models took several days on a multi-CPU workstation, this computational cost is irrelevant to most applications because it has already been done once and for all. Performing all the SCUBA optimization stages for a single backbone took two to three hours using a small number of computing cores (e.g. four cores) of a commonly available multi-core CPU. Thus the design of a few

tens of backbones can be finished in one day using a single CPU of dozens of cores. This amount of computation should not be much of a burden as compared with the computational costs of many other structure modeling tasks.

Referee #3: *"In the introduction the definition of "high designability" is contradictory to that in the abstract and to my understanding. (It's defined here in terms of the "sequence", but AIUI is a property of a structure _independent_ of the sequence.)"*

We have rewritten the sentence in question.

Referee #3: *"Explain in more detail the reasoning for the LVG intermediate representation."*

We have explained our reasoning in the revised manuscript. The LVG sequence simplification scheme has been chosen because it is relatively simple, the respective backbone conformational preferences of leucine and valine are consistent with the corresponding secondary structure types, and the relatively featureless and medium-sized sidechains of leucine and valine make them proper place holders for occupying the backbone interspaces in a backbone-centered model.

Referee #3: *"I would appreciate a table or figure (probably just SI) summarizing the numbers of decoys generated at each stage for each target. It was hard to keep track of in the text."*

A supplementary table (Table S5) has been provided to summarize the numbers of structures considered at different design stages.

Referee #3: *""p. 29, "a majority": What does this mean?*

l. 622: "parallel" not "in parallel""

The unclear word "majority" on P. 29 (original file) is now explained. The wrong use of "in parallel" at line 622 (original file) has been corrected.

Referee #3: *"Overall I think the paper is interesting and demonstrates that a novel approach works, but I do have reservations about the impact of the work. It describes one piece of a complex system with little validation of the specific piece, and does not enable a new capability in protein design (function, scale, accuracy...)"*

For the potential impact of our work, we wonder if the reviewer can consider the following two aspects. First, a major conceptual innovation of our method relative to previous work is that our method does not require existing structures as building blocks. It

can thus enable extensive, unrestricted explorations of the backbone structure space, as illustrated in our design of proteins of novel helical architectures.

Second, the SCUBA+ABACUS2 approach uses energy functions that have been **orthogonally developed to those used in RosettaDesign**, which is the only other extensively validated *de novo* design method. Given the potentially transformative and continuously challenging nature of *de novo* protein design, the impacts of our method will be significant—even in applications for which either RosettaDesign or our method can potentially work.

As for the validation of SCUBA, we believe the most compelling validating results presented in this work are the close agreements between the SCUBA-optimized backbones and the X-ray structures of the designed *de novo* proteins. The computational comparisons of designability between the initial and the SCUBA-optimized backbones discussed above constitute additional supporting data.

Besides these, our revised paper specifically contains the following data for the validation of SCUBA: i) individual NC-NN terms capturing the characteristic correlations in observed protein structures (supplementary fig. S1 shows this for the highest-dimensional SCUBA energy term in the revised manuscript. Other SCUBA terms have similar properties); ii) the deviations from native backbone structures in SCUBA-driven simulations of natural proteins being reasonably small (supplementary fig. S2); and iii) the close reproduction of natural-protein like backbone structures by SCUBA-driven optimizations from artificially constructed initial structures (supplementary fig. S4).

Referee #3: *"The start of the second paragraph is very hard to make sense of. Explain what you're doing clearly. "Conceptual framework" isn't a good start to a paragraph."*

We have rewritten the sentence in question.

Referee #3: *"I found the explanation in 1.2 seemed unnecessarily hard to read."*

To make section 1.2 easier to read, we have broken it into small subsections and inserted subsection titles to guide the readers.

Finally, we would like to thank the reviewers again for their valuable suggestions that greatly helped us to improve our manuscript.

Reviewer Reports on the First Revision:

Referee #1

The authors have provided thoughtful responses to my comments and have modified the manuscript accordingly. The sheer number of successful designs presented in this study is impressive and provides strong evidence that the underlying techniques are robust. The NN-based energy terms employed are novel in that they examine multi-dimensional features that would not be approachable with more traditional statistical approaches. The methods developed here are of sufficient novelty and usefulness to warrant publication in Nature.

Referee #2

The authors have addressed all my concerns, with the newly designed and solved structures further demonstrating that they can generate novel structural motifs.

Referee #3

A Summary of the key results

The paper describes a method for computing an energy function for protein backbones independent of their amino acid types.

The paper demonstrates that from a relatively simple sketch (e.g. derived from secondary structure packing description), optimization under the energy function in two stages can produce a plausible backbone that can then be sequenced using an existing design algorithm.

The SCUBA energy optimization is just one part of a complex process of designing proteins (building the sketch, closing the loops then optimizing under the energy function followed by sequencing, selection using Rosetta forward folding) but does result in moderate success in designed proteins.

The revision has improved the paper in several ways:
adding more structures, stability data and generally addressing the points raised by the reviewers.

B Originality and significance: if not novel, please include reference

The complete protocol is a novel way to achieve protein backbone structures based on a sketch, but somewhat similar effects have been achieved using Rosetta in previous publications.

The paper describes the construction of three topologies, including inserting a novel piece in an existing protein. The designs are experimentally validated with a moderate proportion of the designs succeeding. This protocol does seem to allow a fairly general sketch to be input to the process, so it would be interesting to see in silico demonstrations of more varied sketches being used to drive designs.

The paper does not demonstrate proteins with function.

AIUI, the "neighbour counting neural network", is just a simple fully-connected multi-layer perceptrons regressing a target which is estimated by a kernel function. Giving it a new label gives a misleading semblance of originality but is confusing.

However, I find Fig. 1d confusing. You have lines from e(Q^probe) going to both input and output of the neural network. Fig. 1c is also confusing. Why are you taking counts from four separate regions into the box in 1d? I understand from the formula that you should be comparing the number of grey points within one of probe circles to the number of 'empty' points. Why are the regions of different diameters? It would also be helpful to label the space that the points lie in in the figure (as you mention in line 150).

"the SCUBA term for through-space interaction between two backbone sites is of 14 dimensions": I

think you mean "is a function of 14 dimensions"; the term itself is just a scalar?

The basic method of learning a set of statistical energy terms is similar to that used in Rosetta, except here the energy terms are learned with simple MLP, and most of those terms are chosen to be independent of the residue type. While the authors mention gradients, it's not explicitly said how they are actually used.

C Data & methodology: validity of approach, quality of data, quality of presentation

The machine learning part still feels a bit unclear. (Beyond the points raised earlier).

There is no description of the loss used, nor the optimization algorithm to train the networks, nor measures of training/test accuracy for the functions being fitted. Maybe add the number of actual inputs to each network to S1.

How many training points are derived from the 12000 training proteins? How many "reference distribution" samples were used?
How is the 'non-redundant' set curated? How many parameters do the neural networks have?

As discretized probability distributions, S7 & S8 should be plotted as histograms not line curves.

Can you add natural monomer (or just the training set) proteins' distributions to S3?

l 1011: "random" isn't very helpful without knowing the distributions.

Saying SCUBA is an "all-atom effective energy function" seems to contradict the "side-chain unknown" principle.

E Conclusions: robustness, validity, reliability

Overall I think the paper is interesting and demonstrates that a novel approach works, and I am more convinced than before of its use (though I think this case would be strengthened by showing more sketch variety even if only in silico).

The authors' points about SCUBA driving the process don't all seem valid to me. While the initial sketches may be unphysical, that does not preclude designing to them (though it does mean that the ABACUS2 energies calculated for them are not comparable to those for the structures relaxed with SCUBA). The demonstration of the rosetta relaxed energies does get your point across, and seems worth putting in the paper, but perhaps at larger scale (and with an indication of the N on which you're computing these values).

F Suggested improvements, experiments, data for possible revision

More detail of the neural network part.

G References: appropriate credit to previous work?

H Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

Clarity could be improved. There are some awkward wordings that should be fixed.

Author Rebuttals to First Revision:

Point-by-point responses to comments of reviewer #3

Comment 1. "*...This protocol does seem to allow a fairly general sketch to be input to the process, so it would be interesting to see in silico demonstrations of more varied sketches being used to drive designs.*"

RESPONSE: We greatly appreciate the reviewer's suggestion. We have now completed ABACUS2 sequence design on all the 10 topological sketches presented in Extended Data Figure 4. For each sketch, 10 initial backbones and the same number of SCUBA-optimized backbones have been used. We considered 10 ABACUS2-selected sequences for each backbone. This led to in total 2000 sequences for 200 different backbones of various sketches. Each sequence has been evaluated *in silico* by calculating the per-residue ABACUS2 energy, as well as the per-residue Rosetta energy after structure relaxation. The results are given in the following table, which is also presented as Extended Data Figure 5**c** in the revised manuscript.

ABACUS2 and Rosetta energies of ABACUS2-selected sequences for initial and SCUBA-optimized backbones of different topological architectures.

	Architecture*	1	2	3	4	5
ABACUS2 energy per residue†	Initial	1.49	1.05	1.02	1.01	0.89
	Optimized	0.27†	-0.31	-0.13	-0.62	-0.70
Rosetta energy per residue†	Initial	-0.98	-1.42	-1.43	-1.51	-1.54
	Optimized	-2.65	-2.76	-2.75	-3.05	-3.17

	Architecture*	6	7	8	9	10
ABACUS2 energy per residue†	Initial	0.82	1.01	0.82	0.96	1.33
	Optimized	-0.68	-0.71	-0.58	-0.63	-0.87
Rosetta energy per residue†	Initial	-1.47	-1.46	-1.51	-1.74	-1.90
	Optimized	-3.05	-3.22	-3.05	-3.02	-3.17

*The secondary structure compositions of the architectures are: 1: $E_{10}E_{10}E_{10}E_{10}$, 2: $E_7H_{16}E_7E_7$, 3: $E_7E_7H_{16}E_7$, 4: $E_7H_{16}E_7E_7H_{16}E_7$, 5: $E_{10}H_{20}E_{10}H_{20}E_{10}E_{10}$, 6: $E_7H_{16}E_7H_{16}E_7E_7$, 7: $E_{10}H_{20}E_{10}E_{10}H_{20}E_{10}$, 8: $E_7H_{16}E_7E_7H_{16}E_7$, 9: $H_{15}H_{15}H_{15}$, 10: $H_{21}H_{21}H_{21}H_{21}$. Sketch 10 led to optimized backbones of either left-handed or right-handed twists, as shown in the two boxes presented in Extended Data Figure 4.

†Each energy value has been averaged over 100 sequences selected on a group of 10 initial or optimized backbones (10 sequences selected using ABACUS2 for each backbone), with standard deviation between 0.08 and 0.38. Rosetta energies have been calculated on relaxed structures with selected sequences. The averaged ABACUS2 energy for the optimized sketch 1 backbones is higher than those for the other sketches, as sketch 1 is a plain 4-strand β sheet without a hydrophobic core.

We have also now calculated the per-residue Rosetta energies of the experimentally-examined ABACUS2 sequences designed for the 13 novel all-helical backbones (Extended Data Figure 10). The average value is -3.32±0.07 for the three sequences that have been experimentally verified to be soluble monomers; for the remaining ten sequences the average value is -3.22±0.14.

These *in silico* results show that while the initial backbones of the various sketches are clearly not designable (*i.e.*, sequences designed based on these initial backbones are of high energies), the backbones generated by SCUBA-driven optimization have significantly improved designability, as indicated by their much lower energies compared with the initial backbones.

Besides the above results regarding SCUBA-designed structures, Extended Data Figure 2**b** shows simulation results indicating that natural protein structures are stable when modeled with SCUBA; for example, the average backbone RMSD of structures sampled in the SCUBA simulations from examined native structures of natural proteins are below 2 Å for 26 of 33 proteins of various fold types. The designed and simulated structures together covered structure elements of a wide spectrum of geometries, including sheets formed by parallel, anti-parallel, or mixed-direction strands; helices running in parallel or anti-parallel directions and packed in left-hand or right-hand twisted forms (the novel all-helical proteins exhibit even more diversely packed helices), and packing between helices and strands in varied geometries; and so on. Thus our *in silico* calculations have tested the applicability of SCUBA in a wide range of structural contexts.

We also want to note that without experimental evidence, the *in silico* computed energies themselves may not be sufficient for objectively judging the true physical plausibility of designed proteins. This point is exemplified by the novel all-helical designs, for which sequences of similarly low Rosetta energies (-3.32 $\pm$ 0.07 versus -3.22 $\pm$ 0.14) led to distinctively different experimental outcomes. For this reason, we have relied more on the agreements between the designed and the experimentally solved structures including loops to support the validity of our computational model.

Comment 2. *"The paper does not demonstrate proteins with function."*

RESPONSE: Please kindly note that function was not a primary goal of the current study. Rather, it was to design new architectures/topologies. We focused on this goal because—relative to previous work—a major conceptual innovation of our method is that our method does not require existing structures as building blocks to create new structures. Our method can thus enable extensive explorations of the backbone structure space. The proteins of novel architectures designed here demonstrate the value of our approach in designing broader protein geometries than what is observed in nature.

We agree with the reviewer that the design of proteins with functions is an important topic, which has additional challenges besides those faced by structure-focused design. In particular, protein function in many situations is highly dependent on both the structure and the specific amino acid sequence (as can be seen from the fact that most reported studies of designing proteins of functions have used experiments to extensively screen large numbers of "structure and sequence" variants. For example, a recent study published in *Science* used 100,000 designed variants to screen for SARS-CoV-2 inhibitors (doi: 10.1126/science.abd9909)). How to harness our increasing ability to robustly design new structures for the design of proteins with function constitutes the major topic to be investigated in future studies.

That said, we are confident that our approach has real potential to advance the design of proteins with function for (at least) the following two reasons:

—The first one is technical: our method allows backbones to be designed not by assembling existing fragments, but, rather, based on stochastic dynamics simulations, in which the structure varies continuously over large ranges according to an analytical energy function. In this process, function-related structure restraints can be implemented in a straightforward way.

—The second reason is more general: the SCUBA+ABACUS2 approach used in our study employs energy functions that have been developed in a fashion orthogonal to those employed by RosettaDesign, which is the only other general and extensively validated *de novo* protein design method. Thus, and given the still-challenging nature of *de novo* protein design (as also pointed out by reviewer #1), its low success rates, and the constant need for selection among large numbers of design variants, our method represents a fresh and methodologically distinct approach that is suitable for a wide range of protein design applications, including those where either RosettaDesign or our method can potentially work.

Comment 3. *"the "neighbour counting neural network", is just a simple fully-connected multi-layer perceptrons regressing a target which is estimated by a kernel function. Giving it a new label gives a misleading semblance of originality but is confusing."*

RESPONSE: The reviewer is certainly correct that the neural networks themselves are mathematically just fully connected multi-layer perceptrons; we thank the reviewer for pointing out this potential point of confusion. In our originally submitted manuscript, we used "NC-NN" to refer to the learning process, which comprises i) the neighbor-counting step, followed by ii) the neural network training step; that is, we were not referring to the neural network's mathematical structure. Now, guided by this comment, we have realized that readers could be confused by our content and think that we had meant the latter.

To avoid confusion, in our revised manuscript, we now explained in the introduction section of the main text that NC-NN denotes:

> "…a two-step process of first estimating statistical energy values from raw structure data by kernel-based methods (*i.e.*, neighbor counting), followed by training of neural networks (which are fully-connected three-layer perceptrons, see Fig. 1**d** and the Supplementary Methods) to represent the potentials."

Given admonition from the Editor regarding the need to retain comprehensibility for a general readership, our revised manuscript still needs a short-hand name to refer to the entire (*i.e.*, two-step) approach; we trust that the reviewer can agree that the terms "neighbor counting" and "neural network" are likely more friendly to non-expert readers than the technically more accurate terms "kernel-based density estimation" and "perceptron network" to represent the respective steps, and it is reasonable to put "neighbor counting" and "neural network" together for a short-hand name of this two-step process. Please note that the general term "neural

networks" was also used to refer to "perceptrons" in the first study to introduce the use of fully-connected multi-layer perceptrons to represent high-dimensional energy surfaces (doi:10.1103/PhysRevLett.98. 146401, cited as Ref 29 in our paper). This guided our word choice when preparing the originally submitted manuscript. Now, although we have retained the short acronym "NC-NN", please note that we have now changed expressions like "NC-NN energy terms" to "NC-NN-learned energy terms" to avoid possible confusion. We have also used the more specific term "fully-connected three-layer perceptron neural network" in place of the general term "neural network" in the Supplementary Methods when it is needed for unambiguity.

Comment 4. "*However, I find Fig. 1d confusing. You have lines from e(Q^probe) going to both input and output of the neural network. Fig. 1c is also confusing. Why are you taking counts from four separate regions into the box in 1d? I understand from the formula that you should be comparing the number of grey points within one of probe circles to the number of 'empty' points. Why are the regions of different diameters? It would also be helpful to label the space that the points lie in in the figure (as you mention in line 150).*"

RESPONSE: Again, we thank the reviewer for pointing out these points of confusion.

We have changed Fig. 1**d** to place the lines going to (from) the network input (output) more precisely to indicate that only "Q^probe" goes to the input, while the output goes only to "e".

The drawing of more than one regions in panel 1**c** is to indicate that estimations at different Q^probe points are collectively needed for training the network. To avoid possible confusion that counts from different regions are used for a single estimation, the box holding the energy estimation formula in panel 1**d** has now been changed from one to a stack of several boxes to indicate that neighbor counting within different circles in panel 1**c** are used to estimate the energies at different Q^probe points.

The regions in Figure 1**c** are of different diameters to indicate that different radii of the neighbor counting kernel can be chosen at different Q^probe points. This is useful because this radius affects the trade-off between resolution and statistical uncertainty: while a smaller radius can lead to higher resolution of the model, it also leads to fewer training points counted as neighbors, which results in increased statistical uncertainty. Choosing the radius of the kernel adaptively instead of using a fixed value for all Q^probe points helps to better balance this trade-off.

We have now clarified this in the revised main text:

> "The radius of the kernel can be chosen adaptively based on the local density of the training data, so that the trade-off between resolution and statistical uncertainty of the estimated energies can be balanced."

And in the legend of Figure 1:

> "More than one neighbor-counting circles are drawn in **c** to indicate that single-point energies of many different probing Q points are to be estimated (separately) by NC. These circles are of different diameters to indicate that the radii of the neighbor-counting kernels can be chosen adaptively according to the distributions of the NC training data."

At the end of Supplementary Methods section 1.2.3, we further explained:

> "If necessary, the similarity criteria (or the radius of the kernel) in formula S6 can be adaptively chosen, being stricter for $\boldsymbol{QQ}^{pppppppppp}$ in regions densely populated by training data, and being more relaxed for $\boldsymbol{QQ}^{pppppppppp}$ in sparsely populated regions. Because a smaller radius of the kernel function leads to higher resolution of the resulting model (which is accompanied by increased statistical uncertainty, owing to fewer points being counted as neighbors), the adaptive kernel can balance this trade-off between resolution and statistical uncertainty according to the local distributions of the training data."

We have labeled the box in Figure 1**c** as "$\boldsymbol{Q}$ space".

Comment 5: *"the SCUBA term for through-space interaction between two backbone sites is of 14 dimensions": I think you mean "is a function of 14 dimensions"; the term itself is just a scalar?"*

RESPONSE: We thank the reviewer for helping us to correct this. The term is indeed a scalar function of 14 variables. We have revised the respective sentence in the main text:

> "For example, Extended Data Fig. 1 shows low-dimensional projections of a NC-NN-learned SCUBA energy term depending on 14 variables."

Comment 6: *"The basic method of learning a set of statistical energy terms is similar to that used in Rosetta, except here the energy terms are learned with simple MLP, and most of those terms are chosen to be independent of the residue type. While the authors mention gradients, it's not explicitly said how they are actually used.*

RESPONSE: The reviewer is definitely right that statistical energy terms have been widely used in protein structure modeling, even in models that mainly rely on physics-based energy functions, such as Rosetta.

On the other hand, we would also like to bring to notice that statistical terms used in Rosetta energy function, as other continuous statistical energy terms built by conventional approaches, have each considered only one or two structure variables. Thus they would not be able to correctly describe higher-order correlations.

Even though a fully-connected MLP is much simpler as compared with sophisticated networks such as deep Convolutional Neural Networks or Transformers, it can represent high-dimensional energy surfaces well, as has been shown in a study for complex energy surfaces computed using quantum mechanics (doi:10.1103/PhysRevLett.98.146401, cited as Ref 29 in our

paper). Indeed, it is notable that the MLP terms used here are far more sophisticated than conventional statistical energy terms in order to faithfully represent high-order correlations.

Another point is that the small sizes of the networks used here (relative to more complicated networks) makes them computationally inexpensive, suitable for use in structure sampling and optimization, in which thousands of MLP terms need to be evaluated hundreds of thousands of times or more.

We thank the reviewer for the suggestion to clarify how gradients are used. We now provide an explanation of this in the main text:

> "As the negative gradients of the energy function (with respect to coordinates) can be used as forces, SCUBA can be used to drive Newtonian or Langevin molecular dynamics (MD or SD) simulations of protein structures."

Comment 7: *"There is no description of the loss used, nor the optimization algorithm to train the networks, nor measures of training/test accuracy for the functions being fitted. Maybe add the number of actual inputs to each network to S1."*

We have added a new section 1.2.5 "Details of learning the three-layer perceptron neural networks from data" in the Supplementary Methods, which includes the following content:

> "For each statistical energy term, the loss was the sum of squares for error. The optimization algorithm used was the momentum optimizer, as implemented in the TensorFlow machine learning package[35]. The training/test accuracy were measured by the squares of error averaged over the training/test data points."

We have also added the number of input nodes in Extended Data Figure 2**a** (Table S1 in the previous version).

Comment 8: *"How many training points are derived from the 12000 training proteins? How many "reference distribution" samples were used?"*

RESPONSE: We have now put the information in the newly added Supplementary Methods section 1.2.5. "Details of the learning the three-layer perceptron neural networks from data", which contains information for each of the energy terms:

> "For $e^{\varphi\text{-}\psi\text{-}1}_{NC-NN}$:
>
> a) Observed data points for NC: 1,252,924 backbone positions in the training proteins that are not in helices and strands.
>
> b) Reference data points for NC: not needed for a uniform distribution.
>
> c) Probing points for training NN: 32,400 uniformly distributed points on the 2D plane.

d) NN attributes: 2 input dimensions encoded by 12 input-layer nodes, 16 middle-layer nodes, and 225 parameters. The mean square fitting error is 0.03.

For $e^{\varphi\text{-}\psi\text{-}5}_{NC-NN}$:

a) Observed data points for NC: 1,722,251 five-residue segments in the training proteins.

b) Reference data points for NC: independent combinations of separately drawn points from one 1-D distributions for ψ, three 2-D distributions for (φ, ψ), and one 1-D distributions for φ. For each distribution 50,000 points were drawn.

c) Probing points for training NN: 165,000 observed points from training data and 165,000 points drawn from the reference distribution.

d) NN attributes: 8 input dimensions encoded by 64 input-layer nodes, 72 middle-layer nodes, and 4,753 parameters. The mean square fitting error is 0.52.

For $e^{site-pair}_{NC-NN}$:

a) Observed data points for NC: 27,590,092 backbone pairs with inter Cα distance below 12.1 Å in the training proteins.

b) Reference data points for NC: independent combinations of separately drawn points from one 6-D distribution for the inter-site translations and rotations, and two 4-D distributions, each for the four (φ, ψ) angles surrounding one site. The number of reference points covering the 6-D space is the same as the observed data points. 2,805,520 observed, four-residue segments served as reference points for the 4-D distributions.

c) Probing points for training NN: 95,000 observed points from training data and 176,000 points drawn from the reference distribution.

d) NN attributes: 14 input dimensions encoded by 191 input-layer nodes (the 6-D translation and rotations are first surrogated by inter-atomic distances and then encoded), 64 middle-layer nodes, and 12,353 parameters. The mean square fitting error is 0.60.

For $e^{local\text{-}HB}_{NC-NN}$:

a) Observed data points for NC: 767,524 pairs of three-atom units from the training proteins, with the two units separated by two to four residues along the primary sequence. and O-N distance blow 5.5 Å.

b) Reference data points for NC: 7,675,240 pairs of three-atom units with uniformly distributed relative translations and rotations, and with O-N distance below 5.5 Å.

c) Probing points for training NN: 760,000 observed points from training data and 76,000 points drawn from the reference distribution. Here the number of probing points drawn from the reference distributions has been intentionally chosen to be one order of magnitude less than the probing points taken from the observed data points, so that the training could focus on fitting the lower energy regions (populated

by the probing points taken from the observed data).

d) NN attributes: 5 input dimensions encoded by 35 input-layer nodes, 32 middle-layer nodes, and 1,185 parameters. The mean square fitting error is 1.13 (the variation range of the energy is about 23 in arbitrary unit).

For $e_{NC-NN}^{rotamer}$

a) Observed data points for NC: $4◊10^4$ to $2.8◊10^5$ points from training structures for various sidechain types.

b) Reference data points for NC: $2◊10^5$ points with backbone φ-ψ angles randomly drawn from observed data combined independently with sidechain torsional angle(s) randomly drawn from uniform distributions.

c) Probing points for training NN: $4◊10^4$ to $2◊10^5$ observed points from training data, and the same numbers of points drawn from respective reference distributions.

d) NN attributes: the number of input dimensions is (2 + *number_of_flexible_sidechain_torsional_angles*), each torsional angle is encoded with 6 input-layer nodes. The number of middle-layer nodes is 24. The number of parameters is 1 + 24◊(6◊*number_of_input_dimensions* + 2). The mean square fitting errors are around 0.1 or smaller for various sidechain types.

We note that the NN fitting errors listed above are the training errors. Estimated using data points that had not been used for training, the test errors were similar to the training errors if the test data were composed of the same fractions of observed versus computationally-drawn probing points as the training data. The test errors were smaller than the training errors if the test data were composed solely of observed probing data points. This is understandable because the NC-estimated training energies themselves are noisy. Compared with the computationally drawn probing points, the observed data as probing points are more enriched in regions of higher probability densities, and their NC energies suffer less from statistical uncertainty. Thus, in regions where the observed data points are located, the perceptron models are statistically more accurate than the levels indicated by the overall fitting errors. This is a desired property in applications, because it is the accuracy in these regions that determines the quality of the optimized backbones."

Comment 9: *"How is the 'non-redundant' set curated? How many parameters do the neural networks have?"*

RESPONSE: We have now explained in the Supplementary Methods that the non-redundant set included natural protein structures determined at resolutions of 2.5 Å or above by X-ray crystallography (and with pair-wise mutual sequence identities below 50%). The same protein set had been used to train the ABACUS and ABACUS2 sequence selection models.

The number of parameters of the networks are now given in the newly added Supplementary Methods section 1.2.5, the contents of which are provided in responses to Comments 7 and 8.

Comment 10: *"As discretized probability distributions, S7 & S8 should be plotted as histograms not line curves."*

RESPONSE: We have changed the graphs (now in Extended Data Figure 6) into histograms.

Comment 11: *"Can you add natural monomer (or just the training set) proteins' distributions to S3?"*

RESPONSE: Yes. We have added the distributions in the graph (now Extended Data Figure 5**d**)

Comment 12: *"1011: "random" isn't very helpful without knowing the distributions."*

RESPONSE: We now explained in the Supplementary Methods that these are uniform distributions:

> "Six peptide backbone segments…were placed at uniformly-distributed random positions and in uniformly-distributed random orientations in the same spatial region, their centers of geometries fall within a sphere of radius 15 Å."

Comment 13: *"Saying SCUBA is an "all-atom effective energy function" seems to contradict the "side-chain unknown" principle."*

RESPONSE: We thank the reviewer for pointing out this potential point of confusion. We have accordingly changed the sentence in question by describing SCUBA as:

> "…an effective energy function that analytically depends on atomic Cartesian coordinates."

Comment 14: *"The authors' points about SCUBA driving the process don't all seem valid to me. While the initial sketches may be unphysical, that does not preclude designing to them (though it does mean that the ABACUS2 energies calculated for them are not comparable to those for the structures relaxed with SCUBA). The demonstration of the rosetta relaxed energies does get your point across, and seems worth putting in the paper, but perhaps at larger scale (and with an indication of the N on which you're computing these values)."*

RESPONSE: We have now carried out ABACUS2 sequence selections on 100 initial structures and 100 SCUBA optimized backbones for 10 different sketches. We have added an Extended Data item (Figure 5**c**) to show the corresponding energies (note that we provide details related to this in our response to Comment 1).

Comment 15: *"More detail of the neural network part."*

RESPONSE: We thank the reviewer for his suggestions regarding this issue. We have now provided details in Supplementary Methods section 1.2.5 as described in our responses to Comments 7 and 8.

Comment 16: *"Clarity could be improved. There are some awkward wordings that should be fixed."*

RESPONSE: We have significantly edited the paper for conciseness and clarity without changing its scientific contents. As directed by the Editor, we have substantially reorganized and shortened the summary and main text to be within the 3,000 word limit for a *Nature* article. Figures and tables (including previous supplementary figures and tables) have been rearranged into 4 main figures plus 10 Extended Data items. Please kindly note that we have revised the title of our paper into "A backbone-centered energy function of neural networks for protein design".

We thank the reviewer again for the insightful comments and constructive criticisms, which have greatly helped us to improve our study and our paper.

A list of studies cited in this response:

Cao L, Goreshnik I, Coventry B, et al. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. Science **370**, 6515, doi: 10.1126/science.abd9909 (2020).

Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. Physical Review Letters **98**, 146401, doi:10.1103/PhysRevLett.98.146401 (2007).

Reviewer Reports on the Second Revision:

Referee #3

I'd like to thank the authors for making a number of significant improvements to the paper and for addressing my comments.