nature portfolio

Peer Review File

Protein Sequence Modelling with Bayesian Flow Networks

Corresponding Author: Dr Thomas Barrett

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This paper lies at the intersection of machine learning and biology. Specifically, it adopts a recently proposed generative framework based on a Bayesian Flow Network and trains a large model with 650 million parameters on protein sequences. According to various computational and biological metrics, the model effectively covers known protein types while generating reasonable and diverse novel proteins, significantly outperforming autoregressive and discrete diffusion models. Furthermore, the paper fine-tunes the model on antibody data and compares it with BERT-like models on inpainting tasks, demonstrating comparable results.

My primary expertise lies in machine learning and generative models. Therefore, my review focuses mainly on the machine learning part of the paper. I believe protein generation is a promising and impactful direction, and I am pleased to see the latest machine-learning techniques successfully applied to this field. Overall, I think this paper is promising but I have the following concerns to be addressed.

1. Generally, the paper is well written. However, the authors claim in the abstract and other sections that "no existing approach is proficient for both unconditional and conditional generation," and use this as one of the main motivations for their work. Based on the following content, conditional generation refers to the inpainting task. I understand that BERT and autoregressive models may not be suitable for both tasks simultaneously, but diffusion models should naturally be well-suited for both tasks. Could you provide a more detailed explanation of why existing diffusion models cannot handle both tasks simultaneously? Additionally, even with the proposed method, fine-tuning was performed on the new antibody inpainting task, meaning that the parameters differ between the two tasks. Could the existing diffusion models not perform the inpainting task after fine-tuning? Finally, since this is one of the main motivations, I would suggest that the authors further clarify the importance of having a single model for both tasks. I believe that achieving both tasks together might improve data efficiency, but this may not be a comprehensive view.
2. The discussion of related work on diffusion models and the relationship between BFNs and diffusion models could be improved. First, classic works on discrete diffusion (e.g., [1-2]) should be appropriately cited in a prominent position (e.g., line 35) in the introduction, especially for the methods that have been applied to protein generation and other biological domains. Moreover, recent advancements [3] have shown that, in terms of both training and sampling, BFNs for discrete data are equivalent to a continuous diffusion model defined over latent variables (with a special noise schedule) (see details in concern 3). Therefore, some of the discussion in the introduction regarding continuous diffusion should be revised accordingly. I fully understand that the main contribution of this paper lies in the successful application of (large-scale) BFNs to the important domain of protein generation, rather than in developing a new generative model. However, I suggest that the discussion of generative models in the introduction be as rigorous as possible to avoid any potential misunderstandings.
3. Technically, I am confused by the issues encountered during sampling after training and the subsequent adjustments made to the sampling methods.
4. Technically, I am confused by the Entropy Encoding introduced by the authors in lines 305-312. I do not understand why the original BFN's training and testing processes would result in a mismatch (see more in concern 5). The authors provide a hypothesis, but it lacks further explanation and validation. I suggest including some new theoretical or experimental evidence to support this hypothesis. Additionally, I am not entirely clear on why the proposed method can resolve this issue. While I do not doubt the reproducibility of the paper, I'm curious why certain components of the original BFN did not work, whereas the corresponding adjustments did.
5. The original BFN employed a stochastic sampling method akin to the SDE sampler in diffusion models. In this paper, a heuristic ODE sampler is presented in Equation 16. It is noteworthy that in the recent work [3], BFN was explicitly characterized using a linear SDE, and corresponding first-order and second-order ODE and SDE samplers were derived

using the Fokker-Planck equation. On one hand, the ODE sampler in this paper might also be explicable using the theory from [3]. On the other hand, based on this perspective, some advanced SDE samplers proposed in [3] could correct certain biases in the original BFN SDE sampler, potentially improving experimental results and even addressing some perplexing experimental phenomena (such as concern 4, I'm not sure). Lastly, employing a second-order ODE sampler could effectively accelerate the sampling process, but it is not a necessity for this paper.
6. The paper claims that antibody inpainting is a zero-shot task. If I understand correctly, the proposed method requires fine-tuning on the same type of data, which seems to be a standard generalization problem rather than a typical zero-shot generalization. In machine learning, zero-shot generalization usually refers to performance without any fine-tuning after pre-training. I understand that the authors seem to want to emphasize that the proposed method has stronger generalization capabilities or is more data-efficient compared to the baseline. However, the current wording needs to be revised to clarify this distinction.
7. The paper appears to lack a comparison of parameter sizes between different models. I believe it would be beneficial to include a table systematically comparing the sizes of various models, their training costs, and the amounts of data used for each because these (scaling) parameters significantly affect the performance.
8. The experimental results of this paper are very impressive, but I am curious as to why the BFN achieves such a strong performance. Could this be related to the size of the BFN model (related to concern 4) used in this study? Beyond the experimental results, providing more theoretical or experimental analysis on why BFNs are particularly well-suited for protein generation tasks would significantly enhance the quality of the paper.

References
[1] Austin J, Johnson D D, Ho J, et al. Structured denoising diffusion models in discrete state spaces. Advances in Neural Information Processing Systems, 2021, 34: 17981-17993.
[2] Lou A, Meng C, Ermon S. Discrete diffusion modeling by estimating the ratios of the data distribution. ICML 2024.
[3] Xue K, Zhou Y, Nie S, et al. Unifying Bayesian Flow Networks and Diffusion Models through Stochastic Differential Equations. ICML, 2024.

Reviewer #2

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

Reviewer #3

(Remarks to the Author)
In this manuscript the author(s) introduce Bayesian Flow Networks (BFNs) in the biological domain. Although the BFN methodology appears powerful, the authors demonstrate its use here by taking protein sequence generation as an application, which is a problematic choice in principle because results are difficult to evaluate and gold standards for benchmarking are lacking.

The authors state that the primary goal of their ProtBFN technique is to learn the distribution of natural proteins. However, they admit there is no direct measure to assess the 'naturalness' of generated protein samples, such that they have to resort to computing indirect statistical and biophysical properties which they then compare against distributions that the autors deem natural. This benchmark has a high degree of uncertainty.

Regarding alternative benchmarks, given that "sequence leads to structure leads to function" and given that Alphafold has shown superior predictive power of tertiary protein structure, evaluating the protein space generated by ProtBFN using Alphafold would have been a more direct way to gain information on the information contained in the generated sequences. The authors use Netsurfp to validate their generated sequences, which predicts structural features such as secondary structure but not full tertiary structures. The authors turn closer to structure by using template modelling to relate their generated sequences to experimental protein structures in the CATH database, but this is less powerful and arguably more prone to error than using tertiary structures predicted by Alphafold.

The manuscript is difficult to read, has grammar/spelling problems in various places, and contains a lot of jargon that is not suitable for the wider readership of Nature Communications. This reviewer is wondering whether the manuscript would not be a better fit for a more specialised journal or a computer science venue.

The authors compare their ProtBFN method against the ProtGPT2 and EvoDiff methods for generating protein sequences. However, the latter two methods have been trained on a different database, UniProt50, which, unlike the authors training database UniProtCC, includes proteins of hypothetical or unknown existence. This makes it impossible to assess whether the authors' method performance is due to a more optimal method, as the authors claim, or to different training sets.

The authors test their BFN method on antibodies, by training the networks over specific antibody databases. This is a realistic use case that can show tangible strength of the authors' method. However, although the authors claim to gain superior results, the outcome differs appreciably between the three CDRs (Table 3): While the authors' method slightly outperforms other methods for CDR1 and CDR2, performance for CDR3 is well behind the performance of two contenders (AntiBERTy and AbLang2). It would be interesting to investigate why this is the case. Table 2 also demonstrates that the

third CDR is least amenable to correct prediction by the authors' method.

In summary, it may well be that the authors have a worthwhile and powerful methodology, but this reviewer is not convinced by the validation efforts carried out by the authors to demonstrate its performance with respect to protein sequence generation.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This is my second review of this article. Compared to the initial review, I believe the article has significantly improved completeness and clarity. Specifically, the new analysis on sampling and other technical aspects has addressed the previous concerns. I have no further issues, except for a very minor point: I believe that inpainting under the same data distribution is a natural application of modeling joint probability and then inferring conditional probability, rather than a zero-shot task in the general sense. However, the context is already clear enough, and the authors can decide whether further refinement is necessary. As with my previous opinion, I consider this to be an exciting and interesting piece of work, and I recommend its acceptance.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have dealt with my comments appropriately. Specifically, the use of ESMFold to additionally demonstrate the power of the authors' method at the structural protein level is appreciated.

One minor comment remains: the panel of Figure 2 appear to be referenced incorrectly for the pLDDT score.

My name is Jaap Heringa.

(Remarks on code availability)

We thank the reviewers for their insightful comments and suggestions. We have invested a substantial amount of time and computational resources into directly addressing their concerns. We are pleased to report that in all cases, the extended results further reinforce the advantages of our overall approach and design decisions. We hope that the reviewers will find that our revised manuscript is more robust and transparent with respect to the design decisions. Please note that in all places where we refer to specific line numbers, these refer to the version of the manuscript where changes are tracked.

REVIEWER #1

This paper lies at the intersection of machine learning and biology. Specifically, it adopts a recently proposed generative framework based on a Bayesian Flow Network and trains a large model with 650 million parameters on protein sequences. According to various computational and biological metrics, the model effectively covers known protein types while generating reasonable and diverse novel proteins, significantly outperforming autoregressive and discrete diffusion models. Furthermore, the paper fine-tunes the model on antibody data and compares it with BERT-like models on inpainting tasks, demonstrating comparable results.

My primary expertise lies in machine learning and generative models. Therefore, my review focuses mainly on the machine learning part of the paper. I believe protein generation is a promising and impactful direction, and I am pleased to see the latest machine-learning techniques successfully applied to this field. Overall, I think this paper is promising but I have the following concerns to be addressed.

We greatly appreciate the reviewer's efforts in reviewing our manuscript and hope that they find their concerns addressed in detail. In particular, we have provided substantial new analysis regarding the sampling methods used within the manuscript, and expanded our discussion of generative methods under their advice, and provided new information regarding the training resources of various models used within the paper.

1. Generally, the paper is well written. However, the authors claim in the abstract and other sections that "no existing approach is proficient for both unconditional and conditional generation," and use this as one of the main motivations for their work. Based on the following content, conditional generation refers to the inpainting task. I understand that BERT and autoregressive models may not be suitable for both tasks simultaneously, but diffusion models should naturally be well-suited for both tasks. Could you provide a more detailed explanation of why existing diffusion models cannot handle both tasks simultaneously?

The reviewer is correct that in principle, diffusion-based methods should naturally be well-suited for both tasks. However, diffusion models (in the conventional sense) are not naturally well-suited to discrete data, with the fundamental challenge being how to define and model discrete data at a continuous spectrum of noise levels (e.g. how to apply Gaussian noise to a discrete label). Recently, there are now a plethora of 'discrete diffusion' methods have been

proposed; however these typically rely on either some discrete-time noising process [1] or are defined on a simplex and require some modification of the loss function (for example, in [2], a denoising cross-entropy loss is introduced which does not correspond to a variational loss or a traditional diffusion MSE loss). In contrast, the BFN directly optimises the negative log-likelihood of the data and retains the flexibility of a continuous time process. We identify in our manuscript that the prominent discrete diffusion method, EvoDiff, is empirically unsuited for the unconditional generation task.

Our intention was not to claim that no discrete diffusion method could be developed that is proficient for both unconditional and conditional generation; only that to date this has not been empirically achieved for protein sequence generation. We hope that by providing an extended discussion of discrete diffusion we can provide readers with this context (Lines 39-46), whilst further highlighting the significance that BFN's provide such a significant performance gain over methods such as EvoDiff.

2. The discussion of related work on diffusion models and the relationship between BFNs and diffusion models could be improved. First, classic works on discrete diffusion (e.g., [1-2]) should be appropriately cited in a prominent position (e.g., line 35) in the introduction, especially for the methods that have been applied to protein generation and other biological domains. Moreover, recent advancements [3] have shown that, in terms of both training and sampling, BFNs for discrete data are equivalent to a continuous diffusion model defined over latent variables (with a special noise schedule) (see details in concern 3). Therefore, some of the discussion in the introduction regarding continuous diffusion should be revised accordingly. I fully understand that the main contribution of this paper lies in the successful application of (large-scale) BFNs to the important domain of protein generation, rather than in developing a new generative model. However, I suggest that the discussion of generative models in the introduction be as rigorous as possible to avoid any potential misunderstandings.

Thank you for this suggestion. We have prominently cited these references as suggested and adapted our phrasing of our introductory discussion on generative models. (Lines 44-45).

3. Technically, I am confused by the issues encountered during sampling after training and the subsequent adjustments made to the sampling methods.

We have included an additional ablation of sampling methods within the supplementary material, where we compare the Reconstructed ODE (R-ODE) method proposed in our submission with the Base sampler method for BFNs and those introduced in Xue *et. al.* [3]. These new results empirically demonstrate the practical issues with sampling ProtBFN using existing approaches (broadly speaking; it doesn't work!) and how our decisions regarding R-ODE and filtering substantially improve the pLDDT and perplexity of generated samples. We hope that the reviewer will find these empirical results sufficiently motivating for our design decisions in sampling.

Extended motivation for R-ODE

When sampling the model using the standard procedure introduced in [4], we found that the generated sequences had excellent statistical correlation to the training set in terms of metrics such as amino acid & oligomer frequencies, sequence length and secondary structure annotations. What these metrics all have in common is that they can be computed from only local (in sequence space) residue information. However, when we considered the pLDDT - a metric that is computed over the entire sequence - we saw a marked shortfall compared to natural proteins. Our hypothesis was to generate globally coherent sequences, we needed to shift the generating distribution towards higher likelihood samples - akin to how diffusion models often use low temperature sampling, or autoregressive models use top-k sampling. This motivated our modified Reconstructed ODE (R-ODE) sampling method; where we assume that the most recent prediction from the model is the 'best guess' for the center of the input distribution, and reconstruct the input distribution at that point. In effect, this strategy completely removes any previously accumulated error, as previous predictions are forgotten.

As the reviewer highlights, our primary contribution in this work is not to develop new methodologies for BFN's; but rather new applications. However, to ensure that our design choices are clear and well motivated we have extended the discussion of our R-ODE in the Methods to summarise the above points (Lines 372-381).

New sampling results

We appreciate the reviewer highlighting the recent work of Xue *et al.* [3] which demonstrates how to formulate the sampling of a BFN as an ODE or SDE. To ensure that our proposed R-ODE is necessary, we implemented the proposed ODE and SDE sampling methods from Xue *et al.* (both first and second order) and added comparison to these in the new Section D of the Supplementary Information. Empirically, these results clearly highlight that our R-ODE and filtering substantially improve the pLDDT and perplexity of generated samples.

4. Technically, I am confused by the Entropy Encoding introduced by the authors in lines 305-312. I do not understand why the original BFN's training and testing processes would result in a mismatch (see more in concern 5). The authors provide a hypothesis, but it lacks further explanation and validation. I suggest including some new theoretical or experimental evidence to support this hypothesis. Additionally, I am not entirely clear on why the proposed method can resolve this issue. While I do not doubt the reproducibility of the paper, I'm curious why certain components of the original BFN did not work, whereas the corresponding adjustments did.

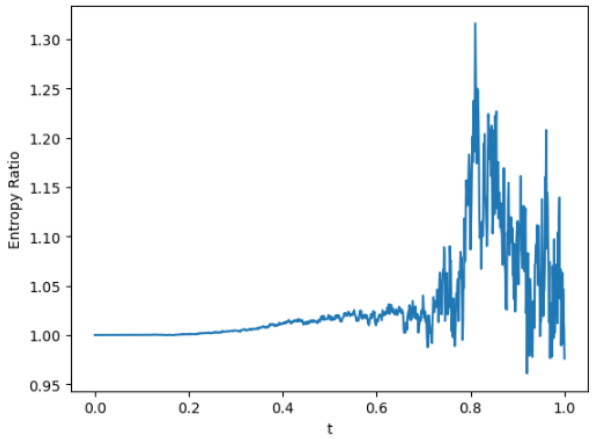
Our use of an entropy encoding instead of a time encoding was an empirical design decision, however we would not go so far as to say the (originally proposed) time encoding "did not work". Rather we noticed the entropy mis-match between training and sampling (details below) and hypothesised that conditioning on entropy directly could therefore help. Preliminary tests

showed that, at the least, training performance was not worsened by this decision and therefore it was kept constant from early in the model's development. We would draw a parallel to how diffusion models can be conditioned on the noise level, rather than time (for example, as proposed in Karras *et al.* [5] and used in recent models such as AlphaFold 3 [6]).

With that said, we primarily include discussion of the entropy encoding to ensure full reproducibility, rather than to claim it as a critical component of our work (we clarify this in the modified text, Lines 355-358). Whilst an ablation with ProtBFN retrained using a time encoding rather than an entropy encoding could further illuminate this point - we ultimately felt this was not necessary as (1) we do not consider this design choice a major contribution of our work and (2) re-training is exceptionally computationally expensive and we elected to commit this resources to training UR50BFN to address reviewer 3's concerns.

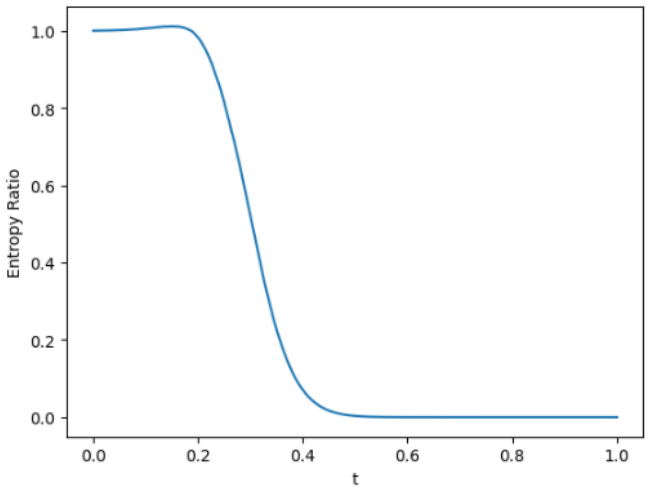
Empirical evidence of entropy mis-match between training and sampling

Nevertheless, we are more than happy to detail the entropy mis-match that motivated our design choice. To test this we can compare the input distribution observed during training (i.e. theta as sampled from the flow distribution) to that during sampling (theta after having been updated with a sample from the receiver distribution). As seen from the figure below, during sampling, when using the original BFN sampler, we see a considerably higher entropy (up to 25% more); especially at later times.



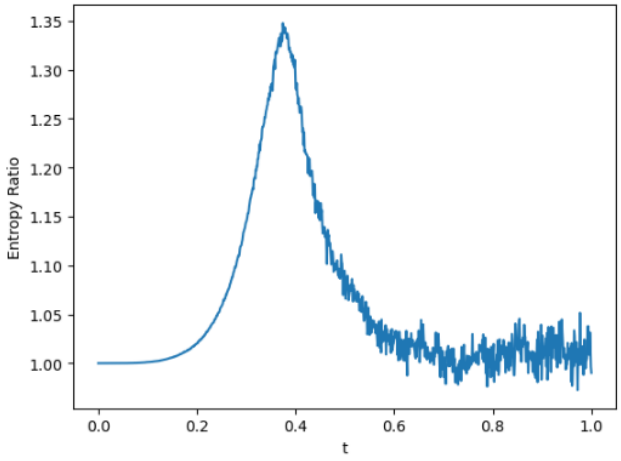
Caption Ratio of the median entropy of the input distribution during sampling to the entropy of the input distribution during training. The figure was obtained using 2000 samples and using the original sampling procedure proposed in [4].

In contrast, as seen in the figure below, during sampling, when using the R-ODE sampler, we see a substantially lower entropy. Our hypothesis is that the R-ODE sampling method is able to align the generated samples with the direction of the underlying isotropic noise, allowing it to achieve even lower entropy than the flow distribution.



Caption Ratio of the median entropy of the input distribution during sampling to the entropy of the input distribution during training. The figure was obtained using 2000 samples and using the R-ODE sampling procedure.

To further test this, we modified the R-ODE sampling method such that the isotropic noise is resampled in every step. In this case, shown in the figure below, we observe the entropy ratio is never below 1, indicating that it is R-ODE's ability to align the prediction with the noise through time that allows it to produce such a dramatic drop of entropy. However, this is clearly a complex topic that would be of interest for future study.



Caption Ratio of the median entropy of the input distribution during sampling to the entropy of the input distribution during training. The figure was obtained using 2000 samples and using the R-ODE sampling procedure, modified so that the isotropic noise is resampled at every step..

In any case, the substantial divergence in entropies, seen with any sampling method, would cause the input distribution and time to mismatch during training. By inputting the overall entropy to the model, instead of time, we effectively sidestep this possible out-of-distribution input.

5. The original BFN employed a stochastic sampling method akin to the SDE sampler in diffusion models. In this paper, a heuristic ODE sampler is presented in Equation 16. It is noteworthy that in the recent work [3], BFN was explicitly characterized using a linear SDE, and corresponding first-order and second-order ODE and SDE samplers were derived using the Fokker-Planck equation. On one hand, the ODE sampler in this paper might also be explicable using the theory from [3]. On the other hand, based on this perspective, some advanced SDE samplers proposed in [3] could correct certain biases in the original BFN SDE sampler, potentially improving experimental results and even addressing some perplexing experimental phenomena (such as concern 4, I'm not sure). Lastly, employing a second-order ODE sampler could effectively accelerate the sampling process, but it is not a necessity for this paper.

Please refer to our response to point 3.

6. The paper claims that antibody inpainting is a zero-shot task. If I understand correctly, the proposed method requires fine-tuning on the same type of data, which seems to be a standard generalization problem rather than a typical zero-shot generalization. In machine learning, zero-shot generalization usually refers to performance without any fine-tuning after pre-training. I understand that the authors seem to want to emphasize that the proposed method has stronger generalization capabilities or is more data-efficient compared to the baseline. However, the current wording needs to be revised to clarify this distinction.

When we refer to zero-shot, we are referring to the specific task of in-painting, not the choice of data. AbBFN was never trained specifically to perform in-painting on antibodies, this just arises naturally as a capability due to our ability to sample from the conditional distribution having trained only on the unconditioned joint distribution of all variables. We have added additional text to clarify this point (Lines 246-250).

7. The paper appears to lack a comparison of parameter sizes between different models. I believe it would be beneficial to include a table systematically comparing the sizes of various models, their training costs, and the amounts of data used for each because these (scaling) parameters significantly affect the performance.

We have added a new section to our supplementary information (Section G) directly addressing this question. We cover the compute resources and training data for each model discussed and evaluated within the manuscript. In particular, ProtBFN uses ~1.5x as much data as ProtGPT2 and EvoDiff, and similar overall GPU hours to ProtGPT2 in training. AbBFN is a much larger model than AntiBERTy and AbLang2, but we note that the overall compute cost is not that much greater than AntiBERTy, which also uses substantially more data.

REVIEWER #2

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

We greatly appreciate the reviewer's efforts in co-reviewing the manuscript and wish them luck in the early stages of their career!

REVIEWER #3

In this manuscript the author(s) introduce Bayesian Flow Networks (BFNs) in the biological domain. Although the BFN methodology appears powerful, the authors demonstrate its use here by taking protein sequence generation as an application, which is a problematic choice in principle because results are difficult to evaluate and gold standards for benchmarking are lacking.

The authors state that the primary goal of their ProtBFN technique is to learn the distribution of natural proteins. However, they admit there is no direct measure to assess the 'naturalness' of generated protein samples, such that they have to resort to computing indirect statistical and biophysical properties which they then compare against distributions that the autors deem natural. This benchmark has a high degree of uncertainty.

We would like to thank the reviewer for their comments on our manuscript. We appreciate the concerns they raised regarding our choice of benchmarking and comparison. We have invested a substantial amount of time and compute into addressing their concerns through the training of a new model. We have also made substantial modifications to the text of the manuscript to address their concerns regarding clarity.

To the point of the suitability of protein sequence generation as a use case for BFN's - we fully agree that this is a challenging domain with complexities that prevent it from being an easy testing ground for new deep learning innovations. However, the original and still key goal of our work was improved generative modelling for proteomics; for which we considered (and believe our work now shows) BFN's to be a powerful algorithm.

We appreciate the reviewer highlighting the particular challenge of evaluating a generative model in a domain such as proteomics. Indeed, any generative model - whether for proteomics or beyond - aims to learn the ("natural") distribution of the data on which it is trained. How to evaluate these models is always a challenge; re-generation of exactly the training data clearly not desirable, however samples should appear as if they are drawn from the same distribution of the training data. The established approach is therefore to choose one, or several, key metrics of the data and compare the distribution of real ("natural") and generated samples. For

example, the FID score's [7] commonly used to evaluate the quality of generative image model follow exactly this approach.

However, there is no such widely accepted single metric for protein sequences - and, indeed, we agree with the reviewer that and single metric can't guarantee the quality of the generations (for example; we could generate sequences that have high confidence predicted structures as measured by pLDDT's, but this would not guarantee other properties such as suitable diversity in the secondary structure). It is for this reason, we endeavoured to provide many different analyses of our generations.

- Statistical properties (e.g. Fig 2a,b,c) and biochemical properties (Sup Fig 1) of primary protein sequences.
- Similarity of primary sequences to the training data (Fig 2f) and the established curated database of natural protein sequences (Table 1). We also repeat perform comparison using the learned representation of a leading protein language model, ESM-2 (Fig 3 and Sup Table 1).
- We consider the secondary structure properties of our generations (Fig 2e and Sup Fig 2).
- We also predict the full tertiary structures of our generations (highlighted by the Reviewer as a key test, discussed below). We consider the confidence of these folds (Fig 2d) and, once again, compare our generations to the established curated database of experimentally resolved structures, CATH (Fig 4 and Sup Fig 4). Finally, we also investigate specific properties of protein sequences that could be expected to be challenging to model by confirming the presence of coherent mult-domain interactions (Sup Fig 5).

We believe this is a rigorous and suitable best effort to effectively evaluate a generative sequence model that incorporates the benchmarking requested by the reviewer (detailed below).

Regarding alternative benchmarks, given that "sequence leads to structure leads to function" and given that Alphafold has shown superior predictive power of tertiary protein structure, evaluating the protein space generated by ProtBFN using Alphafold would have been a more direct way to gain information on the information contained in the generated sequences. The authors use Netsurfp to validate their generated sequences, which predicts structural features such as secondary structure but not full tertiary structures. The authors turn closer to structure by using template modelling to relate their generated sequences to experimental protein structures in the CATH database, but this is less powerful and arguably more prone to error than using tertiary structures predicted by Alphafold.

We strongly agree with the reviewer that evaluation of the generated protein space through the lens of tertiary structures is important for understanding its practical and functional diversity. Whilst secondary structure prediction with Netsurfp is the first structural validation we perform (Fig 1e), we highlight that a significant portion of our analysis – Fig 4(a-e) and the corresponding section of the main text, Supplemental Section C and Figures 4, 5, 7(d-e) and 8(a,d,e) – directly

leverages predicted tertiary structures from an AlphaFold-like model (ESMFold). We emphasise that this is fully aligned with the reviewers' requested analysis of using predicted tertiary structures and that at no point do we use template modelling. We apologise if the original text was not clear on these points and have ensured the revised manuscript is clear; moreover we provide further details of our analysis choices below.

Tertiary structure prediction and analysis

We leverage ESMFold [9] in place of AlphaFold; but otherwise we do provide the detailed analysis of tertiary structures requested by the reviewer.

Whilst the AlphaFold models have provided leading performance through structure prediction competitions such as CASP, ESMFold is another widely adopted protein folding model that we choose to use for two primary reasons.

1. Our use case is to predict the structure of *de novo* generated protein sequences. It has been observed that on *de novo* designed and orphan sequences, the advantage of AlphaFold over leading single-sequence methods such as ESMFold is greatly reduced, or even AlphaFold can perform worse (for example, we refer the reviewer to [8], specifically Fig 3). This is attributed to the fact that AlphaFold takes advantage of templates from PDB and MSAs derived from multiple sequence databases which have been shown to influence the results of the models. When suitable templates or MSA's are unavailable, the performance can be correspondingly worse. ESMFold by contrast is a single-sequence, template-free deep learning algorithm that is not explicitly conditioned on inputs derived from external databases.
2. Practically speaking, large scale predictions ESMFold are significantly faster than AF2 whilst maintaining similar performance.

With regards to our comparison of generated structures to the CATH database; this is also done using the tertiary structures predicted with ESMFold. Comparison to experimentally validated functional domains is performed using two well established metrics (TM score and SSAP score) and detailed in Fig 4 of the main manuscript. This is critical because whilst the pLDDT metric (Fig 2d) is very useful to assess the confidence of the modeling pipeline, it can fail short in assessing the 'naturalness' of the structure. However, by comparing our generated sequences with experimentally solved functional domains we can assess how feasible our generated sequences are.

Use of Netsurfp and secondary structure prediction

NetSurfP is a well established peer-reviewed computational method (cited more than 100 times in less than 2 years) that is only one of several methods we use to comprehensively assess our generated sequences. Whilst we fully agree with the reviewer that considering predicted tertiary structures (also provided in the manuscript and discussed above) is highly important; by including NetSurfP we ensure that our analysis spans primary sequence, secondary and tertiary

structure, effectively saturating the amount of metrics available for single chain protein generation.

The manuscript is difficult to read, has grammar/spelling problems in various places, and contains a lot of jargon that is not suitable for the wider readership of Nature Communications. This reviewer is wondering whether the manuscript would not be a better fit for a more specialised journal or a computer science venue.

We have reviewed the grammar and spelling of our manuscript and removed a substantial amount of ML-specific jargon and terminology. We have also edited the language more generally and made changes throughout the text to improve clarity. We hope that the reviewer will find the use of language more suited to the broad and multi-disciplinary audience of Nature Communications. Should any specific concerns remain we would be more than happy to directly address them.

The authors compare their ProtBFN method against the ProtGPT2 and EvoDiff methods for generating protein sequences. However, the latter two methods have been trained on a different database, UniProt50, which, unlike the authors training database UniProtCC, includes proteins of hypothetical or unknown existence. This makes it impossible to assess whether the authors' method performance is due to a more optimal method, as the authors claim, or to different training sets.

Understanding the origin of ProtBFN's empirical performance is clearly of utmost importance. After examining the official repositories of ProtGPT2 and EvoDiff, we have not identified the code to replicate their training process. If we need to re-implement them, this will open doors to even more concerns about the fair comparison between the different methods. We therefore decided to address the reviewer's concerns by evaluating the BFN methodology on Uniref50 data. In this light, the following analysis has been added to the Supplementary information, Section F:

- We have trained a variant of ProtBFN, UR50BFN, on Uniref50 instead of UniProtCC.
- We have evaluated samples generated from UR50BFN against those of ProtGPT2 (ablating model) and shown that the BFN methodology we propose substantially improves the naturalness and diversity of generated samples in comparison to the autoregressive approach. In particular, we observe overall higher pLDDT scores (Supp. Fig. 7d) and SSAP scores (Supp. Fig. 7e) in comparison to ProtGPT2, indicating stronger overall global coherence. The proteins generated by UR50BFN are also more diverse and provide better coverage of UniRef50 by sequence identity (Supp. Table 2). 33.9% of UR50BFN sequences have 50% identity with some element of UniRef50, in comparison to only 15.7% of ProtGPT2 sequences.
- We have evaluated samples generated from UR50BFN against those of ProtBFN (ablating data source) and shown that the decisions we have made with respect to our data are responsible for a substantial portion of the overall performance of ProtBFN. For example, in Fig. 7d and 7e, we now observe lower overall pLDDT and SSAP scores, respectively, indicating that the choice of data source has a substantial impact on the

overall global coherence of proteins. Similarly, with respect to Supp. Table 2, we observe lower cluster hits and coverage scores when comparing UR50BFN to ProtBFN, which suggests that the UniProtCC dataset (which does not omit non-centroid cluster members) allows models to provide much stronger overall coverage of the protein space.

In view of this new analysis, we hope that the reviewer will agree with us that both our choice of data and our methodology are important for the overall performance that we present in the manuscript.

The authors test their BFN method on antibodies, by training the networks over specific antibody databases. This is a realistic use case that can show tangible strength of the authors' method. However, although the authors claim to gain superior results, the outcome differs appreciably between the three CDRs (Table 3): While the authors' method slightly outperforms other methods for CDR1 and CDR2, performance for CDR3 is well behind the performance of two contenders (AntiBERTy and AbLang2). It would be interesting to investigate why this is the case. Table 2 also demonstrates that the third CDR is least amenable to correct prediction by the authors' method.

We thank the reviewer for this observation and agree that the discrepancy between the different regions is noteworthy.

Much of modern protein sequence modelling relies heavily on coevolutionary restraints, wherein the model learns to approximate the underlying conditional probabilities of different positions in a protein given the rest of the sequence. This is also evidenced in our study by the fact that training ProtBFN with careful sampling of our curated dataset results in higher performance. However, antibodies are not subject to the same coevolutionary constraints as other proteins. The CDR-H3 loop is generated through a combination of V-, D-, and J-genes, with the junctions between these genes being subject to non-templated nucleotide addition. That is, no evolutionary information is available to the model when trying to inpaint these junctions. On the other hand, CDR-H1/2 are generated entirely by the underlying V-genes, and so can be more accurately inferred by such a model relying on coevolutionary information (i.e. once a model is able to infer the underlying V-gene in an antibody from the remainder of the sequence, guessing the sequences of these loops is a much easier task). We argue that the lower predictability of CDR-H3 loops due to this diversity and randomness is the likely reason behind AbBFN's comparatively lower performance on this loop. We also note that as AntiBERTy and AbLang2 do not share the exact sequences that they were trained on, it is highly likely that these models have been exposed to our test sequences during their respective training procedures. Due to the immense diversity of CDR-H3 loops, these models will therefore have the advantage of having been exposed to the exact sequences that we test on, leading to higher performance specifically on this loop.

As this is a point that will be of interest to the immunoinformatics community, we have added a detailed explanation on this observation starting on line 261.

References

[1] Austin, Jacob, et al. "Structured denoising diffusion models in discrete state-spaces." *Advances in Neural Information Processing Systems* 34 (2021): 17981-17993.
[2] Mahabadi, Rabeeh Karimi, et al. "Tess: Text-to-text self-conditioned simplex diffusion." *arXiv preprint arXiv:2305.08379* (2023).
[3] Xue, Kaiwen, et al. "Unifying Bayesian Flow Networks and Diffusion Models through Stochastic Differential Equations." arXiv preprint arXiv:2404.15766 (2024).
[4] Graves, Alex, et al. "Bayesian flow networks." arXiv preprint arXiv:2308.07037 (2023).
[5] Karras, Tero, et al. "Elucidating the design space of diffusion-based generative models." Advances in neural information processing systems 35 (2022): 26565-26577.
[6] Abramson, Josh, et al. "Accurate structure prediction of biomolecular interactions with AlphaFold 3." Nature (2024): 1-3.
[7] Heusel, Martin, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." Advances in neural information processing systems 30 (2017).
[8] Kandathil, Shaun M., Andy M. Lau, and David T. Jones. "Machine learning methods for predicting protein structure from single sequences." Current Opinion in Structural Biology 81 (2023): 102627.
[9] Lin, Zeming, et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model." Science 379.6637 (2023): 1123-1130.

We are grateful to the reviewers for their continued engagement with our work and constructive discussions during the review process. Below we address the final few points raised in the last round of feedback.

Reviewer #1

This is my second review of this article. Compared to the initial review, I believe the article has significantly improved completeness and clarity. Specifically, the new analysis on sampling and other technical aspects has addressed the previous concerns… As with my previous opinion, I consider this to be an exciting and interesting piece of work, and I recommend its acceptance.

Thank you for the kind words about our work.

I have no further issues, except for a very minor point: I believe that inpainting under the same data distribution is a natural application of modeling joint probability and then inferring conditional probability, rather than a zero-shot task in the general sense. However, the context is already clear enough, and the authors can decide whether further refinement is necessary.

We fully agree with the reviewer's characterization of the task and that while “zero-shot” is a suitable description within our context, we do not want to confuse the meaning to readers first encountering our work. For this reason we have updated the main text of our manuscript to only use the phrase “zero-shot” in the results section, where the specific task is being discussed and the context is clear. Practically, that means we have removed the following references from other sections:

- **Introduction:** We update the phrase “To validate the zero-shot conditional generation capabilities of our model…” to “To validate the conditional generation capabilities of our model…”.
- **Methods: Fine-tuning AbBFN+**: We update the phrase “The exponential moving average parameters are used for zero-shot inpainting results.” to “The exponential moving average parameters are used for inpainting results.”.
- **Discussion** We update the phrase “Combined with the arbitrary zero-shot conditional generation we demonstrate in this paper…” to “Combined with the arbitrary conditional generation we demonstrate in this paper…”

Reviewer #3

The authors have dealt with my comments appropriately. Specifically, the use of ESMFold to additionally demonstrate the power of the authors' method at the structural protein level is appreciated.

One minor comment remains: the panel of Figure 2 appear to be referenced incorrectly for the pLDDT score.

My name is Jaap Heringa.

We are grateful for Professor Heringa's feedback and believe the paper has been improved as a result of the review process. We have updated the incorrect reference to Figure 2 in the revised manuscript.