nature portfolio

Peer Review File

Learning millisecond protein dynamics from what is missing in NMR spectra

Corresponding Author: Professor Dorothee Kern

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

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

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
Despite advances in predicting protein structures from sequence, the biological activities of proteins and other biomolecules also crucially depend on dynamics occurring over timescales spanning twelve orders of magnitude. This manuscript represents a timely and ambitious attempt to harness AI to learn from NMR measurements of micro-to-millisecond timescale motions, a class of dynamics often tightly coupled to biological function. These experiments are technical and time-consuming, and are currently available only for a limited number of proteins. The authors develop the clever idea that missing resonances reported in routinely collected spectra carry such information, thus greatly expanding the data available for this analysis. This is an original and noteworthy idea, especially in the era of AI– I am sure others in the field will wonder "why did I not think of that?". The authors also make a valuable contribution by highlighting the need for standardized reporting of NMR dynamics datasets.

Although the work sets an important milestone in the field, I was disappointed by the lack of prospective experimental tests, i.e. testing the model by generating and experimentally verifying de novo predictions. Such prospective testing is now standard in the development and testing of predictive models. Without such tests, it is difficult to assess whether the model has learned generalizable features of micro-to-millisecond motions or merely recapitulates patterns in the existing data.

Given that the authors are leaders in NMR dynamics measurements, this seems a reasonable request. For example, the authors could test their model on a protein with known resonance assignments but lacking Rex data or predict missing assignments for small relatively easy to assign proteins. Even a single case study would significantly strengthen the manuscript's central claims. Alternatively, if such prospective testing is not feasible, the authors should provide a compelling justification. The authors should also make the model and associated software available for others to test and apply to new systems.

Additional comments:

1. Throughout the manuscript, the authors need to clarify that they are not predicting "micro-to-millisecond timescale motion" but rather predicting which residues exhibit micro-to-millisecond timescale motions as detected by NMR chemical exchange methods. There is an entire universe of micro-to-millisecond timescale motions which will not be detected by these NMR experiments either because the difference in chemical shift or population of the alternative state are too small. The method also does not predict the timescale of the motion or the population of alternative states even if they are detectable by NMR. So again – predicting which residues are NMR active in these experiments, not predicting micro-to-millisecond timescale motion in general.

2. The authors should list and examine all other plausible reasons for why a resonance might be missing from the NMR spectrum. For example, a resonance might be overlapped with another resonance. This possibility could be directly tested by predicting chemical shifts from the 3D structure and seeing whether the resonance falls in a crowded region of the spectrum. In addition, a resonance might be missing not due to 15N Rex broadening but rather due to broadening of the amide proton. Does this complicate the integration of data sets specifically measuring 15N broadening versus 1H?

3. A related comment: the Rex contribution depends in a complex manner on the population of the alternative conformation as well as the difference in chemical shit and exchange kinetics and whether RF is being applied or not. Thus, in each R1rho experiment, for example, whether one detects Rex or not crucially depends on the chosen spin lock powers and offsets, or if using CPMG, on the range of tauCP values employed. Thus, depending on the experimental setup, one may or may not detect Rex; was anything done to control for these variables?

4. The authors can be more quantitative about the criteria that need to be satisfied for a peak to be missing from the NMR spectrum. For example, a resonance with high R2 due to a high molecular weight protein or anisotropic diffusion will require a smaller Rex contribution before it falls below the detection threshold. Thus a resonance could be missing for the same Rex from a large protein but not in a small one. This is likely to confound the analysis. Perhaps the analysis could be done by binning different molecular weight proteins. In addition, a peak might be missing on one spectrometer but not another, either due to differences in sensitivity or due to differences in Rex when working at different fields. It would be good to explain how this aspect of the analysis was handled.

5. On p 5, the authors rightly point out that a residue may have a ‘normal’ R2 value despite competing slow and fast motions canceling their contributions. The authors state that if anything R2/R1 is most likely underestimating Rex; what was the basis for this claim?

6. I found the analysis presented on pages 10/11 to be difficult to follow and highly technical without driving the key points with compelling statistics backing the claims being made.

7. It would be helpful to include comparisons to pre-existing models that predict protein dynamics like DynaMine or elastic models. Especially since the AUROC scores for Dyna-1 are not that high, usually ranging around 0.6-0.7, this comparison may provide stronger evidence that Dyna-1 is a step forward in predicting protein dynamics.

8. The authors may want to point out that resonances such as adenine-C2 are often missing from HSQC spectra in RNA due to line broadening caused by protonation of adenine-N1, so that this approach could hold promise in applications to nucleic acids as well.

Minor comments:

Fig. 1c: It would be beneficial to include quantitative metrics on how well HYDRONMR predictions agree with the experimental data, particularly since the proposed model relies on values derived from these predictions. The criteria used to define different scales of motion are also unclear. Are these definitions based on previously established thresholds in the field?

Fig. 1d: The grey distribution in Fig. 1d is not labeled. Does it correspond to residues without observed exchange?

Fig. 1d: The sequence conservation distributions appear broad and overlapping. The assertion that residues exhibiting μs–ms motion are "highly conserved" is insufficiently supported based on this data.

Fig. 4a: This subfigure is difficult to follow and could be better explained in the manuscript/figure description.

Ext. Data Fig. 6b: Based on the data presented in Ext. Data Fig. 6b, it appears that Dyna-1 and other models perform only modestly, with AUROC values below 0.6 in predicting exchange residues. This should be acknowledged more directly and factored into the interpretation of the results.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
Generally this is a good paper that explores ways of capturing protein dynamics in ML models, offering downstream improvements in speed, accuracy as well as interpretation advantages.

The paper is poorly written, with many avoidable typos, and several poorly supported statements. The paper, with more attention to writing, could develop the main points more carefully.

The paper is also poorly developed with respect to connection to prior results and alternate approaches.

I would imagine that simple alternate predictors, such as number of contacts, surface accessible surface area and secondary structure could combine to make predictors of the dynamic quantities you measure. You show some of these metrics individually, but do not show how these could be composed into a very simple predictor. Please make comparison to other methods and simple metrics more prominent to help readers interpret you main findings and models. You test AF2-pair, ESM-2, and ESM-3-based models (billions of parameters to predict a single number per residue), but not much simpler ones that might work well.

The data presented shows that you have an ability to predict some dynamics related quantities from NMR data, but the paper does not really illustrate/support how this code could be used in practical settings. How do you sample residue conformations from this model, are they biophysically accurate/plausible, how do you differentiate error (of either kind) from the biophysically relevant dynamics you focus on?

What other NMR methods or dynamics methods would work here? J-couplings from NMR, HD exchange from NMR and/or other measures that convolve accessibility and dynamics could be used as validation vignettes or test cases.

Using unassigned residues, and finding ways to bring these missing data back into your analysis is a great idea.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
Summary of the key results:

You describe a method to predict us-ms motions in proteins based on their sequence, on the assumption that peaks missing in 15N spectra indicate such dynamics. This assumption is related to (sparse) relaxation data.

Originality and significance:

Whilst this is certainly an interesting take on the dynamics problem, some of the concepts you claim to introduce are not new. Consider for example the dynasome, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0033931, where the relation between dynamics and function was already explored and introduced. You should acknowledge this seminal work. In terms of significance, there are major issues with your dataset, described below, which make it impossible to assess whether your method works properly and is generic.

Data & methodology:

- The protein environment can have a very large influence on its behavior, for example phosphorylation can switch a disordered region to a folded one. You do not really address this aspect in relation to your method - dynamics is after all determined by the energy landscape of the protein, which is variable. How persistent are these regions you predict, and how dependent on your filtering of the BMRB dataset? You only included monomeric datasets with no cofactors present. Did you also check for pH, temperature, possible other agents in the solution (e.g. urea) and such, as these can have a large influence on protein behavior?

- Your assumption is that missing 15N signals are due to us-ms motions, but which proportion of these is likely missing due to signal overlap, which would in any case be more common in loops? Also here, for intermediate exchange, the field strength of the NMR spectrometer does play a role, please comment on this aspect.

- You have a bias towards residues loops in your dataset, as mentioned on line 199-200 - this also makes sense from an intermediate exchange perspective (see also previous point). Instead of many individual, personally picked protein cases can you present the statistics of these data and how this might affect the performance measures of your method in relation to your test data? How long are these loops? How long are the helices? You have to more extensively describe the dataset and explore the possible bias therein.

- What is your reasoning to employ a 0.65 cutoff for the hetNOE? And how does this precisely statistically relate to the ‘missing peaks’? How did you ‘visually assess’ each dataset - such a procedure is not reproducible, why were those entries removed? This part is not well described and vague, you should precisely describe why you took these decisions and how. Also the four cases you present in Figure 1 - even though such a qualitative comparison is not relevant here (you can use the statistics), if you chose these 4 why did you pick them?

- pLDDT is maybe a reasonable predictor of backbone order parameters by employing the structure-based Bruschweiler method, but on a dataset of 10 proteins this is not very convincing. Consider also the large-scale perspective, for example https://doi.org/10.1016/j.jmb.2024.168900. Further in relation to AF2, you say the ‘glimpses that AF2 might have learned information about multiple conformations abound’. There is also other work that instead shows that this is only possible if it has memorised the conformation, see for example https://www.nature.com/articles/s41467-024-51801-z. Please incorporate these more critical aspects in your work.

- In relation to lines 213-216, it has been shown that B-factors at cryogenic temperatures (which constitutes most of the x-ray diffraction determined structures in the PDB) have no relation to dynamics at room temperature, please see https://www.pnas.org/doi/abs/10.1073/pnas.1323440111.

- In relation to lines 209-210, even though NMR structures were not used in training AF2, there is of course overlap with structures determined by x-ray crystallography, where the core fold will be very similar. There is therefore significant overlap between the data in the BMRB and in the ‘x-ray PDB’. Please clarify how you dealt with this - did you only use the BMRB entries that only had NMR structures? You also have to use a 30% SI cutoff for creating training dataset for methods such as

yours, as is standard in the field. Anything higher will confer significant bias on training/test.

- The level of conservation is also dependent on the multiple sequence alignment you use, and the evolutionary depth thereof. Can you comment on this aspect? For example, the AF2 pair representation is doing very well as baseline, why would this be the case for the problem you are addressing?

Appropriate use of statistics and treatment of uncertainties

In terms of machine learning, I refer to the DOME recommendations (https://www.nature.com/articles/s41592-021-01205-4), which provide an essential guide to reporting data and performance measures. You have to cross-validate and report the spread of each fold (at least 5 here), for example - this is possible with the dataset you have.

Conclusions: robustness, validity, reliability

Given the issues with likely bias in the dataset, it is not possible to ascertain whether the method is robust, valid nor reliable.

Suggested improvements: experiments, data for possible revision

- Overall better explanations of choices made to create the datasets. Improve filtering of datasets and use much higher SI separation between the protein sequence
- Reporting of dataset statistics - how large is the bias toward loop residues and how does this influence your results (also relation to test set)
- If you have scatter plots like in Figure 1, please also report correlations

References: appropriate credit to previous work?

No, see methods section.

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

Whilst I appreciate your enthusiasm in conveying the story of how you got to developing the method, it would be scientifically more appropriate if the data and approach speak for itself. Already in the abstract phrases like ‘we realised an observable for us-ms motion was hiding in plain sight’, ‘We made the bold assumption’ are unnecessary and distract from the facts. Also later (e.g. line 249) ‘To our surprise’, and others.

(Remarks on code availability)

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors have addressed my original comments and concerns, and the revised manuscript is substantially improved and suitable for publication in Nature.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
Summary of the key results:

You illustrate on the basis of NMR data that protein amino acid residues with us-ms exchange tend to be more conserved in evolution and relate this observation to missing HN assignments in chemical shift lists from the BMRB. On the assumption that such missing HN assignments are related to said us-ms motions, you develop a sequence-based predictor to identify such residues for any sequence, and further relate these predictions to independent NMR relaxation data by a mix of large-scale statistical performance analyses and detailed case study illustrations.

Originality and significance:

The relation between amino acid conservation and slower us-ms timescale movements is very interesting, and important for further developments to distinguish different timescale ranges in protein dynamics. Also novel, though indeed obvious in hindsight, is your insight to use missing HN assignments as large-scale evidence for slower us-ms timescales. However, I am still not convinced by the performance of the predictor, where underlying biases and performance limitations are not sufficiently explored.

Data & methodology:

You point out that missing HN assignments could have other reasons than us-ms timescale movements, as evidenced by essentially the difference in p(missing) vs p(exchange). Yet in the manuscript you advance the point that your p(missing) does reliably capture these us-ms timescale movements, without in my opinion critically investigating the reasons behind that (see points below) nor reporting all appropriate performance measures (for this, see comment in next section).

1. The performance on the relaxation dataset is significantly lower, with AUROCs in the 0.6-0.7 range, which is acceptable but not great, indicating that there are indeed other reasons for the missing HN that you might be capturing. In addition, for the CPMG dataset, about 6 proteins seem well predicted but 4 are at 0.6 or below, which really is not very good evidence that your predictor works across the board. You then use details in individual protein case studies to illustrate it does work, but this is not statistical evidence across the board for the performance. In the text these two types of approach (illustrative case study vs. statistical performance) are intermingled, which is confusing in terms of assessing your work, as these are not equivalent in relation to predictive performance - in the end the paper is about the predictor after all. It would make the paper much more effective and clear if you would separate the actual evidence from the illustrative case studies, which are nice but in the end always anecdotal. In relation to the per-protein performances, it would be interesting to report the correlation between the % of known missing assignments to the AUROC of the predictor for each protein.

2. In relation to the bias point I previously brought up, if the missing HN assignments are primarily in loops, then this would constitute a clear bias - even if slower motions primarily occur there. If you are trying to predict which people tend to run slower (the dynamics in your case), this is more likely in older people (the loops) and you have a way to determine people's age somehow (the pLM in relation to the loops) then the bulk from your overall performance can come from age detection (loops), not from detecting run speed (slower motions). You do show per-ESM2-predicted secondary structure (SS) categories for the relaxation dataset, which is interesting, but not for the overall dataset - I would very much like to see what the % of missing residues is per SS category, and what the performance of your method is per sub-SS category. Also note that in our experience predictors like ESM2 tend to over-predict helix for dynamic regions, which might influence your results.

3. Incomplete sequence-level separation between evaluation sets: If I interpreted this correctly, the lower sequence-identity thresholds (50% and 30%) were applied only to the training set, while the validation and test sets were originally separated at an 80% identity cutoff and then seem to have been kept fixed. Consequently, moderate-to-high sequence similarity can remain between validation and test proteins, allowing information gained during validation-based model selection to indirectly transfer to the test set. This asymmetry reduces the effective independence of the test evaluation and may lead to optimistic performance estimates. This is also indicated by the similar validation and test AUROC values, which suggest limited independence and are consistent with residual similarity between these sets. While not definitive proof of leakage, this pattern weakens claims of strong generalization and suggests that the test set may be easier than intended.

Appropriate use of statistics and treatment of uncertainties:

I again refer to the DOME recommendations (https://www.nature.com/articles/s41592-021-01205-4), which provide an essential guide to reporting data and performance measures. This not only relates to the cross-validation (which you addressed, thanks) but also to reporting more than only the AUROC to enable a better assessment of where the (lack of) performance originates from. Given that your p(missing) can also be converted into a classifier using the best possible cutoff, it would also be of interest to explore these cutoffs and relate this to your data in the context of a classifier, report TPR, FPR, and so forth to get a better idea of where your prediction works (and where not). Whilst classification does of course have its drawbacks and I am not suggesting you should use this as the main output of your method, it also helps to assess performance - for example is the best p(missing) the same for all individual proteins or is the best cutoff very different? This would already make interpretation of the p(missing) much more difficult.

Conclusions: robustness, validity, reliability:

Whilst the method seems to work to identify missing assignments, the reasons for this are multiple and are not explored critically enough in the paper, especially given the low performance on independent data and varying results on different proteins.

Suggested improvements: experiments, data for possible revision:

See above.

References: appropriate credit to previous work?:

OK.

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

Thank you for adapting the tone of the text.

(Remarks on code availability)

Referee #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Referees' comments:

Referee #1:

Despite advances in predicting protein structures from sequence, the biological activities of proteins and other biomolecules also crucially depend on dynamics occurring over timescales spanning twelve orders of magnitude. This manuscript represents a timely and ambitious attempt to harness AI to learn from NMR measurements of micro-to-millisecond timescale motions, a class of dynamics often tightly coupled to biological function. These experiments are technical and time-consuming, and are currently available only for a limited number of proteins. The authors develop the clever idea that missing resonances reported in routinely collected spectra carry such information, thus greatly expanding the data available for this analysis. This is an original and noteworthy idea, especially in the era of AI– I am sure others in the field will wonder “why did I not think of that?”. The authors also make a valuable contribution by highlighting the need for standardized reporting of NMR dynamics datasets.

Although the work sets an important milestone in the field, I was disappointed by the lack of prospective experimental tests, i.e. testing the model by generating and experimentally verifying de novo predictions.

Such prospective testing is now standard in the development and testing of predictive models. Without such tests, it is difficult to assess whether the model has learned generalizable features of micro-to-millisecond motions or merely recapitulates patterns in the existing data.

We thank the reviewer for the thorough and supportive review, highlighting and very well describing the originality and major impact of this manuscript for biology. In response to this insightful and endorsing review, we performed exactly the additional NMR experiments on even two proteins for which **no** dynamics data were ever recorded, to deliver novel experimental verification and to showcase the practical application of Dyna-1.

Given that the authors are leaders in NMR dynamics measurements, this seems a reasonable request. For example, the authors could test their model on a protein with known resonance assignments but lacking Rex data or predict missing assignments for small relatively easy to assign proteins. Even a single case study would significantly strengthen the manuscript's central claims. Alternatively, if such prospective testing is not feasible, the authors should provide a compelling justification.

The authors should also make the model and associated software available for others to test and apply to new systems.

These are excellent suggestions, and we want to answer them point by point:

1. We fully agree with the reviewer that testing the Dyna-1 model with experimental data is key. For exactly this reason, we had curated relaxation data for 133 proteins (RelaxDB) and CPMG data for 10 proteins (RelaxDB-CPMG). These proteins were from our held-out test set, for which we ensured Dyna-1 was not trained with them or any similar proteins (defined by a 30% sequence similarity, 0.5 structural TM-score cutoff). Therefore, this approach delivered 143 proteins for such experimental verification of our predictive model, and we evaluated how the predictions from Dyna-1 compared to actual NMR relaxation experiments (see Fig. 4,5).

2. For a de novo prediction to showcase the practical application of Dyna-1 (requested by the editor as well), we followed the reviewer's exact suggestion to choose proteins with resonance assignments but lacking Rex data. These proteins came from our held-out test set (therefore, no more than 30% sequence similarity to training set and no structure homologues in the training set). For the revised manuscript, we expressed, purified and experimentally measured micro-millisecond backbone dynamics by CPMG dispersion. To be extra rigorous (the reviewer had suggested to only experimentally test one protein), we experimentally tested two proteins with de novo dynamics predictions, one for which Dyna-1 predicts major intrinsic dynamics and one with the other extreme, little dynamics. New text paragraph and figure in the manuscript reproduced below:

Prospective experimental test of Dyna-1.

"To validate the predictive capabilities of Dyna-1, we conducted a prospective experimental test after public release of the model and initial review of the manuscript. We selected proteins from the "mBMRB test" set to characterize their dynamics, as these are proteins that are already assigned but many do not have published relaxation data. **Fig. 6a** depicts the mean p(exchange) from each protein in the mBMRB test set vs. sequence length, plotted by alpha helical / beta-strand content. This demonstrates that the predicted mean p(exchange) from Dyna-1 is broad for any given sequence length or helix / strand content. For proteins at the top and bottom range of mean p(exchange), we discarded proteins as candidates if they were fully intrinsically disordered, did not have published expression/purification methods sections, and already had some relaxation data publicly available. From the remaining proteins, we selected Chitinase 19 from *Bryum coronatum*[62] (BMRB: 11466), which had a high degree of p(exchange) (**Fig. 6b**), and

hypothetical protein yjbJ[63] (BMRB entry: 5105), which had a low degree of p(exchange) (**Fig. 6c**). Performing ^{15}N relaxation dispersion CPMG demonstrated that, resembling Dyna-1's predictions, the chitinase did have exchange in region that are crucial for biologic activity—in its active site and the loops known to bind chitin (Figure 6b,e)—while the yjbJ only had detectable exchange for residues 2 and 3 (Fig. 6c,d, supplemental dataset 1). Evaluated at a per-residue level, Dyna-1 obtained an AUROC of 0.62 for chitinase and an AUROC of 0.72 for yjbJ."

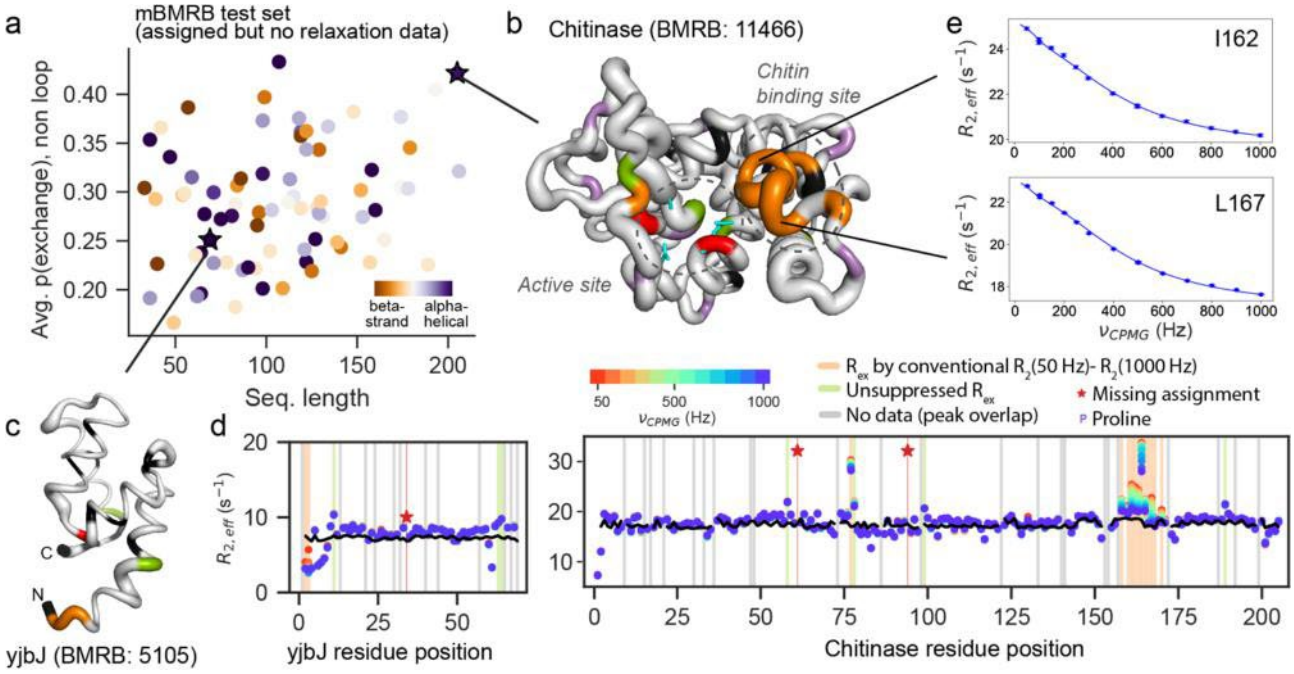


Fig. 6: Dyna-1 has predictive power for enzymatic motions in prospective experimental test. a) Dyna-1 proteins in mBMRB test set vary in net p(exchange) predicted across sequence length and helix/strand content. We selected two proteins to experimentally test Dyna-1 predictions: (b) Chitinase 19 from Bryum coronatum, and (c) hypothetical protein yjbJ. Residue coloring in (b,c,d) is same as in Fig. 5. (d). ^{15}N CPMG relaxation dispersion only showed dispersion curves for a two N-terminal residues in yjbJ, whereas chitinase has extensive exchange in the active site and chitin binding site (see also supplemental dataset 1). (e) Representative ^{15}N CPMG dispersion data for I162 and L167 in the chitin binding site region."

3. We are happy to have already made the relaxation data and the software available for public use in March 2025. Currently, the inference code is on GitHub for any researcher to directly download and run. Additionally, there are available Google Colab notebooks and a HuggingFace Space where researchers can directly run both the ESM-2 and ESM-3 versions of Dyna-1 in their web browser immediately. As stated in the text, we aim to release all remaining training code and datasets upon publication, but have provided the code and data to the reviewers of this manuscript. We have already gotten excellent feedback from the larger community using Dyna-1, and usage is high.

Crucially, non-NMR people, non-structural biologists, and non-computationalists can easily access and use these models. The user only needs to input the sequence of the protein and a structure model! We shared the same sentiment with the reviewer for the impact for the broader community to easily predict protein dynamics!

Furthermore, while we were writing this response, Evolutionary Scale (the authors of the ESM models) announced they will open-source the ESM models upon which Dyna-1 is based. The former license

previously prohibited us from releasing Dyna-1 under a more permissive open-source license, and therefore we are excited to soon also open-source Dyna-1.

HuggingFace Space: https://huggingface.co/spaces/gelnesr/Dyna-1
Google Colab: https://colab.research.google.com/github/WaymentSteeleLab/Dyna-1/blob/main/colab/Dyna_1.ipynb
GitHub: https://github.com/WaymentSteeleLab/Dyna-1

Datasets: https://huggingface.co/datasets/gelnesr/RelaxDB
Model weights: https://huggingface.co/gelnesr/Dyna-1/tree/main

Additional comments:

1. Throughout the manuscript, the authors need to clarify that they are not predicting "micro-to-millisecond timescale motion" but rather predicting which residues exhibit micro-to-millisecond timescale motions as detected by NMR chemical exchange methods. There is an entire universe of micro-to-millisecond timescale motions which will not be detected by these NMR experiments either because the difference in chemical shift or population of the alternative state are too small. The method also does not predict the timescale of the motion or the population of alternative states even if they are detectable by NMR. So again – predicting which residues are NMR active in these experiments, not predicting micro-to-millisecond timescale motion in general.

That was our exact intention in our writing of the original manuscript to communicate these points. We apologize that we may have taken this knowledge of what NMR chemical exchange measures as general knowledge by now (for 30 plus years Rex is being plotted onto protein and RNA structures). We thank the reviewer for this comment, and in response, we edited the entire manuscript to make this point even clearer. To make it very clear from the beginning, we edited the abstract (new text in bold)

Strikingly, these models also predict µs ms motion **exchange** measured via NMR relaxation experiments, **indicative of µs-ms dynamics**.

2. The authors should list and examine all other plausible reasons for why a resonance might be missing from the NMR spectrum. For example, a resonance might be overlapped with another resonance. This possibility could be directly tested by predicting chemical shifts from the 3D structure and seeing whether the resonance falls in a crowded region of the spectrum.

We had carefully considered possible reasons for a missing resonance assignment and excluded all data in the training for which systematic reasons, other than exchange broadening, would result in missing assignments (deuteration, proteins with more than 12 missing in a row, unassigned N- and C-termini). These are described in the paper, in detail in the methods section. Overlapping peaks do not fall in this category for folded proteins, since it is next to impossible that the NH, Ca, Cb, and CO are all identical for a residue. Therefore, even for overlapped peaks in a ^{15}N ^{1}H HSQC, the residue will have an assignment if not severe R_{ex} prohibits either the peak not to be there at all, or relaxes during the INEPT transfer in the 3D experiments.

There are other reasons resonances might be missing besides the plausible reasons we described in the paper, but if these errors are random, as opposed to systemic, as for instance assignment errors would be, this noise could actually enhance training. We have added a line to comment on this:

"We note that while there will doubtlessly also be noise in these labels arising from mis-assignment or other errors, there is a well-established literature showing that in certain regimes of machine learning, noise can actually assist model convergence as a form of regularization[27,28]."

In addition, a resonance might be missing not due to 15N Rex broadening but rather due to broadening of the amide proton. Does this complicate the integration of data sets specifically measuring 15N broadening versus 1H?

Our model was trained using labels only where the N had no assignment. A missing N assignment could be due to exchange experienced by either the N or attached amide H (as pointed out by the reviewer), as both would lead to magnetization loss in the INEPT transfer steps. This would not have a negative effect on our model.

The reviewer's question led us to realize something we had not previously considered: for our experimental validation in which we compare the Dyna-1 predictions with 15N relaxation data, the AUROC might actually even be underestimated for cases where Rex is only seen for 1H due to chemical shift differences only in 1H, not 15N. Hence, it could be that the model would perform even better than reported. Thank you for this question! We added a sentence about this scenario:

"We note that because our RelaxDB-CPMG dataset is curated with ^{15}N relaxation data, our evaluations on the experimental data do not consider cases where R_{ex} is only seen for ^{1}H caused by chemical shift differences only in ^{1}H and not ^{15}N, therefore underestimating other possible chemical exchange. Hence, Dyna-1 performance might be better than reported."

3. A related comment: the Rex contribution depends in a complex manner on the population of the alternative conformation as well as the difference in chemical shift and exchange kinetics and whether RF is being applied or not. Thus, in each R1rho experiment, for example, whether one detects Rex or not crucially depends on the chosen spin lock powers and offsets, or if using CPMG, on the range of tauCP values employed. Thus, depending on the experimental setup, one may or may not detect Rex; was anything done to control for these variables?

Yes, this is exactly correct. Therefore, we want to reiterate that we used missing assignments for ~2700 proteins, and not Rex, to train our Dyna-1 model, hence avoiding technical settings used by different authors. We then used the 133 proteins for which actual relaxation data were reported (RelaxDB benchmark), and 10 proteins where Rex was measured by CPMG to evaluate the performance of our model. Rex data were never used in training the model.

We agree that the tauCP range in CPMG may or may not detect Rex. A major point of Figure 5 was that if we carefully considered residues with R_{2eff} values that did not decay, but remained elevated, the overall performance of Dyna-1 increased, indicating that Dyna-1 had predictive power for residues with exchange that conventional analysis would have missed.

In summary, these dependencies of Rex on these factors only play a role in our experimental validation, i.e. what the AUROC is, and not the new model itself. As described above, one might miss Rex in the experimental data of the 134 proteins for the reasons described by the reviewer, which could possibly lead to an increase in AUROC.

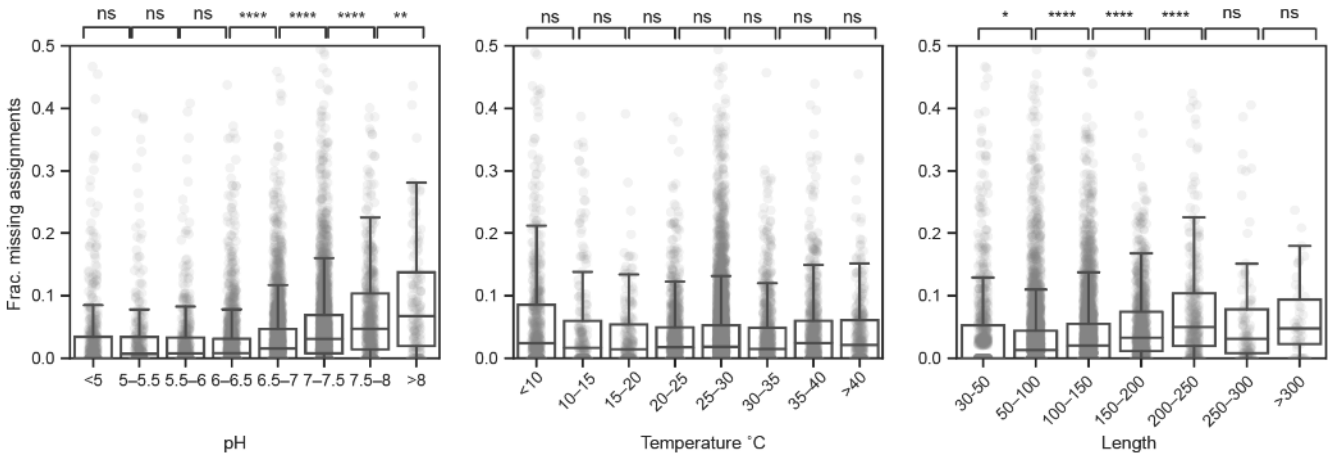
4. The authors can be more quantitative about the criteria that need to be satisfied for a peak to be missing from the NMR spectrum. For example, a resonance with high R2 due to a high molecular weight protein or anisotropic diffusion will require a smaller Rex contribution before it falls below the detection threshold. Thus a resonance could be missing for the same Rex from a large protein but not in a small one. This is likely to confound the analysis. Perhaps the analysis could be done by binning different molecular weight proteins. In addition, a peak might be missing on one spectrometer but not another, either due to differences in sensitivity or due to differences in Rex when working at different fields. It would be good to explain how this aspect of the analysis was handled.

Our criteria for if a peak is missing from an NMR spectra is simply for each residue if there was a chemical shift deposited in the BMRB (assigned) or not (not assigned i.e missing).

Thanks to the comments by the reviewer, we edited the text to make it even clearer why we thought it could be reasonable to use missing assignments as a source of data for ms exchange:

“Of course, residues that are completely exchange-broadened represent a subset of all residues with exchange. If we use residue counts in RelaxDB as a rough proxy for the relative abundances of detectable Rex versus missing assignments, we observe that 5% of residues have elevated R2/R1, whereas 4% have missing assignments (Figure 1d). This rough estimate concludes that nearly half of the residues with exchange have missing assignments, which certainly is not all, but also not negligible.
We curated a new dataset: the mBMRB. In this dataset, we removed proteins that might have systematic, non-exchange reasons for assignments being missing. ... This dataset comprises “positives” which are missing assignments curated as best as possible, but of course the “negatives" in this dataset are not unambiguously negatives: many residues are assignable while also experiencing exchange.”

Second, we have newly included statistics on three parameters that might affect the prevalence of missing assignments: temperature, pH, and sequence length in Extended Data Fig. 3d:



We note that the fraction of missing assignments does increase with sequence length, but data is much sparser for proteins over length 250 and it is difficult to conclude how we might actionably use this to improve the model in its current architecture. Such improvements are the focus of future work.

5. On p 5, the authors rightly point out that a residue may have a ‘normal’ R2 value despite competing slow and fast motions canceling their contributions. The authors state that if anything R2/R1 is most likely underestimating Rex; what was the basis for this claim?

For the 133 proteins (RelaxDB) for which we only have R_2, R_1, and het NOE as experimental marker to evaluate the performance of Dyna-1, we can only use the elevated R_2/R_1 ratio as experimental measure for us-ms exchange, as us-ms exchange leads to an increase in R_2 when compared to that from overall anisotropic or isotropic tumbling. Crucially, ps- low ns motion can lead to a decrease in R_2, hence the opposite effect. Therefore, it is highly likely there will be a number of residues with such cancellation in the experimental data. For those residues for which Dyna-1 would correctly predict us-ms motions, the experimental label would be incorrectly labeling them as "silent", hence driving the AUROC down. We extended our explanation in the text (new text in bold):

"As expected, many residues with **low hetNOE have a dR_2/R_1 less than zero, as motions at the ps-ns timescale lead to both a lower hetNOE and a reduced R_2 value**. A few residues have both fast motions detected via hetNOE and elevated dR_2/R_1, indicative of µs-ms exchange. If anything, we anticipate that this method of looking at elevated R_2/R_1 is under-estimating exchange for residues where motions are present at more than one timescale, **as the presence of µs-ms exchange in elevating R_2 and the presence of ps-ns motion decreasing R_2 would cancel."**

6. I found the analysis presented on pages 10/11 to be difficult to follow and highly technical without driving the key points with compelling statistics backing the claims being made.

We thank the reviewer for this feedback as it helped us to revise this whole section for the general readership! In the revised manuscript we have condensed how we describe results in these two pages, reduced technical jargon and better emphasize and explain key concepts such as transfer learning. Representative updated text below:

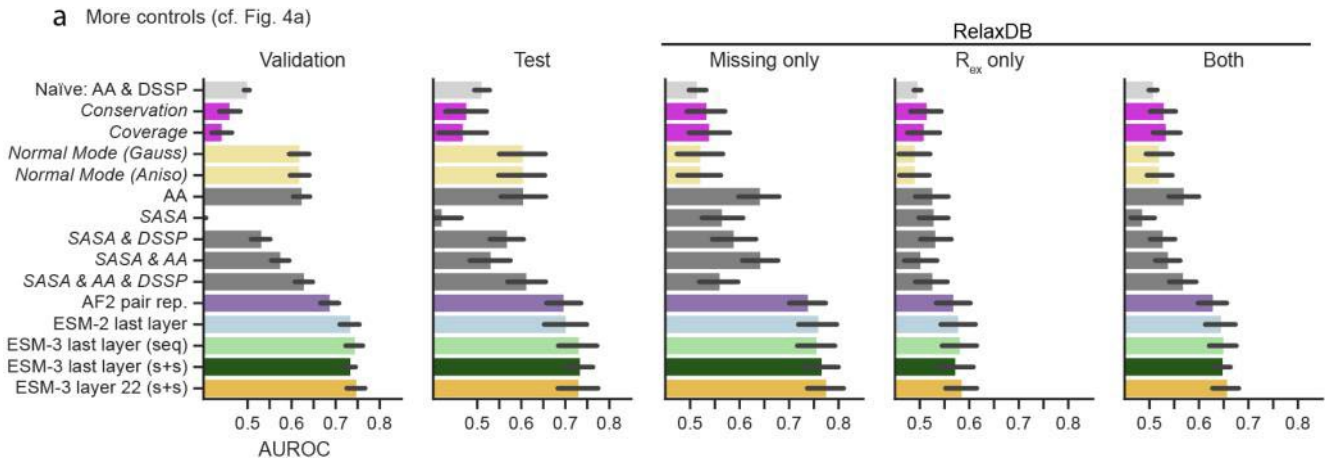
"We were curious to explore the predictive power of existing models on this task, since models such as AF2[6] and ESM-2[34] have shown predictive capabilities beyond the tasks for which they were trained. For example, the protein language model ESM-2 was trained in the self-supervised task of masked language modeling yet has predictive power for secondary structure, 3D structure, and functional labels. In this "transfer learning" paradigm, extracted representations from pretrained models are used to train new models with different datasets on related downstream tasks[30]. We designed architectures to test structure-focused representations from AF2 pair representations ("AF2, pair rep."), sequence-only representations in ESM-2, and interactions between sequence and structure by using the multimodal language model ESM-3 (see Methods, Fig. 3b, Extended Data Fig. 4). We first compared these architectures by training with our curated NMR dataset (mBMRB) at 80% sequence identity cutoff; after assessing the models' predictive capability on this novel task given a majority of the available data, we then retrained models with a more stringently filtered training set (see below).
We observed that all three of these pretrained models out-performed other baseline models (Fig. 3c). The best-performing model was layer 22 of ESM-3 if provided both sequence and structure models as input, which achieved an AUROC of 0.77. Intriguingly though, the last layer of ESM-2 model achieved an AUROC of 0.75, and the AF2-pair model, which did not use MSAs, and rather only structure input, achieved 0.71. This suggests that both sequence and structure information provide utility in this new task."

7. It would be helpful to include comparisons to pre-existing models that predict protein dynamics like DynaMine or elastic models. Especially since the AUROC scores for Dyna-1 are not that high, usually ranging around 0.6-0.7, this comparison may provide stronger evidence that Dyna-1 is a step forward in predicting protein dynamics.

Based on the reviewer's suggestion, we have added studies evaluating the ability of fluctuations from elastic network models to predict missing assignments–in particular an Anisotropic Normal Mode model and Gaussian Normal Mode model. In addition, we also compute alternative, simpler models as baselines. These new results further strengthen the original conclusion that training with actual NMR dynamics data

for us-ms motion results in a new model with superior performance over any alternate, simpler predictors. We found the additional comparison of elastic network models to be insightful, and have added this figure in **Extended Data Figure 6a.**

We did not evaluate DynaMine as it is designed to predict backbone dynamics on the picosecond-nanosecond (ps-ns) timescale, whereas Dyna-1 specifically targets microsecond-millisecond (µs-ms) conformational exchange, representing a fundamentally different dynamical regime. Note that ps-ns dynamics can be quickly and directly sampled these days with MD simulations, negating the need for a predictive model for this time-scale.



Extended Data Figure 6. (a) Validation, test, and RelaxDB datasets comparing further benchmarks at training cutoff of 30% sequence identity, TMscore 0.5. Benchmark models in italics are not in Fig. 4a.

In Methods:

Baseline training. Our baseline training aimed to test simpler representations of protein sequences and/or structure. We use the mBMRB data to train these methods. The results of these baseline models on the mBMRB-Validation, mBMRB-Test, and RelaxDB datasets are depicted in Extended Data Fig. 6a. None outperform the models based on pre-trained models.

Naïve baselines: The simple, naïve baseline uses amino acid frequencies in the mBMRB training data as weighting of a random classifier.

- The simplest is the AA baseline, which computes the frequency of each amino acid being missing from the mBMRB training data. We then predict for our validation and test set whether a residue is missing at random, weighted by that relative frequency.
- We evaluate similar random classifier using frequency of secondary structure (loop, sheet, helix) as calculated by DSSP, and a combination termed the "AA & DSSP" baseline using the frequency of each amino acid type given its secondary structures (loop, sheet, helix).

Simple classifiers: We use a similar Transformer classifier as that described for the final Dyna-1 architecture trained on simpler representations of the input mBMRB training data. Each representation is described.

- "OHE AA": we input each protein as a one-hot encoded sequence, where each residue is represented by a unique combination of 0s and 1s, into the transformer classifier.
- "SASA" we evaluate solvent-accessible surface area using the Shrake-Rupley method[69] implemented in MDTraj[70]. These values are encoded by a simple one-dimensional MLP of hidden size 128 and then input into the transformer classifier.

- The “SASA & DSSP”, “SASA & AA”, and “SASA & AA & DSSP” concatenate the outputs of the MLP with a one-hot encoding of the secondary structure, sequence, and secondary structure and sequence, respectively. This concatenated embedding is subsequently used as input into a transformer and trained accordingly.

Other representations:

- Normal mode analysis (NMA) is a computational technique that has long been of interest for modelling dynamic processes in proteins[78]. To evaluate NMA, we use ProDy[79] to compute the top 20 normal modes of the models generated using ESMFold (for the mBMRB data) or AF2 (for the RelaxDB data) using the Anisotropic Network Model (ANM) and Gaussian Network Model (GNM) and build the Hessian matrix and Kirchhoff matrix, respectively. We then calculate the mean square fluctuation and evaluate the AUROC and AUPRC using these values. No model is trained.
- To use sequence conservation as a baseline, we computed sequence conservation scores and coverage index per protein as described in ref. [68] and evaluated the AUROC and AUPRC using these values as logits. No model is trained.

8. The authors may want to point out that resonances such as adenine-C2 are often missing from HSQC spectra in RNA due to line broadening caused by protonation of adenine-N1, so that this approach could hold promise in applications to nucleic acids as well.

The reviewer is on point that the same idea of using missing peaks in RNA spectra can be used, and in fact we hope that this idea and our publication inspires the same approach for predicting RNA dynamics!

Minor comments:

Fig. 1c: It would be beneficial to include quantitative metrics on how well HYDRONMR predictions agree with the experimental data, particularly since the proposed model relies on values derived from these predictions.

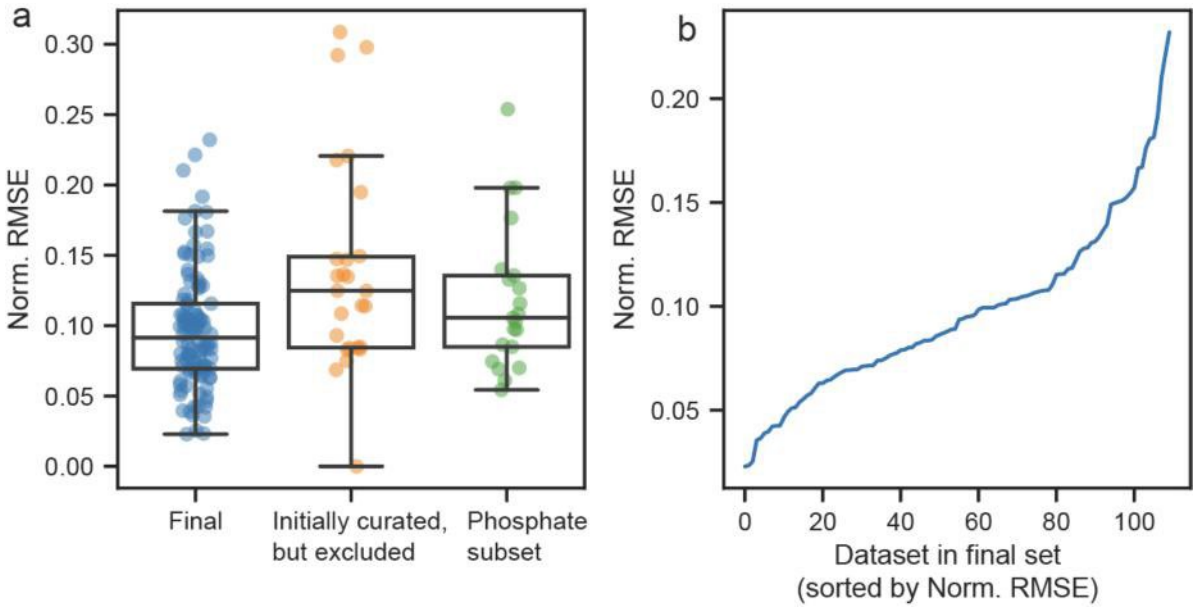
We thank the reviewer for this feedback. In this revision we have made our HYDRONMR-based fitting significantly more quantitative.

We realized in revisiting our fitting that the fit parameter – the ratio of tumbling time between experimental and calculated R_2/R_1 – has a closed-form solution, whereas we were previously iteratively fitting this. We have also increased the rigor with which we do outlier-calling to infer where there is elevated R_2/R_1: previously we were using any value greater than one standard deviation above the mean of dR_2/R_1, but we realized this was skewed by outliers. We instead now estimate model uncertainty using interquartile range (more robust to outliers), and call any residue where R_2/R_1 is greater than the combined experimental error and model uncertainty.

Hence, our RelaxDB labels have changed slightly, yet our conclusions remain the same. Our overall AUROC values for RelaxDB have remained largely the same or even gone up slightly in some cases (previously: average over whole dataset not including phosphate buffer set was AUROC of 0.65, now it is 0.66).

We now include a normalized RMSE to evaluate fit for residues without dynamics labels. Summary statistics of this in the final dataset are included in Extended Data Figure 2. Norm RMSE is also included

for each dataset in Extended Data Fig. 1 now.



Extended Data Figure 2: (a) RMSE normalized to median R_2/R_1 per protein for the final RelaxDB dataset, protein datasets that were initially curated but excluded, and proteins with biological function involving phosphate binding, but measured in phosphate buffer. (b) Norm. RMSE distribution of final RelaxDB dataset used for evaluation.

From Methods (pg. 40): "In the final dataset, 65/111 datasets had a norm. RMSE of less than 0.1, and 107/111 had a norm. RMSE of less than 0.2 (Extended Data Fig. 2b). A norm. RMSE of 0.1 can be interpreted as the experimental data for residues that did not have dynamics assigned were on average within 10% of the predicted HYDRONMR value for R_2/R_1."

The criteria used to define different scales of motion are also unclear. Are these definitions based on previously established thresholds in the field?

Yes, they are. Regarding designation for ps-ns motion: We have added 3 citations dating back to 1999 that use a hetNOE cutoff of 0.65 for folded proteins in their data analysis:

"We also assigned a label for residues with ps-ns motion as residues with hetNOE < 0.65 as used in refs. 65-67 to designate residues with ps-ns motion."

[51] Raza, T. et al. Insights into the NF-κB-DNA Interaction through NMR Spectroscopy. ACS Omega 6, 12877-12886 (2021). https://doi.org:10.1021/acsomega.1c01299

[52] Korn, S. M. et al. (1)H, (13)C, and (15)N backbone chemical shift assignments of the nucleic acid-binding domain of SARS-CoV-2 non-structural protein 3e. Biomol NMR Assign 14, 329-333 (2020). https://doi.org:10.1007/s12104-020-09971-6

[53] Johnson, E. C., Lazar, G. A., Desjarlais, J. R. & Handel, T. M. Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin. Structure 7, 967-976 (1999). https://doi.org:https://doi.org/10.1016/S0969-2126(99)80123-3

For µs-ms motion, we have discussed above how we have substantially revisited our fit to reflect best statistical practices on our part. We hope that by making our data publicly available, the field will devise more standard model-agnostic ways to interpret data at scale, which is what is needed for machine learning efforts.

Fig. 1d: The grey distribution in Fig. 1d is not labeled. Does it correspond to residues without observed exchange?

We have updated Fig 1d to include a label for the grey distribution.

Fig. 1d: The sequence conservation distributions appear broad and overlapping. The assertion that residues exhibiting µs–ms motion are "highly conserved" is insufficiently supported based on this data.

We agree with the reviewer, we have updated this text to say “more conserved”.

Fig. 4a: This subfigure is difficult to follow and could be better explained in the manuscript/figure description.

Yes, we have removed the cartoons in Fig. 4a. and revised how it is discussed.

Ext. Data Fig. 6b: Based on the data presented in Ext. Data Fig. 6b, it appears that Dyna-1 and other models perform only modestly, with AUROC values below 0.6 in predicting exchange residues. This should be acknowledged more directly and factored into the interpretation of the results.

This is indeed an important point going back to the question we had asked earlier, given that Dyna-1 is trained on missing assignments, can it 1. predict missing assignments, and 2. also predict residues with an assignment but having measurable Rex? This figure shows that unsurprisingly it performs best at missing assignments, but strikingly also has decent predictive power for assigned residues with Rex, but lower AUROC for the latter category. Thanks to the suggestion we edited the text to emphasize this result more:

“To test if the model indeed generalized beyond the data with which it was trained, we held out missing assignments and asked how well it could predict which assigned residues have exchange (Fig. 4aii). **Performance does decrease across all models compared to predicting missing assignments, but** AF2-pair, ESM-2, and ESM-3-based models achieved AUROC above the controls, indicating that they can generalize learning from predicting missing assignments (e.g. p(missing)) to predicting exchange within assigned residues (e.g. p(exchange))”.

We have also made AUROC the metric shown in the main text to match the other metrics.

We want to emphasize that this question of whether Dyna-1 has predictive power for assigned residues with R_{ex} was an important one in this study and we did not know if our model can generalize. This result of higher AUROC than baselines, although modest, is a key result. Clearly this opens the next improvement in the future for Dyna-2, a way to find enough experimental data for including residues with Rex in training the next model.

Referee #2:

Generally this is a good paper that explores ways of capturing protein dynamics in ML models, offering downstream improvements in speed, accuracy as well as interpretation advantages.

The paper is poorly written, with many avoidable typos, and several poorly supported statements. The paper, with more attention to writing, could develop the main points more carefully. The paper is also poorly developed with respect to connection to prior results and alternate approaches.

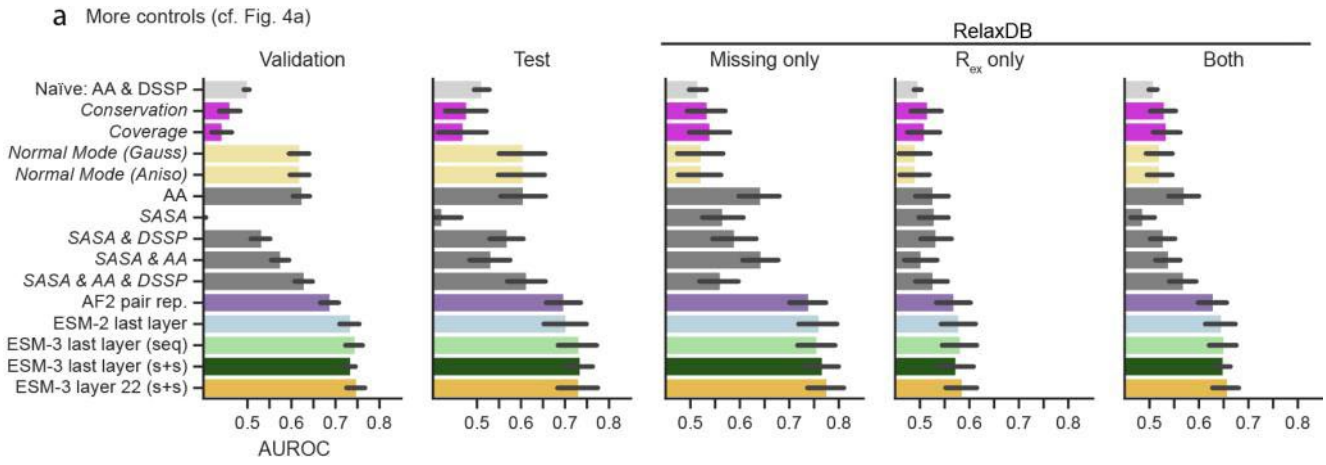
We thank the reviewer for the general endorsement of our manuscript, and comments to edit the writing. These suggestions made the vastly revised manuscript much better, and the suggested new experiments and analysis further supported the main points and conclusions. We thank the reviewer for the constructive pointers!
In the revised manuscript we majorly edited the text to make the key points much clearer to readers from the ML field, the NMR and structural biology fields. In this exciting new area where experiments meet AI, we need to find best ways to communicate between these fields, and from comments from the NMR expert and the expert in deep learning it is apparent that major achievements in this manuscript were not well enough communicated.

This paper presents a novel task: can a model predict which residues are exchange-broadened or have measurable Rex, two observables from NMR that indicate us-ms dynamics. Existing tasks that could get confused with this task are 1) predicting backbone flexibility, a catch-all for faster ps-ns motions, and 2) predicting multiple conformations (with no information on timescales). Predicting protein dynamics is a frontier for molecular biology, and part of tackling that frontier is clarifying these different problems. To explain from the start the fundamental connections and differences to prior results and alternate approaches, we have edited our introductory text and included references to other works that have aimed to tackle those other tasks for readers unfamiliar with these distinctions:

“This task: namely, can a model predict the presence of a concerted, millisecond-timescale process, is distinct from other existing tasks: predicting backbone flexibility or order parameters[13,14]; or predicting multiple conformations without information on timescales[15,16].”

I would imagine that simple alternate predictors, such as number of contacts, surface accessible surface area and secondary structure could combine to make predictors of the dynamic quantities you measure. You show some of these metrics individually, but do not show how these could be composed into a very simple predictor. Please make comparison to other methods and simple metrics more prominent to help readers interpret you main findings and models. You test AF2-pair, ESM-2, and ESM-3-based models (billions of parameters to predict a single number per residue), but not much simpler ones that might work well.

This is a good suggestion, and we have added more baselines testing such simpler models, described in a new section under Methods (see below). None of these models reached the performance obtained by our final model Dyna-1. These new results further strengthen the original conclusion that training with actual NMR dynamics data for us-ms motion results in a new model with far superior performance over alternate, simpler predictors. We found these baselines insightful and thank the reviewer for suggesting to perform such comparative analysis. These new analysis are added to the **Extended Data Figure 6a** (new baselines are in italics).



Extended Data Figure 6. (a) Validation, test, and RelaxDB datasets comparing further benchmarks at training cutoff of 30% sequence identity, TMscore 0.5. Benchmark models in italics are not in Fig. 4a.

In Methods:

Baseline training. Our baseline training aimed to test simpler representations of protein sequences and/or structure. We use the mBMRB data to train these methods. The results of these baseline models on the mBMRB-Validation, mBMRB-Test, and RelaxDB datasets are depicted in Extended Data Fig. 6a. None outperform the models based on pre-trained models.

Naïve baselines: The simple, naïve baseline uses amino acid frequencies in the mBMRB training data as weighting of a random classifier.

- The simplest is the AA baseline, which computes the frequency of each amino acid being missing from the mBMRB training data. We then predict for our validation and test set whether a residue is missing at random, weighted by that relative frequency.
- We evaluate similar random classifier using frequency of secondary structure (loop, sheet, helix) as calculated by DSSP, and a combination termed the "AA & DSSP" baseline using the frequency of each amino acid type given its secondary structures (loop, sheet, helix).

Simple classifiers: We use a similar Transformer classifier as that described for the final Dyna-1 architecture trained on simpler representations of the input mBMRB training data. Each representation is described.

- "OHE AA": we input each protein as a one-hot encoded sequence, where each residue is represented by a unique combination of 0s and 1s, into the transformer classifier.
- "SASA" we evaluate solvent-accessible surface area using the Shrake-Rupley method[69] implemented in MDTraj[70]. These values are encoded by a simple one-dimensional MLP of hidden size 128 and then input into the transformer classifier.
- The "SASA & DSSP", "SASA & AA", and "SASA & AA & DSSP" concatenate the outputs of the MLP with a one-hot encoding of the secondary structure, sequence, and secondary structure and sequence, respectively. This concatenated embedding is subsequently used as input into a transformer and trained accordingly.

Other representations:

- Normal mode analysis (NMA) is a computational technique that has long been of interest for modelling dynamic processes in proteins[78]. To evaluate NMA, we use ProDy[79] to compute the top 20 normal modes of the models generated using ESMFold (for the mBMRB data) or AF2 (for the RelaxDB data) using the Anisotropic Network Model (ANM) and Gaussian Network Model (GNM)

and build the Hessian matrix and Kirchhoff matrix, respectively. We then calculate the mean square fluctuation and evaluate the AUROC and AUPRC using these values. No model is trained.

- To use sequence conservation as a baseline, we computed sequence conservation scores and coverage index per protein as described in ref. [68] and evaluated the AUROC and AUPRC using these values as logits. No model is trained.

The data presented shows that you have an ability to predict some dynamics related quantities from NMR data, but the paper does not really illustrate/support how this code could be used in practical settings. How do you sample residue conformations from this model, are they biophysically accurate/plausible, how do you differentiate error (of either kind) from the biophysically relevant dynamics you focus on?

The practical settings for using this code is that the user only inputs the sequence and structural model of the protein, and Dyna-1 predicts on a per-residue basis the probability of us-ms motions plotted onto the structural model. The code is implemented into a HuggingFace Space, making it very easy to use for anybody. As described in the manuscript, Dyna-1 predicts where us-ms motions occur, but does not sample 3D conformations. The latter task would clearly be the ultimate goal delivering the full free energy landscape for any protein, from which the field is still quite far away.

Importantly, we have now performed a novel prospective experimental test which demonstrates the practical use of Dyna-1 in its intended purpose of predicting the presence of micro-millisecond exchange:

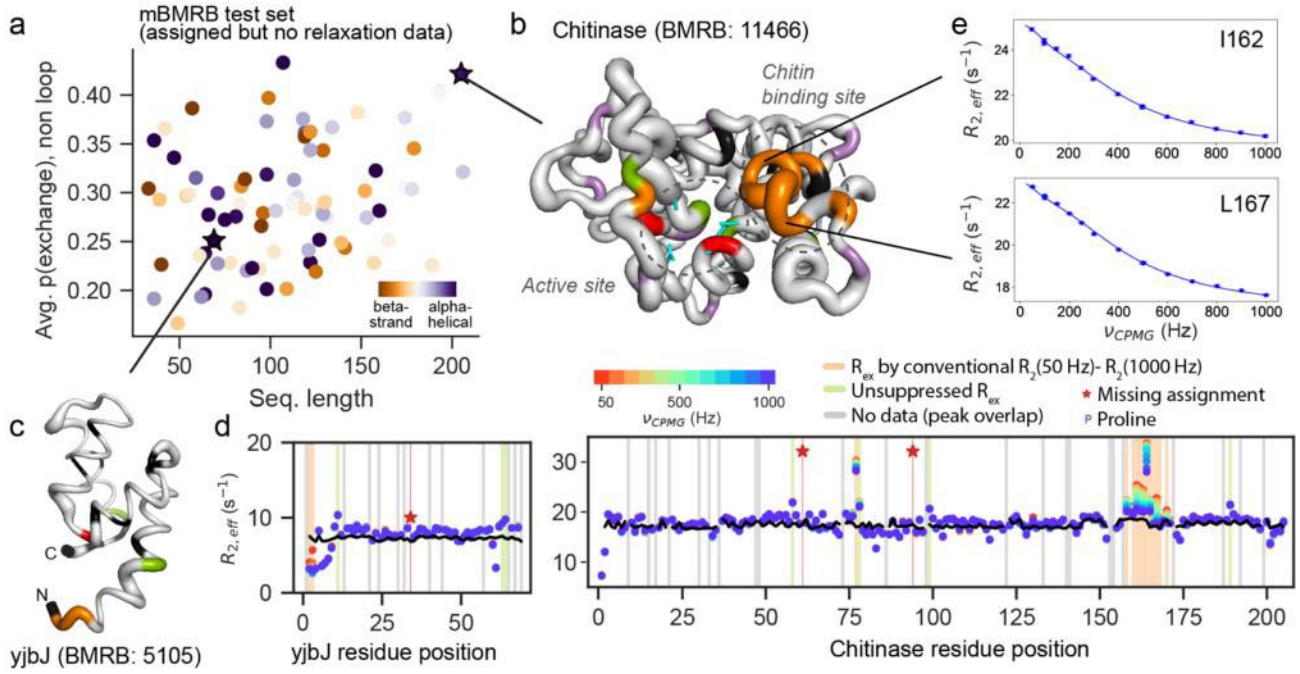


Fig. 6: Dyna-1 has predictive power for enzymatic motions in prospective experimental test. a) Dyna-1 proteins in mBMRB test set vary in net p(exchange) predicted across sequence length and helix/strand content. We selected two proteins to experimentally test Dyna-1 predictions: (b) Chitinase 19 from Bryum coronatum, and (c) hypothetical protein yjbJ. Residue coloring in (b,c,d) is same as in Fig. 5. (d). ^{15}N CPMG relaxation dispersion only showed dispersion curves for a two N-terminal residues in yjbJ, whereas chitinase has extensive exchange in the active site and chitin binding site (see also supplemental dataset 1). (e) Representative ^{15}N CPMG dispersion data for I162 and L167 in the chitin binding site region.

New text below:

Prospective experimental test of Dyna-1.

To validate the predictive capabilities of Dyna-1, we conducted a prospective experimental test after public release of the model and initial review of the manuscript. We selected proteins from the "mBMRB test" set to characterize their dynamics, as these are proteins that are already assigned but many do not have published relaxation data. **Fig. 6a** depicts the mean p(exchange) from each protein in the mBMRB test set vs. sequence length, plotted by alpha helical / beta-strand content. This demonstrates that the predicted mean p(exchange) from Dyna-1 is broad for any given sequence length or helix / strand content. For proteins at the top and bottom range of mean p(exchange), we discarded proteins as candidates if they were fully intrinsically disordered, did not have published expression/purification methods sections, and already had some relaxation data publicly available. From the remaining proteins, we selected Chitinase 19 from *Bryum coronatum*[62] (BMRB: 11466), which had a high degree of p(exchange) (**Fig. 6b**), and hypothetical protein yjbJ[63] (BMRB entry: 5105), which had a low degree of p(exchange) (**Fig. 6c**). Performing ^{15}N relaxation dispersion CPMG demonstrated that, resembling Dyna-1's predictions, the chitinase did have exchange in region that are crucial for biologic activity—in its active site and the loops known to bind chitin (Figure 6b,e)—while the yjbJ only had detectable exchange for residues 2 and 3 (Fig. 6c,d, supplemental dataset 1). Evaluated at a per-residue level, Dyna-1 obtained an AUROC of 0.62 for chitinase and an AUROC of 0.72 for yjbJ.

We note that we have already gotten excellent feedback from the larger community using Dyna-1 for hypothesis generation. While we were writing this response, Evolutionary Scale (the authors of the ESM models) announced they will open-source the ESM models upon which Dyna-1 is based. This previously prohibited us from releasing it under a more permissive open-source license, and therefore we are excited to soon also open-source Dyna-1 to make it more practically useful.

What other NMR methods or dynamics methods would work here? J-couplings from NMR, HD exchange from NMR and/or other measures that convolve accessibility and dynamics could be used as validation vignettes or test cases.

We chose R_{ex} since this is the **only** direct experimental measure for us-ms motion. J couplings contain no time-domain information, and HD exchange is an indirect method since several processes need to happen: (i) The amide proton needs to be solvent exposed in one of the conformations (which is often not the case for ms interconversion between two folded states, (ii) and a deuterated hydroxyl, OD^-, needs to react with the amide proton faster than the protein dynamics process of interest).

NMR relaxation (R_2, R_{ex}) is currently the only experimental method for which us-ms motions can be measured at atomic resolution in solution, at equilibrium conditions, and for which enough data are available. This is why we sought to curate as many sets of Rex measurements as we could for model testing. The only other method that is currently being developed that could monitor these dynamics might be time-resolved quenched cryo-EM. Here, a reaction needs to be initiated, often by photoactivation, and then the time evolution to equilibration is measured by quenching the reaction at different time points, freezing the sample on the grid, and then measuring.

Using unassigned residues, and finding ways to bring these missing data back into your analysis is a great idea.

We are excited to see that the reviewer highlights and appreciates the key idea and foundation of how to train this new dynamics model with the most powerful NMR experimental data.

Referee #3:

Summary of the key results:

You describe a method to predict us-ms motions in proteins based on their sequence, on the assumption that peaks missing in 15N spectra indicate such dynamics. This assumption is related to (sparse) relaxation data.

We respectfully disagree that our statements regarding our model's performance are based on "sparse" relaxation data. First, we trained our model with 3000 proteins using a sequence cutoff of 30%, Second, to evaluate our model, we curated an order of magnitude more R1/R2/hetNOE datasets than have been available in one place; 133 proteins, as well as 10 proteins for which we could find CPMG data. Though we very much wish for more data on frontier problems such as this one, benchmarks of this size are used routinely as evaluations in other contexts. A widely cited paper interpreting AF2 (Roney, Ovchinnikov PRL 2022) uses a benchmark of 133 proteins. A new method for fitting x-ray ensembles (Wankowicz ... Fraser eLife 2024) demonstrated their method with a test set of 144 proteins, to just mention two examples.

Originality and significance:

Whilst this is certainly an interesting take on the dynamics problem, some of the concepts you claim to introduce are not new. Consider for example the dynasome, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0033931, where the relation between dynamics and function was already explored and introduced. You should acknowledge this seminal work.

This citation (dynasome) performs molecular dynamics on systems only up to 80 ns, a time scale we clearly described to be well covered by direct MD simulations, and orders of magnitude different to our novel task of predicting us-ms motions where no predictive models exist, and which cannot be sampled by brute force unbiased all atom MD simulations. This 2018 paper is also by no means the first paper to show that function is related to dynamics. We have added a reference in the introduction to the truly seminal work by the pioneer in this field Frauenfelder which is an earlier work (1991) establishing that function emerges from dynamics.

"The functions of proteins often depend on their ability to interconvert between multiple conformations[3]".

[3] Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The energy landscapes and motions of proteins. Science 254, 1598-1603 (1991). https://doi.org:10.1126/science.1749933

In terms of significance, there are major issues with your dataset, described below, which make it impossible to assess whether your method works properly and is generic.

We have responded to the misinterpretation and concerns about the dataset below.
In addition, we also performed additional NMR experiments on two proteins for which **no** dynamics data were ever recorded and no homologues in our training dataset to deliver novel experimental verification, demonstrate generalizability, and show its practical application.

Data & methodology:
- The protein environment can have a very large influence on its behavior, for example phosphorylation can switch a disordered region to a folded one. You do not really address this aspect in relation to your method

- dynamics is after all determined by the energy landscape of the protein, which is variable. How persistent are these regions you predict, and how dependent on your filtering of the BMRB dataset?

Key concepts of energy landscapes that have been defined and characterized over the last 40 years are as follows: protein environments such as pH, temperature, agents in solution, or even phosphorylation only shift equilibria (change relative populations) and alter the rate constants of interconversion, meaning that the same residues are still dynamic on this time-scale independent of changes in pH, temperature etc. That is the foundation of the accepted energy landscape concept.

The reviewer's first statement of phosphorylation switching a disordered region into a folded one demonstrates a misunderstanding of this concept, and hence our data and methodology: Phosphorylation shifts the relative populations between the disordered state and the folded conformation and **not,** as described by the reviewer, that the non-phosphorylated protein only samples the disordered state and the phosphorylated protein only samples the folded state.

We have added text in the introduction to clarify this task and the fact that pH and temperature would not alter the presence/absence of a dynamic process.

"This task: namely, can a model predict the presence of a concerted, millisecond-timescale process, is distinct from other existing tasks: predicting backbone flexibility or order parameters[13,14]; or predicting multiple conformations without information on timescales[15,16]. External conditions such as temperature, pH, and cofactors will change the rate and relative populations of such a process, but not the underlying process itself."

You only included monomeric datasets with no cofactors present. Did you also check for pH, temperature, possible other agents in the solution (e.g. urea) and such, as these can have a large influence on protein behavior?

We included monomeric datasets with no cofactors for curating RelaxDB, the main validation set that our claim for predicting millisecond motion is contingent upon, as we wanted to ascertain as much as possible that those datasets contained exchange only emanating from intramolecular millisecond motions as opposed to multimerization or ligand binding. We also excluded datasets from RelaxDB if they were performed under denaturing conditions for precisely the reviewer's reasoning. See our answer above for pH and temperature.

- Your assumption is that missing 15N signals are due to us-ms motions, but which proportion of these is likely missing due to signal overlap, which would in any case be more common in loops?

It is highly unlikely that missing ^{15}N assignments are due to signal overlap, since it is almost impossible that the NH, Ca, Cb, and CO are all identical for different residues. Therefore, even for overlapped peaks in a ^{15}N ^{1}H HSQC, the residue will have an assignment if not severe R_{ex} prohibits either the peak to be there at all, or relaxes during the INEPT transfer in the 3D experiments.

Also here, for intermediate exchange, the field strength of the NMR spectrometer does play a role, please comment on this aspect.

Yes, this is well-known–that the value R_{ex} depends on the difference in chemical shift in Hz and hence on the NMR spectrometer field strength. Therefore we devised a method to compare experimental R2/R1 datasets to R2/R1 values arising from rigid tumbling only and statistically determine residues with increased experimental R_2/R_1 ratios, see Fig. 1 and our updates described below making this fitting more rigorous.

- You have a bias towards residues loops in your dataset, as mentioned on line 199-200 - this also makes sense from an intermediate exchange perspective (see also previous point).

The reviewer is referring to the observation that loops have more peaks missing as a “bias” – it is not possible to say if this is “bias” or the fact that loops have on average more ms motions than other parts.

Instead of many individual, personally picked protein cases can you present the statistics of these data and how this might affect the performance measures of your method in relation to your test data? How long are these loops? How long are the helices? You have to more extensively describe the dataset and explore the possible bias therein.

We are very puzzled by this comment. We do not personally pick protein cases in the manuscript, we report the AUROC for every protein in the test sets: in the RelaxDB (133 proteins), and Relax CPMG (10 proteins). The statistics are reported in the AUROC and AUPRC plots in Fig. 3c and 4b and 5d. We also already presented AUROC as binned by secondary structure type and sequence conservation, in Figs. 4d and f.

We do not understand the question of “how long are these loops, how long are the helices? “ The loops and the helices in the proteins have different lengths; also lengths of loops and helices have no relation to millisecond motions in proteins.

- What is your reasoning to employ a 0.65 cutoff for the hetNOE? And how does this precisely statistically relate to the ‘missing peaks’?

We chose a generally accepted conservative number for the hetNOE in folded proteins as marker for extensive ps-low ns motions. We have now added 3 citations dating back to 1999 that use this cutoff in their data analysis:

“We also assigned a label for residues with ps-ns motion as residues with hetNOE < 0.65 as used in refs. 65-67 to designate residues with ps-ns motion.”

[51] Raza, T. et al. Insights into the NF-κB-DNA Interaction through NMR Spectroscopy. ACS Omega 6, 12877-12886 (2021). https://doi.org:10.1021/acsomega.1c01299

[52] Korn, S. M. et al. (1)H, (13)C, and (15)N backbone chemical shift assignments of the nucleic acid-binding domain of SARS-CoV-2 non-structural protein 3e. Biomol NMR Assign 14, 329-333 (2020). https://doi.org:10.1007/s12104-020-09971-6

[53] Johnson, E. C., Lazar, G. A., Desjarlais, J. R. & Handel, T. M. Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin. Structure 7, 967-976 (1999). https://doi.org:https://doi.org/10.1016/S0969-2126(99)80123-3

Note that we do not use residues with these fast motions to train a model for fast motions, we only use them for a statistical analysis of the 133 proteins in the RelaxDB dataset (test set) to evaluate conservation (Fig. 1e). It does not relate to missing peaks, but we explained that for evaluation, residues with extensive fast motions (hetNOE<0.65) could lead to an underestimation of us-ms exchange in the RelaxDB data set and therefore potentially only an underestimation of the Dyna-1 performance. The reason is that fast and slow motions have the opposite effect on R2 potentially canceling the R2 elevation due to exchange.

How did you ‘visually assess’ each dataset - such a procedure is not reproducible, why were those entries removed? This part is not well described and vague, you should precisely describe why you took these decisions and how.

We have updated our methods section to more precisely include how we discerned which experimental datasets we removed for RelaxDB and why. We note again, that this data set is only used for evaluation, not training of Dyna-1, hence does not influence the trained model.

“Filtering datasets resistant to fitting protocol. After curating 163 datasets initially, we realized in the process of finding a standardized way to fit all of them that some would not be well fit for a variety of reasons. For instance, datasets ISDHN3, MECP2, CBP6C, 5330, and MH35LIF were missing many R2 and R1 values, datasets 19356, CBP6N, and CBP6C all had hetNOEs < 0.65, and datasets 18773, 6758, and 5762 had relatively high Norm. RMSE values (see Extended Data Fig. 2a for comparison of all). We show the raw data for all the excluded datasets in Extended Data Fig. 1.”

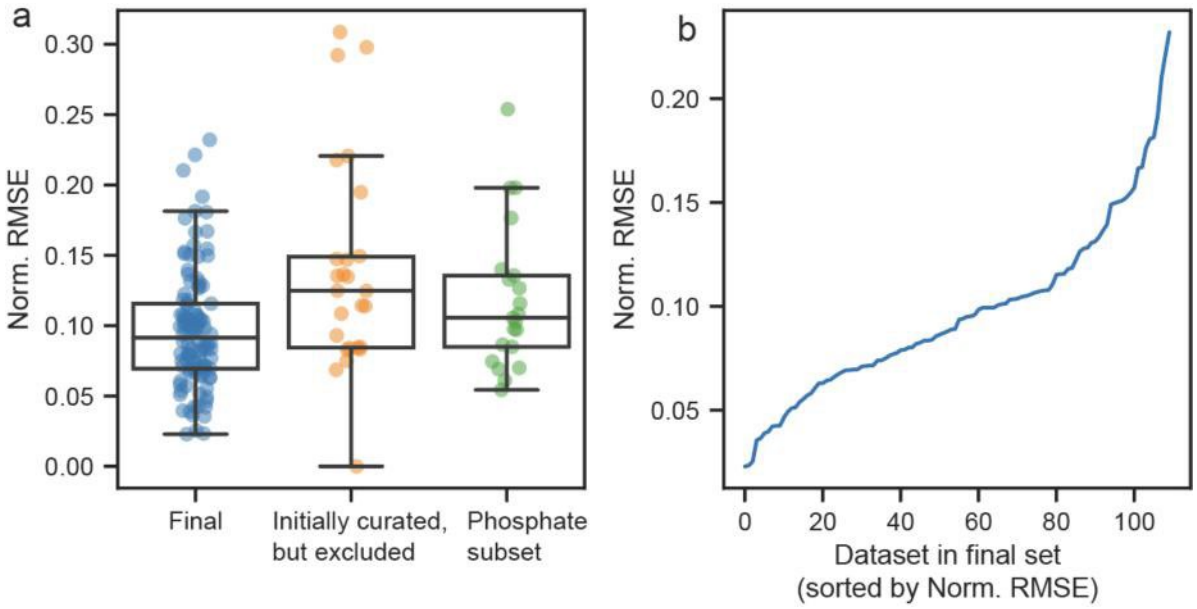
Importantly, in this revision we have made our HYDRONMR-based fitting significantly more quantitative:

We realized in revisiting our fitting that the fit parameter – the ratio of tumbling time between experimental and calculated R2/R1 – has a closed-form solution, whereas we were previously iteratively fitting this. We have also increased the rigor with which we do outlier-calling to infer where there is elevated R2/R1: previously we were using any value greater than one standard deviation above the mean of dR2/R1, but we realized this was skewed by outliers. We instead now estimate model uncertainty using interquartile range (more robust to outliers) and call any residue where R2/R1 is greater than the combined experimental error and model uncertainty.

Hence, our RelaxDB labels have changed slightly, yet our conclusions remain the same. Our overall AUROC values for RelaxDB have remained largely the same or even gone up slightly in some cases (previously: average over whole dataset not including phosphate buffer set was AUROC of 0.65, now it is 0.66).

We now include a normalized RMSE to evaluate fit for residues without dynamics labels. Summary statistics of this in the final dataset are included in Extended Data Figure 2. Norm RMSE is also included

for each dataset in Extended Data Fig. 1 now.



Extended Data Figure 2: (a) RMSE normalized to median R_2/R_1 per protein for the final RelaxDB dataset, protein datasets that were initially curated but excluded, and proteins with biological function involving phosphates measured in phosphate buffer. (b) Norm. RMSE distribution of final RelaxDB dataset used for evaluation.

From Methods (pg. 40): "In the final dataset, 65/111 datasets had a norm. RMSE of less than 0.1, and 107/111 had a norm. RMSE of less than 0.2 (Extended Data Fig. 2b). A norm. RMSE of 0.1 can be interpreted as the experimental data for residues that did not have dynamics assigned were on average within 10% of the predicted HYDRONMR value for R_2/R_1."

Also the four cases you present in Figure 1 - even though such a qualitative comparison is not relevant here (you can use the statistics), if you chose these 4 why did you pick them?

To visualize to a non-NMR expert reading the paper what these data look like and how one can discern Rex in elevated R2/R1 compared to R2/R1 ratios from overall tumbling, we had to pick a few proteins to display in the figure: we included 4 diverse proteins (small, bigger, long and anisotropic, more isotropic) and biologically-relevant proteins. Note that for the analysis in this figure, we analysed ALL proteins in Fig. 1e with appropriate statistics. We do not understand how the reviewer missed our analysis and the statistics displayed and explained in this figure, with the figure caption : "A benchmark of 133 protein NMR backbone relaxation measurements..."
We also note that all datasets are represented in Extended Data Figure 1.

- pLDDT is maybe a reasonable predictor of backbone order parameters by employing the structure-based Bruschweiler method, but on a dataset of 10 proteins this is not very convincing. Consider also the large-scale perspective, for example https://doi.org/10.1016/j.jmb.2024.168900.

We wanted to give credit to the Bruschweiler group for their previous finding, but we agree that because it was indeed a study on only 10 proteins, this is not a large sample size and we have removed the reference.

Further in relation to AF2, you say the 'glimpses that AF2 might have learned information about multiple conformations abound'. There is also other work that instead shows that this is only possible if it has memorised the conformation, see for examplehttps://www.nature.com/articles/s41467-024-51801-z. Please incorporate these more critical aspects in your work.

Our intent with this sentence was to note in the introduction that there is significant interest in using deep learning approaches to predict multiple conformations in the field. We have removed this line from the introduction. We also note that the above commentary (https://www.nature.com/articles/s41467-024-51801-z) has since been rebutted in a peer-reviewed paper at 10.1016/j.jmb.2025.169376, demonstrating that AF2 does use coevolutionary signal to predict multiple states, and not merely a result of memorization.

- In relation to lines 213-216, it has been shown that B-factors at cryogenic temperatures (which constitutes most of the x-ray diffraction determined structures in the PDB) have no relation to dynamics at room temperature, please see https://www.pnas.org/doi/abs/10.1073/pnas.1323440111.

We fully agree with the reviewer, that was the intention of our sentence.

- In relation to lines 209-210, even though NMR structures were not used in training AF2, there is of course overlap with structures determined by x-ray crystallography, where the core fold will be very similar. There is therefore significant overlap between the data in the BMRB and in the 'x-ray PDB'. Please clarify how you dealt with this - did you only use the BMRB entries that only had NMR structures?

There was no model training dependent on the structure data analyzed in Figure 2. The intent of our analysis in Figure 2 was to understand the similarity between missing assignments and residues missing from X-ray and EM datasets. We intentionally wanted to compare BMRB datasets and PDB structures, and allowed up to 10% sequence dissimilarity to find possible matches.

We previously had a sentence stating "NMR structures were removed from the AF2 training set, signifying that much of the mBMRB comprises proteins that AF2 and other algorithms adopting similar training splits have not been trained on." We have since realized that NMR structures actually were used to train current versions of AF2, and have removed this sentence. However, this does not change anything about our conclusions from this figure – that missing assignments represent distinct information from unresolved residues or residues with high B-factors in X-ray and EM structures.

You also have to use a 30% SI cutoff for creating training dataset for methods such as yours, as is standard in the field. Anything higher will confer significant bias on training/test.

We 100% agree with the reviewer with this cutoff, and we **did** use a 30% sequence identity cutoff, and additionally restricted structural homology by employing a 0.5 TM-score cutoff for training the final version of Dyna-1. This is stated in the main text: "We did not observe any individual proteins with significant jumps in AUROC between our most stringent training set (30% sequence identity, 0.5 TM-score cutoff) and our most lenient (80% sequence identity, 1.0 TM-score cutoff), which would have indicated model memorization (Extended Data Fig. 4d)."

- The level of conservation is also dependent on the multiple sequence alignment you use, and the evolutionary depth thereof. Can you comment on this aspect?

To perform the conservation analysis depicted in Figure 1, we used the MSA generation routine used in ColabFold. This was optimized for the purposes of ColabFold and performed the same for all proteins analyzed. This procedure was described in the Methods.

For example, the AF2 pair representation is doing very well as baseline, why would this be the case for the problem you are addressing?

Our final model, Dyna-1, takes into account both sequence and structure information via the multimodal language model ESM-3. This matches our expectation that both are useful for this task, and the AF-pair model's performance supports this.

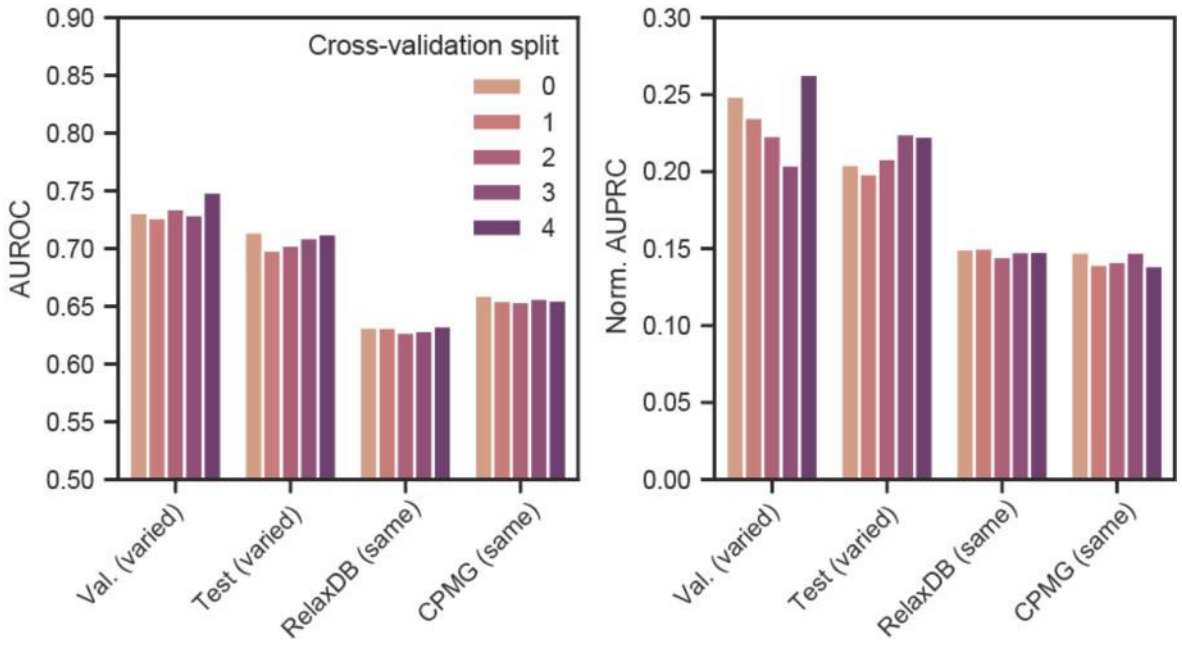
We have rephrased our results to state this:

"The best-performing model was layer 22 of ESM-3 if provided both sequence and structure models as input, which achieved an AUROC of 0.77. Intriguingly though, the last layer of ESM-2 model achieved an AUROC of 0.75, and the AF2-pair model, which did not use MSAs, and rather only structure input, achieved 0.71. This suggests that both sequence and structure information provide utility in this task."

Appropriate use of statistics and treatment of uncertainties

In terms of machine learning, I refer to the DOME recommendations (https://www.nature.com/articles/s41592-021-01205-4), which provide an essential guide to reporting data and performance measures. You have to cross-validate and report the spread of each fold (at least 5 here), for example - this is possible with the dataset you have.

The paper the reviewer cites states that the simplest approach for training is to perform independent train/test/val splits. This is an accepted standard in the deep learning field, in particular because of how time- and compute-intensive it is to train deep learning models. The paper further proposes that an alternative is cross-validation. To address this comment, we now also performed such 5-fold cross-validation, and find there is minimal difference in split performance across all of our validation and test sets, see figure below:



Conclusions: robustness, validity, reliability

Given the issues with likely bias in the dataset, it is not possible to ascertain whether the method is robust, valid nor reliable.

We are not sure what bias the reviewer is referring to. In curating our datasets used in training Dyna-1 from the BMRB, we included as many possible datasets as we could after filtering for deuterated and methyl-labeled datasets, two experimental schemes where there are clear reasons 15N backbone assignments would be missing not due to ms exchange. There is no bias in the training set of 3000 proteins (mBMRB) using a cutoff of less than 30% sequence identity.

For curating relaxation data as the test set (RelaxDB), we also included all data we could find that represented motions of apo, monomeric, folded proteins. These are the motions we were most interested in analyzing and where we felt most confident any measured Rex would represent intramolecular millisecond motions rather than other artefacts. We used all proteins with available ^{15}N relaxation data available to date, hence no bias.

We tested whether Dyna-1 is robust, valid and reliable by using 3 different test sets in the original paper, shown as AUROC in Fig. 3 (predicting missing assignments), Fig. 4 (predicting ms motions by comparing to 133 proteins with experimental data, RelaDB) Fig. 5 (predicting ms motions by comparing to 10 proteins with CPMG data).

During revision we have performed additional experimental validation with two proteins, described

below: *Prospective experimental test of Dyna-1*.

"To validate the predictive capabilities of Dyna-1, we conducted a prospective experimental test after public release of the model and initial review of the manuscript. We selected proteins from the "mBMRB test" set to characterize their dynamics, as these are proteins that are already assigned but many do not have published relaxation data. **Fig. 6a** depicts the mean p(exchange) from each protein in the mBMRB test set vs. sequence length, plotted by alpha helical / beta-strand content. This demonstrates that the predicted mean p(exchange) from Dyna-1 is broad for any given sequence length or helix / strand content. For proteins at the top and bottom range of mean p(exchange), we discarded proteins as candidates if they were fully intrinsically disordered, did not have published expression/purification methods sections, and already had some relaxation data publicly available. From the remaining proteins, we selected Chitinase 19 from *Bryum coronatum*[62] (BMRB: 11466), which had a high degree of p(exchange) (**Fig. 6b**), and hypothetical protein yjbJ[63] (BMRB entry: 5105), which had a low degree of p(exchange) (**Fig. 6c**). Performing ^{15}N relaxation dispersion CPMG demonstrated that, resembling Dyna-1's predictions, the chitinase did have exchange in region that are crucial for biologic activity—in its active site and the loops known to bind chitin (Figure 6b,e)—while the yjbJ only had detectable exchange for residues 2 and 3 (Fig. 6c,d, supplemental dataset 1). Evaluated at a per-residue level, Dyna-1 obtained an AUROC of 0.62 for chitinase and an AUROC of 0.72 for yjbJ."

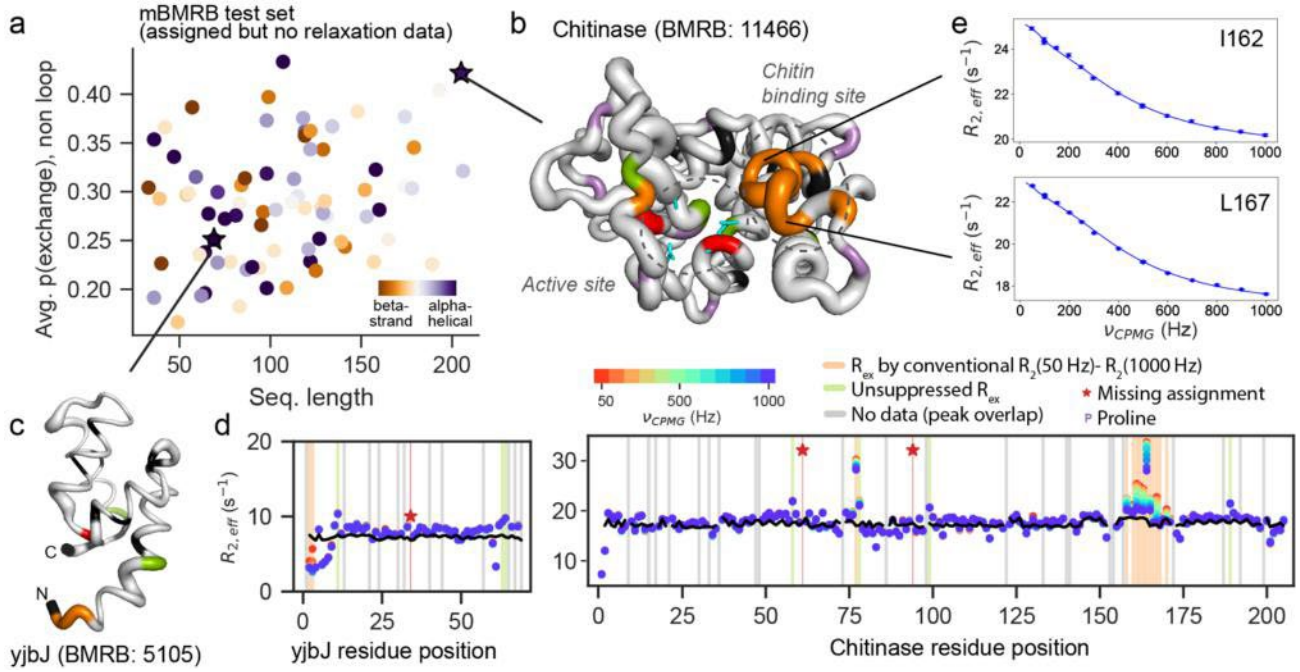


Fig. 6: Dyna-1 has predictive power for enzymatic motions in prospective experimental test. a) Dyna-1 proteins in mBMRB test set vary in net p(exchange) predicted across sequence length and helix/strand content. We selected predictions for two proteins to experimentally test: (b) Chitinase 19 from Bryum coronatum, which has high mean p(exchange), and (c) hypothetical protein yjbJ, which has low mean p(exchange). Residue coloring in (b,c,d) is same as in Fig. 5. (d) CPMG relaxation dispersion reveals yjbJ has negligible measurable relaxation, whereas chitinase has extensive exchange in the active site and chitin binding site. (e) Representative 15N CPMG dispersion data from residues I162 and L167 in the chitin binding site region.

Suggested improvements: experiments, data for possible revision

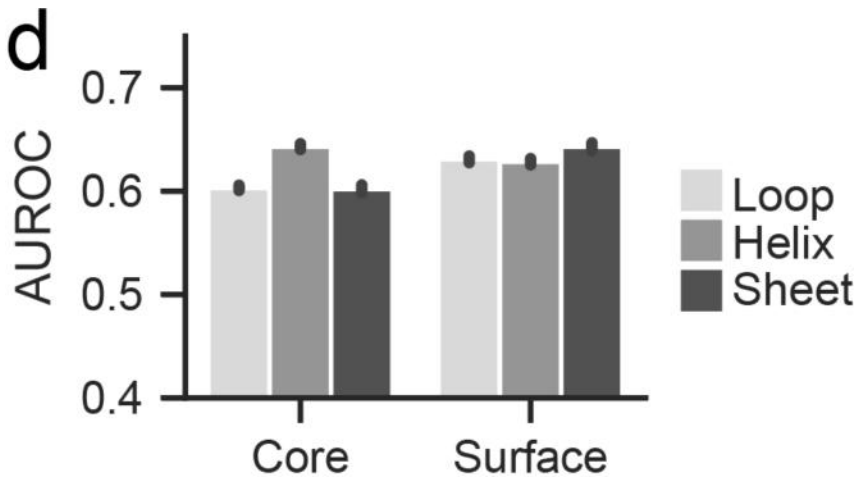
- Overall better explanations of choices made to create the datasets. Improve filtering of datasets and use much higher SI separation between the protein sequence.

1.See above for detailed explanation how we created and filtered the data sets, which was largely included in the original manuscript, but we made it even clearer in the revision.
2. The Dyna-1 model was trained using 30% sequence identity (SI) and 0.5 TM-score structure-based cutoff, as described in the main text and mentioned above. We have attempted to make this even more clear and explicitly stated in the main text:

"The final model leverages the embeddings of layer 22 of ESM-3, uses both sequence and structure as inputs, and was trained with the most stringent data split (30% sequence identity, 0.5 TM-score cutoff)."

- Reporting of dataset statistics - how large is the bias toward loop residues and how does this influence your results (also relation to test set)

See our answer to this comment made above already.
We are not sure what the reviewer means by "bias toward loop residues". However, we had reported AUROC over loop and non-loop residues in RelaxDB in Figure 4d (reproduced below):



- If you have scatter plots like in Figure 1, please also report correlations

There are no correlations necessary to report for the scatter plot in Figure 1 because we were not making claims about correlative trends in this data – this was intended to show that for residues with elevated R2/R1, many are in regions of hetNOE over 0.65.

References: appropriate credit to previous work?

No, see methods section.

We disagree with the suggestion of citing the dynasome reference and https://www.nature.com/articles/s41467-024-51801-z. These papers are not pertinent for our manuscript, see our explanation above.

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

Whilst I appreciate your enthusiasm in conveying the story of how you got to developing the method, it would be scientifically more appropriate if the data and approach speak for itself. Already in the abstract phrases like 'we realised an observable for us-ms motion was hiding in plain sight', 'We made the bold assumption' are unnecessary and distract from the facts. Also later (e.g. line 249) 'To our surprise', and others.

The reviewer arrives at the right conclusion of our choice of writing style in a few places, conveying in an interesting way to the reader how we got to developing the method, and to emphasize the key step. While our personality prefers such a writing style (also in reading papers as it makes it more enjoyable, relatable and teachable in our opinion), we acknowledge this reviewer requests changing it to a more generic style. For example, we have edited the text in these places:

We wished to address the lack of experimental dynamics benchmarks by curating as many datasets as we could of standardized datasets of R_{ex} measurements.

This suggested to us that if we could find these with more experimental data on µs-ms dynamics,

In setting out to curate as many datasets as we could of standardized R_{ex} measurements, we started with the experiment where R_{ex} is perhaps most directly measured

The developed model, Dyna-1 is able to predicts biologically relevant µs-ms motions as assessed on proteins previously characterized by many independent labs and with multiple types of experiments to measure µs-ms motion collected by many independent labs.

We next turned to a more abundant type of experiments that have been collected for the last 30 years, which we will refer to as R1/R2/NOE experiments

We wished to compared the information in these “missing peak” labels

After the painstaking work of finding a set of assignments to match each dataset in RelaxDB (and rejecting several candidate datasets where we could not reliably identify corresponding published assignments), we realized

Due to the challenges in discerning different types of motion in single-field-strength R_1/R_2/NOE experiments, we lastly wished sought to test Dyna-1 on proteins with µs-ms motions characterized by CPMG relaxation dispersion.

We were thrilled to see noticed that Importantly, Dyna-1 predicted higher p(exchange)

Response to Arbitrating Referee Comments on Referee #3

We thank the arbitrating referee for their careful reading of Referee 3's comments and for indicating which points they consider valid. All points marked with an asterisk (*) were fully addressed in our previous revision. We provide a brief account of each below, omitting points that the arbitrating referee deemed a "minor issue".

1. Dynasome reference (Originality and significance)

> *Referee #3: "Consider for example the dynasome … where the relation between dynamics and function was already explored and introduced. You should acknowledge this seminal work."*
>
> *Arbitrating referee (*): "This reference is distantly related to the manuscript — and one might even argue that it is not necessary to cite. It by no means challenges the novelty of the claims in the submitted paper."*

We agree with the arbitrating referee and did not add a citation to the dynasome paper in our revision. The dynasome study uses molecular dynamics simulations on timescales up to 80 ns, which are orders of magnitude shorter than the µs–ms motions addressed by our work. Instead, we added a citation to the foundational Frauenfelder, Sligar & Wolynes (*Science*, 1991) paper for the statement that protein function depends on conformational interconversion.

2. Dataset statistics, loop bias quantification, and scatter plot correlations

> *Referee #3: "Overall better explanations of choices made to create the datasets … Reporting of dataset statistics … If you have scatter plots like in Figure 1, please also report correlations."*
>
> *Arbitrating referee (*): "The above are good suggestions."*

All three points from reviewer 3 were addressed as follows. Black: original reviewer 3 comment, blue: our response

- Overall better explanations of choices made to create the datasets. Improve filtering of datasets and use much higher SI separation between the protein sequence.

1.See above for detailed explanation how we created and filtered the data sets, which was largely included in the original manuscript, but we made it even clearer in the revision.

[From above in response:]

"In curating our datasets used in training Dyna-1 from the BMRB, we included as many possible datasets as we could after filtering for deuterated and methyl-labeled datasets, two experimental schemes where there are clear reasons 15N backbone assignments would be missing not due to ms exchange. There is no bias in the training set of 3000 proteins (mBMRB) using a cutoff of less than 30% sequence identity.

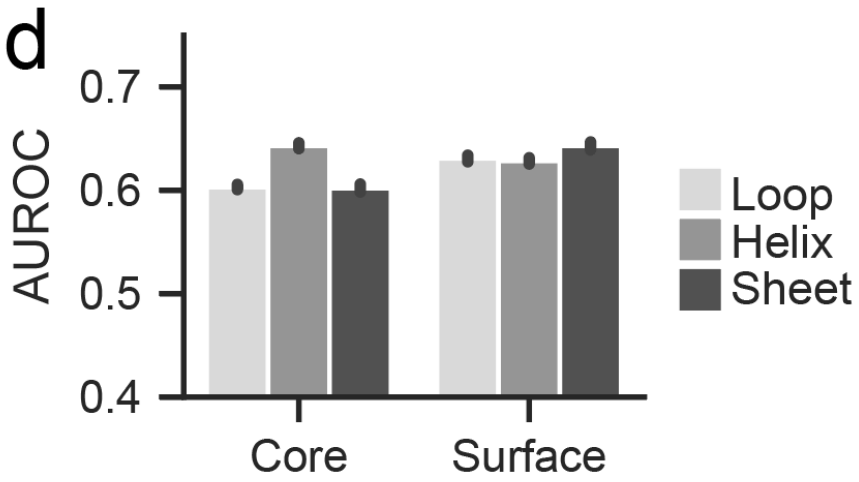
For curating relaxation data as the test set (RelaxDB), we also included all data we could find that represented motions of apo, monomeric, folded proteins. These are the motions we were most interested in analyzing and where we felt most confident any measured Rex would represent intramolecular millisecond motions rather than other artefacts. We used all proteins with available ^{15}N relaxation data available to date, hence no bias."

2. The Dyna-1 model was trained using 30% sequence identity (SI) and 0.5 TM-score structure-based cutoff, as described in the main text and mentioned above. We have attempted to make this even more clear and explicitly stated in the main text:

"The final model leverages the embeddings of layer 22 of ESM-3, uses both sequence and structure as inputs, and was trained with the most stringent data split (30% sequence identity, 0.5 TM-score cutoff)."

- Reporting of dataset statistics - how large is the bias toward loop residues and how does this influence your results (also relation to test set)

We are not sure what the reviewer means by "bias toward loop residues". However, we had reported AUROC over loop and non-loop residues in RelaxDB in Figure 4d (reproduced below):



- If you have scatter plots like in Figure 1, please also report correlations

There are no correlations necessary to report for the scatter plot in Figure 1 because we were not making claims about correlative trends in this data – this was intended to show that for residues with elevated R2/R1, many are in regions of hetNOE over 0.65.

3. Language and tone in the abstract and main text

> *Referee #3: "Already in the abstract phrases like 'we realised an observable for us-ms motion was hiding in plain sight', 'We made the bold assumption' are unnecessary and distract from the facts. Also later 'To our surprise', and others."*
>
> *Arbitrating referee (*): "I agree with the comments above."*

This was addressed in our prior revision. We edited multiple passages throughout the manuscript to replace informal narrative phrases with direct scientific language. Examples of changes are included below:

We wished to address the lack of experimental dynamics benchmarks by curating ~~as many datasets as we could of~~ standardized datasets of R_{ex} measurements.

This suggested to us that ~~if we could find these~~ with more experimental data on µs-ms dynamics,

In setting out to curate ~~as many~~ datasets ~~as we could~~ of standardized R_{ex} measurements, we started with the experiment where R_{ex} is perhaps most directly measured

The developed model, Dyna-1 ~~is able to~~ predicts biologically relevant µs-ms motions as assessed on proteins previously characterized by many independent labs and with multiple types of experiments to measure µs-ms motion ~~collected by many independent labs~~.

We next turned to a more abundant type of experiments ~~that have been collected for the last 30 years~~, which we will refer to as R1/R2/NOE experiments

We ~~wished to~~ compared the information in these "missing peak" labels

After ~~the painstaking work of~~ finding a set of assignments to match each dataset in RelaxDB ~~(and rejecting several candidate datasets where we could not reliably identify corresponding published assignments)~~, we realized

Due to the challenges in discerning different types of motion in single-field-strength R_1/R_2/NOE experiments, we lastly ~~wished~~ sought to test Dyna-1 on proteins with µs-ms motions characterized by CPMG relaxation dispersion.

~~We were thrilled to see noticed that~~ Importantly, Dyna-1 predicted higher p(exchange)