Supplementary Information: A Highly Limited Amino Acid Library from Asteroid Bennu Yields Wide-ranging Protein Folds

Stuart Bartlett*[1,2], Arushi Gupta[1], Dominic Phung[1], Josh Goldford[3], Jiani Yang[1], Klara Hlouchova[4], Woodward Fischer[1], and Yuk L. Yung[1]

[1]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA
[2]SETI Institute, Mountain View, CA, USA
[3]Dayhoff Labs, Boston, MA, USA
[4]Faculty of Science, Charles University, Prague, Czech Republic

February 28, 2026

Sensitivity Analyses

Comparison of Structural Metrics

Since AI protein modeling software are trained through a process of minimising prediction errors, there are several associated metrics of structural fit that the models provide. Additionally, the TM-score is one among several protein structural comparison metrics. Therefore, we performed a sensitivity study to compare various metrics to ensure a baseline level of consistency.

We focus initially on two well-matched structures from very small AA libraries: 1) A Succinyl CoA Synthase with PDB code 1OI7 formed using only A, G, P, and V[1] (the minimal library constrained by measurements of the Bennu asteroid or Murchison meteorite [2]), and 2) A Malate Dehydrogenase with PDB code 7CGD, constructed from an AA library of size 5 (A, D, G, S, and V) guided by Miller-Urey experimental results [3].

*Corresponding author: sbartlett@setiap.org

[1]In the following we use standard single-letter amino acid codes: Alanine (A), cysteine (C), aspartic acid (D), glutamic acid (E), phenylalanine (F), glycine (G), isoleucine (I), leucine (L), methionine (M), proline (P), serine (S), threonine (T), valine (V).

We compare 5 prediction quality and structural comparison metrics:

1. The ProteinMPNN Fit Score: ProteinMPNN's internal measure of expected structural match between target backbone and backbone arising from suggested sequence [1].

2. Alphafold2 pTM: The predicted template modeling score (as opposed to the actual template modeling score measured after structure prediction) that the Alphafold2 algorithm expects between the modeled and target structures [4]. Since the algorithm itself doesn't know the target structure, this is essentially a guess on its part, a measure of its confidence in the predicted structure.

3. Template modeling (TM) score: A global measure of structural match between two proteins, normalised (insensitive to) the overall size of the structures [7, 8].

4. Root Mean Square Distance (RMSD): a simple measure of the root mean square distance between corresponding atoms in the two protein structures, given in units of Angstroms.

5. Maximum Substructure score (MaxSub-Score): A measure of the degree of structural match between the closest-aligned substructures in the modeled and target proteins. This essentially zooms in on the substructures in the two proteins that have the most closely aligned residues and measures the structural match between them [6].

Supplementary Table 1 displays these metrics for 5 of our 10 modelled sequences for the 1OI7 Succinyl CoA Synthase structure, when modeled using the minimum Bennu asteroid library: A, G, P and V. We see that the ProteinMPNN fit scores are essentially uniform across all 5 suggested sequences, suggesting that this metric cannot be relied upon for assessing structural match. This may be due to the fact that ProteinMPNN's training only gives it an 'implicit understanding' of the sequence to structure map. It may be relying upon complex relationships between the vectors in its training embedding space, which themselves have a complex relationship to the protein structure that would arise from such sequences.

The next metric, the Alphafold2 pTM score, shows high values for sequences 2 and 4, and low values for the other three. This implies that it is not confident in its final predicted structure, presumably because the sequence is far from anything it encountered in its training, with perhaps additional modeling difficulties. The pTM correlates well with the next metric, the actual TM-score, also high for sequences 2 and 4, and low for the other three. In turn these scores are also in line with the RMSD values, which has an optimal lower bound of 0 for a perfect match. Hence the values of 1.51 and 1.67 Å imply a very good match, in agreement with the pTM and TM-scores. Finally, the MaxSub-scores also correlate with the pTM, TM-score and RMSD for these structures.

Metric		1	2	3	4	5
ProteinMPNN	Fit Score	1.43	**1.43**	1.43	1.43	1.41
AlphaFold2	pTM	0.24	**0.81**	0.25	0.78	0.24
Comparison	TM-score	0.21	**0.94**	0.22	0.94	0.23
	RMSD	2.77	**1.51**	2.51	1.67	2.66
	MaxSub-score	0.12	**0.84**	0.12	0.81	0.16

Supplementary Table 1: Evaluation metrics for 5 sequences generated for the 1OI7 backbone using ProteinMPNN and predicted with AlphaFold2. The AA library used here was the minimum Bennu asteroid set: A, G, P and V. Note that the second column of scores, shown in bold, corresponds to the structure displayed in Fig. 4a of the main text.

As a second example, we list the respective scores for 5 sequences targeting the Malate Dehydrogenase 7CGD structure, modeled using the 5 AA library constrained by the Miller volcanic spark discharge experiment: G, A, S, D and V [3]. As above, we see very little variation in the proteinMPNN Fit Scores, but total correspondence between the other metrics, which show that the structure in column 5 (shown in Fig. 4b of the main text) is a very good match to the target.

Metric		1	2	3	4	5
ProteinMPNN	Fit Score	1.51	1.54	1.56	1.50	**1.60**
AlphaFold2	pTM	0.30	0.26	0.31	0.28	**0.76**
Comparison	TM-score	0.19	0.19	0.17	0.20	**0.90**
	RMSD	3.10	3.16	2.79	2.70	**2.03**
	MaxSub-score	0.07	0.84	0.07	0.07	**0.74**

Supplementary Table 2: Evaluation metrics for 5 sequences generated for the 7CGD backbone using ProteinMPNN and predicted with AlphaFold2. The AA library used here was the 5 AA library constrained by the Miller volcanic spark discharge experiment: G, A, S, D and V [3]. Note that the fifth column of scores, shown in bold, corresponds to the structure shown in Fig. 4b of the main text.

The above results suggest that the pTM, TM-score, RMSD and MaxSub-score are all useful, correlated metrics of the goodness of fit between pairs of protein structures. We employed the TM-score for our final analysis due to its widespread usage and demonstrated accuracy [7, 8]. We illustrate the correspondence between the TM-score and RMSD by plotting the RMSD-scores for all enzymes analysed in Supplementary Figure 1 and Supplementary Figure 2. Rather than plotting the actual RMSD, which attributes lower values to better matched structures, we plot the following inverted score function: $1-RMSD/8$.

This function assigns higher values to better matches, with the perfect score of 1 corresponding to an RMSD of 0 Å, and a minimum score of 0 corresponding to an RMSD of 8 Å or more (results are not sensitive to small variations in this cutoff value). Note that we assign a threshold 'good match' value of $RMSD_{gm} = 3$ Å, shown as dotted black lines in Supplementary Figure 1 and Supplementary Figure 2.

We see that for both the TM-score and RMSD, we achieve full matching for all protein folds at library sizes of 6, 7, 7 and 8 AAs for the Bennu, Murchison, Miller-Urey volcanic and Miller-Urey H_2S-rich libraries, respectively. However, there is slight disagreement between the metrics in terms of the lack of monotonic increase in scores as a function of library size for the RMSD. This could be due to a sizing effect, since the TM-score is designed to be insensitive to the overall protein size. Subsets of our primitive analogue structures may be larger or smaller than their targets, and this could be leading to poorer fit according to the RMSD. Future work will compare structures at an atomic level of detail to elucidate the origin of these differences.

Comparison of Structure Prediction Fidelity using ESMFold

One of the strengths of Alphafold2 and similar structure prediction tools is its use of 'Multiple sequence alignment' (MSA). This method identifies large numbers of protein sequences that are related to the target sequence and aligning them. AlphaFold2's neural network then interprets those patterns to infer long-range contacts and folds the AA chain into a structure that best fits the evolutionary clues (sometimes using known template structures as additional hints).

MSA can be a powerful asset for structure prediction tools, but can also lead to hallucinations, wherein small segments of a sequence are 'over-interpreted' to predict the resulting 3D structure despite considerable sequence differences in other regions. This highlights the fact that such prediction tools have not explicitly learned the physics of protein folding in order to make their predictions, but rely heavily on sequence homologies or overlaps to make educated guesses about 3D structure.

Given the strong constraints on AA libraries used in the present work, many of our sequences are likely quite different to the vast majority that were used to train models like Alphafold2, but likely contain small segments that do resemble known sequences. Hence, caution is warranted when applying MSA-based structure prediction tools to these kinds of 'out of sample' datasets.

We therefore conducted a sensitivity study of MSA biases by re-running our entire pipeline and libraries with 'ESMFold' instead of Alphafold2. ESMFold is a fast, MSA-free protein structure predictor that uses a protein language model to infer structure from a single sequence, and is available as a Colab notebook. It's data-driven and generally slightly less accurate than AlphaFold2 when rich evolutionary information is available.

The results of our ESMFold re-analysis are shown in Supplementary Figure 3 and Supplementary Figure 4. We see that all of our previous results derived from Alphafold2 predictions are re-capitulated apart from 2 exceptions. In Supplementary Figure 4a, we see that 3 protein families remain below threshold score values for the Miller-Urey volcanic library even with 7 AAs present. There are several potential explanations for this discrepancy, including the following:

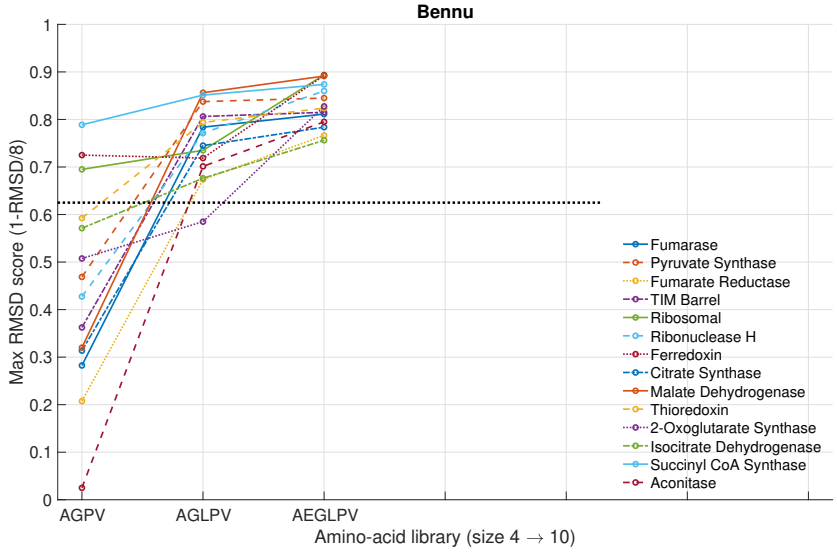
1. Since our two pipelines include the ProteinMPNN step in series, it is possible that, by chance, the sequences generated during the ESMFold runs were simply lower quality than the sequences (or one particularly good sequence) produced during the Alphdfold2 runs.

2. ESMFold was unable to make good structure predictions due to the sequences being too different from those that the model was trained on.

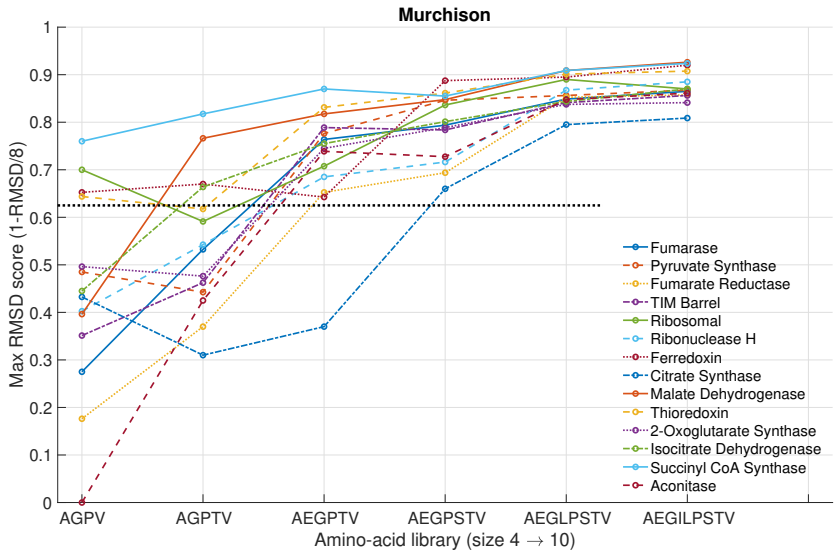
3. EMSFold's predictions are accurate but the higher scoring Alphafold2 results are due to an over reliance on MSA and hence could be hallucinatory.

In addition, the ESMFold predictions shown in Supplementary Figure 3b suggest that a library of 8 AAs is required to reproduce all folds when using constraints from the Murchison meteorite (a library size of just 7 AAs was required when using Alphafold2).

The fact that the vast majority of the results from our Alphafold2 analysis are re-produced when using ESMFold, namely that we achieve full structural matching for libraries of size 6, 8 and 8 for the Bennu, Murchison and Miller-Urey H_2S-rich libraries, respectively, and $\sim 79\%$ matching for the 7 AA Miller-Urey volcanic library, lends support to the robustness of our conclusions.

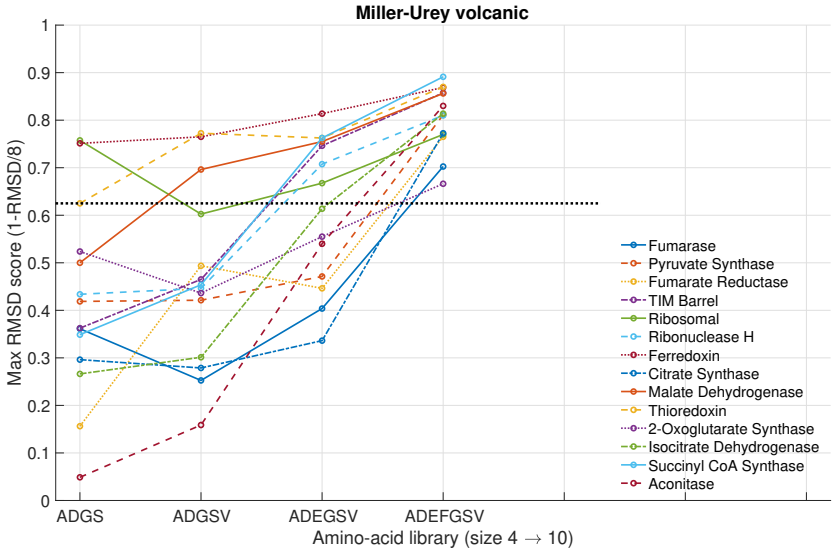


(a) Bennu asteroid, [2]

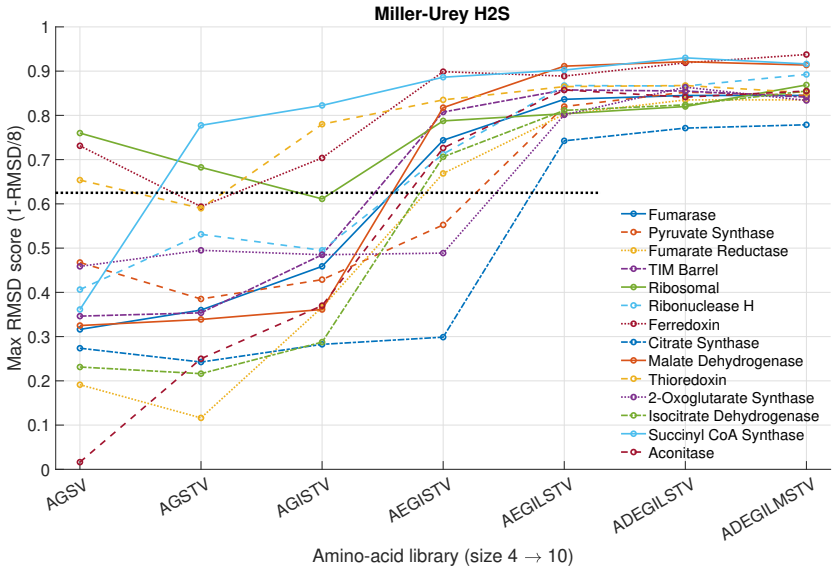


(b) Murchison meteorite, [2]

Supplementary Figure 1: Maximum RMSD scores ($1 - RMSD/8$) as a function of AA library size. AA libraries constrained by AA abundances from a) Bennu asteroid and b) Murchison (CM2) meteorite measurements [2]. The dotted, horizontal black line indicates the threshold score of $RMSD_{gm} = 3$ Å. Source data are provided as a Source Data file.

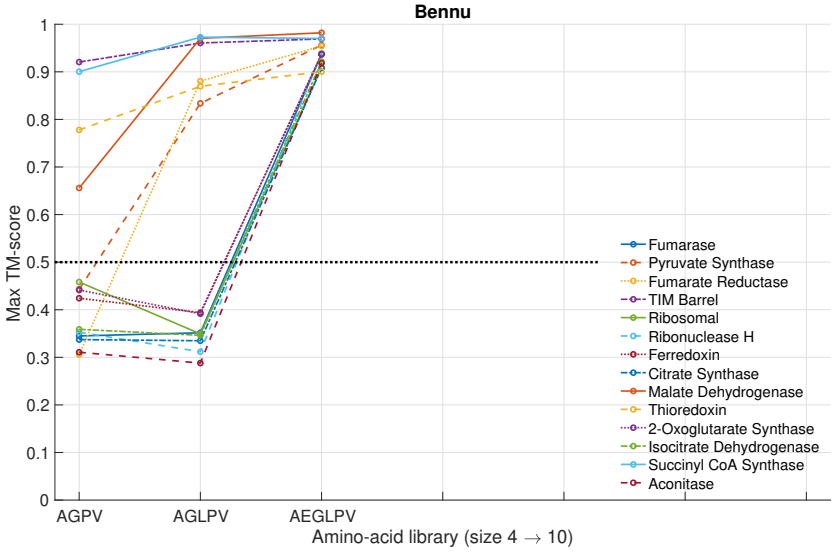


(a) Miller-Urey volcanic [3]

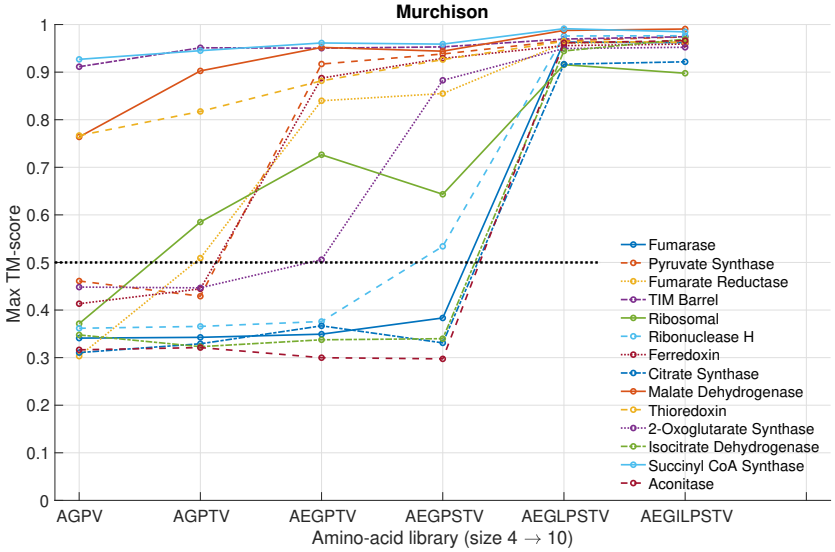


(b) Miller-Urey H_2S-rich [5]

Supplementary Figure 2: Maximum RMSD scores ($1 - RMSD/8$) as a function of AA library size. AA libraries constrained by: a) 1953–54 University of Chicago Miller-Urey experimental samples [3], b) 1958 Miller-Urey H_2S-rich spark discharge experiment [5]. The dotted, horizontal black line indicates the threshold score of $RMSD_{gm} = 3$ Å. Source data are provided as a Source Data file.

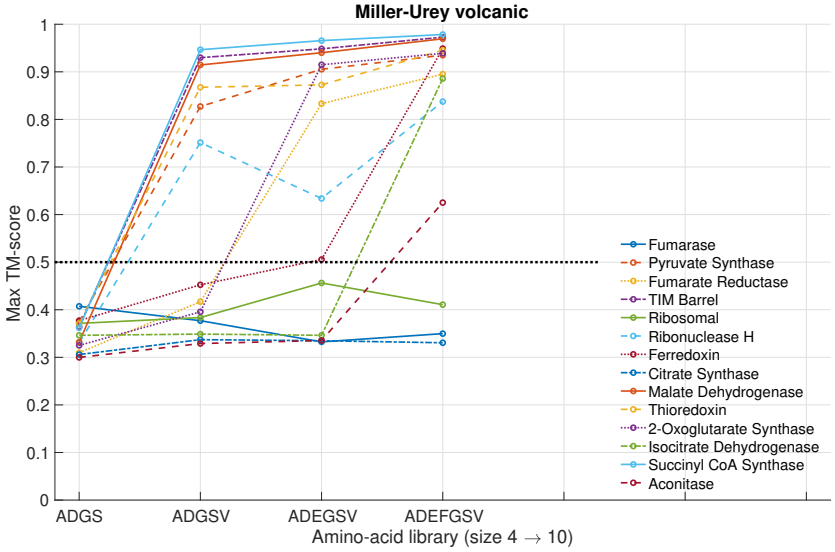


(a) Bennu asteroid, [2]

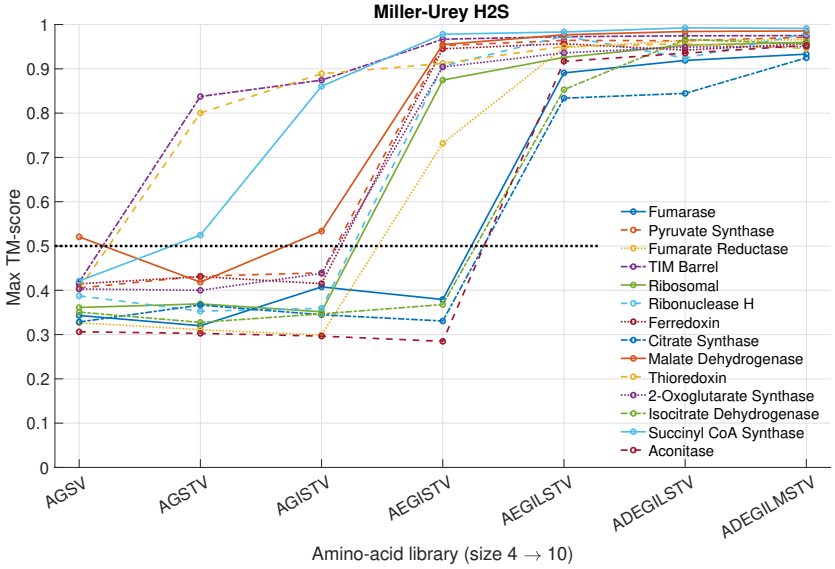


(b) Murchison meteorite, [2]

Supplementary Figure 3: Maximum TM-Scores as a function of AA library size when using ESMFold for structure prediction. AA libraries constrained by AA abundances from a) Bennu asteroid and b) Murchison (CM2) meteorite measurements [2]. Note that the horizontal axes are the same for both subfigures but the Bennu library only extends to a maximum size of 6 AAs, which is why the curves appear to terminate earlier than those in the lower panel. The dotted, horizontal black line indicates the threshold score of 0.5, above which a pair of structures are considered 'essentially the same' [7]. Source data are provided as a Source Data file.



(a) Miller-Urey volcanic [3]



(b) Miller-Urey H_2S-rich [5]

Supplementary Figure 4: Maximum TM-Scores as a function of AA library size when using ESMFold for structure prediction. AA libraries constrained by: a) 1953–54 University of Chicago Miller-Urey experimental samples [3], b) 1958 Miller-Urey H_2S-rich spark discharge experiment [5]. Note that the horizontal axes are the same for both subfigures but the volcanic library only extends to a maximum size of 7 AAs, which is why the curves appear to terminate earlier than those in the lower panel. The dotted, horizontal black line indicates the threshold value of 0.5, above which a pair of structures are considered 'essentially the same' [7]. Source data are provided as a Source Data file.

Supplementary References

[1] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187. URL `https://www.science.org/doi/abs/10.1126/science.add2187`.

[2] Daniel P Glavin, Jason P Dworkin, Conel M O'D Alexander, José C Aponte, Allison A Baczynski, Jessica J Barnes, Hans A Bechtel, Eve L Berger, Aaron S Burton, Paola Caselli, et al. Abundant ammonia and nitrogen-rich soluble organic matter in samples from asteroid (101955) bennu. *Nature Astronomy*, pages 1–12, 2025.

[3] Adam P Johnson, H James Cleaves, Jason P Dworkin, Daniel P Glavin, Antonio Lazcano, and Jeffrey L Bada. The miller volcanic spark discharge experiment. *Science*, 322(5900):404–404, 2008.

[4] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

[5] Eric T Parker, Henderson J Cleaves, Jason P Dworkin, Daniel P Glavin, Michael Callahan, Andrew Aubrey, Antonio Lazcano, and Jeffrey L Bada. Primordial synthesis of amines and amino acids in a 1958 miller h2s-rich spark discharge experiment. *Proceedings of the National Academy of Sciences*, 108(14):5526–5531, 2011.

[6] Naomi Siew, Arne Elofsson, Leszek Rychlewski, and Daniel Fischer. Maxsub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, 16(9):776–785, 09 2000. ISSN 1367-4803. doi: 10.1093/bioinformatics/16.9.776. URL `https://doi.org/10.1093/bioinformatics/16.9.776`.

[7] Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with tm-score= 0.5? *Bioinformatics*, 26(7):889–895, 2010.

[8] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.