RFDiffusion performance benchmarking for local high-end workstation computers

Mario Axel López Aguiñaga[1][0000−0001−8501−4847], María Teresa Romero Gutiérrez[1,2][0000−0002−4882−7491], José Alejandro Morales Valencia[1][0000−0001−8088−9724], and Omar Paredes[1][0000−0002−5382−6127]

[1] University of Guadalajara, Universitary Center of Exact Sciences and Engineering, Engineering and, Boulevard General Marcelino García Barragán 1421, Olímpica, 44430 Guadalajara, Jalisco, México. axel.lopez@alumnos.udg.mx
[2] University of Guadalajara, Universitary Center of Tlajomulco, Division of Technological Development and Engineering, Carretera Tlajomulco - Santa Fé Km. 3.5 No.595 Lomas de Tejeda, 45641 Tlajomulco de Zúñiga, Jalisco, México.

Abstract. RFDiffusion is a denoising diffusion model that allows users to create de novo protein structures from tridimensional Gaussian noise. This model is continuously revolutionizing the way we design proteins that are potentially capable of approaching an immense variety of modern day problems such as drug design, drug delivery, residue management, biological manufacturing and vaccine design. However; given that no information related to the computational performance of this tool is provided neither in the original Github repository created by the baker-lab nor in the literature related to it and also the lack of a cloud-based version of this tool makes its use a risky task as there is no way to find out if the local computers specifications of a laboratory will be enough to drive a proper protein design experiment. In order to evaluate the computational performance of RFDiffusion we installed it in a high-end workstation hosted in our laboratory and performed a series of hallucination experiments, these experiments consisted in fixing a lower limit for the protein's length in 100 amino acids and if the computer did not display any kind of error in the terminal then 50 additional aminoacids are added to the lower limit. The results show that the operational limit of RFDiffusion is around the 1600 amino acids in length which can be considered an acceptable limit given that most of proteins do not surpass the 1000 amno acids of length and if they do it is in the form of proteic complexes.

Keywords: Hallucination models · Protein design · Artificial Intelligence · Benchmark.

Introduction

RFDiffusion is a denoising diffusion model that allows the creation of de novo tridimensional protein structures by using a cloud of tridimensional Gaussian noise[5, 19]. RFDiffusion uses the protein folder model RoseTTAFold to predict

https://doi.org/10.61728/AE20254971

the tridimensional structures for fragments of already known proteins contained inside a database, in turn, these predicted protein fragments are used by the diffusion denoising part of the model as tridimensional templates to reshape the Gaussian noise once and then discarded to generate a new tridimensional template and repeat this process iteratively until a number of iterations automatically determined by RFDiffusion is reached and the protein hallucination process is stopped.[18, 7] With this process RDFdiffusion is capable of creating protein structures that do not exist in the nature and that can be designed to approach a wide variety of functions such as drug delivery mechanisms, residue management, biological manufacturing and vaccine design[6]. However; with the computationally expensive costs that are associated with running a generative artificial intelligence model the questions about the availability of local use for RFDiffusion arise and the lack of an online version for this tool only fuel these questions even more. To respond these questions in this work we installed RFDiffusion using the provided instructions in the official GitHub in a workstation computer with a high-end GPU inside and made an experiment where we hallucinated proteins with a iteratively-increasing amino acid chain length to find out where are the limits of the RFDiffusion and if we can generate proteins within those limits.

Materials & Methods

Given the fact that the generative artificial intelligence models rely heavily on the GPU capabilites of the computer where they are installed we decided that the experiments must be done in a workstation computer as they are more powerful than a conventional computer but more accessible than an entire dedicated server. The Biodigital Innovation Laboratory allowed the access to one of their most powerful high-end workstations in order to make the RFDiffusion benchmarking. The workstation specifications are depicted in the table 1 meanwhile the benchmarking experiment is illustrated in the figure 1.

Table 1. Computational specificatons. These especifications correspond to the workstation computer used for the benchmarking of the RFDiffusion tool.

Component	Model	Features
CPU	Intel Xeon(R) Silver	40 cores
GPU	Nvidia GeForce RTX 3090	24 GB VRAM
RAM	Kingston 9965757-004.A00G	270 GB RAM
OS	System76 Pop!_OS	22.04 LTS

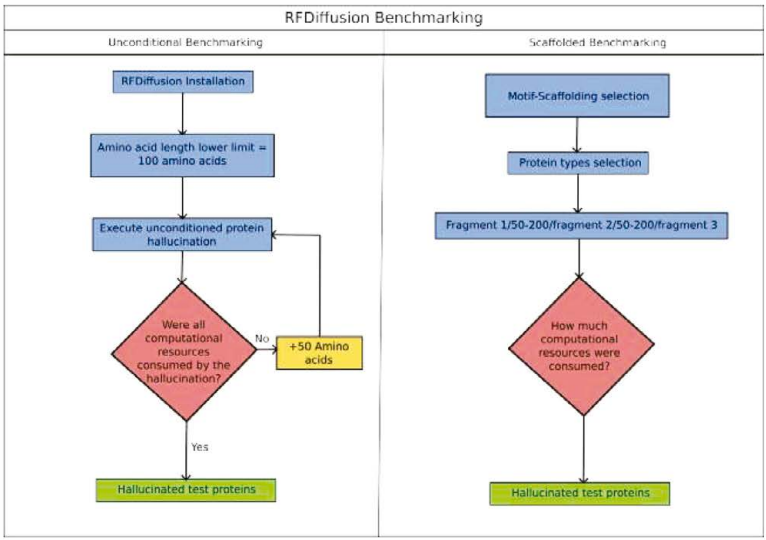

Fig. 1. Schematic representation of the steps used for the RFDiffusion performance benchmarking. These experiments are intended to find the quantity of amino acids where RFDiffusion is no longer capable to generate hallucinated proteins due to a lack of available computational resources and also see the quantity of resources consumed when the hallucination process takes fragments of an already known protein structure as templates.

1 RFDiffusion Installation

We installed the RFDiffusion tool by using the instructions contained inside the official GitHub (`https://github.com/RosettaCommons/RFdiffusion`) provided by the bakerlab.

Summarizing the contents of the instructions we first cloned the RFDiffusion repository into our test workstation. Consequently we downloaded all of the 8 necessary model weights to run this tool properly. Later we set up a conda environment to contain the SE(3) transformer that RFDiffusion requires to work. Finally, via the command terminal we activated such environment and downloaded the actual SE(3) transformer into the test computer by using Python pip.

2 Lower limit definition

The lower limit for the hallucination experiment was determined by reading the original RFDiffusion paper where the authors stated that the tool was capable to hallucinate small proteins with a length of 100 aminoacids sucessfully.

3 Unconditioned benchmarking

With the tool installed in the test computer we proceeded to activate the conda environment that contains the SE(3) transformer and executed the following command lines into the Linux terminal:

```
./scripts/run_inference.py 'contigmap.contigs=[i-i]'
inference.output_prefix=test_outputs/test
inference.num_designs=1
```

Where i is the length of the amino acid chain to hallucinate, in this case the starting point was fixed to 100 amino acids. Given that RFDiffusion hallucinates unconditionated proteins with a length that varies randomly within a defined interval we set both ends of the interval to be the same number (that's why in contig map appears an $[i-i]$), with this we ensure that the hallucinated proteins will have the same increasing length.
Now, as mentioned before this hallucination experiment is iterative and the condition to pass to the next iteration, in this case, to hallucinate a protein 50 amino acids larger than the previous one is that there must not be any error output in ther command terminal. Not having any kind of error output in the terminal means that the computer has not consumed all of the computational resources available (VRAM) to generete a new protein, by other hand, if the computer consumes all of the resources available we obtain an error output where the process cannot continue and consequently we define this amino acid length as the operational limits of the tool.

4 Scaffolding benchmarking

To test the computational resources consumption of RFDiffusion when a series of protein structures is given to it as templates to hallucinate proteins based on them we downloaded 9 tridimensional structures corresponding to the different types of proteins described on the table 2 from the Protein Data Bank.

Table 2. Protein structures and types used for RFDiffusion Scaffolding benchmarking.

Protein type	PDB ID	Description	Selected chain
Enzyme	1BPB	DNA polymerase β [11]	A
Structural	1K6F	Collagen helices [2]	A
Signaling	1BEN	Insulim [13]	D
Regulatory	1AIE	P53 tumor supression factor [9]	A
Transport	3GOU	Hemoglobin [17]	A
Sensorial	1NP7	Cryptochrome [3]	A
Motor	1GOJ	Fast kinesin [14]	A
Defense	1FL5	Antibody 28B4 [20]	A
Storage	1OVA	Chicken Albumin [16]	A

Given the fact that almost all of these downloaded structures represent protein complexes we selected only one of their respective chains to make the scaffolding benchmarking by splitting these structures into 3 fragments which are then connected by hallucinated protein fragments with a random length between the 50 and the 200 amino acids.
The conda environment that contains the SE(3) transformer was activated and the following command lines were introduced intro the Linux terminal to direct RFDiffusion into using the motif-scaffolding mode:

```
../scripts/run_inference.py
inference.output_prefix=test_outputs/protein_benchmark_1
inference.input_pdb=1bpb.pdb
'contigmap.contigs=[A91-150/50-200/A152-175/50-200/A182-
inference.num_designs=1

../scripts/run_inference.py
inference.output_prefix=test_outputs/protein_benchmark_2
inference.input_pdb=1k6f.pdb
'contigmap.contigs=[A1-10/50-200/A11-20/50-200/A21-29]'
inference.num_designs=1

../scripts/run_inference.py
inference.output_prefix=test_outputs/protein_benchmark_3
inference.input_pdb=1ben.pdb
'contigmap.contigs=[D1-10/50-200/D11-20/50-200/D21-28]'
inference.num_designs=1

../scripts/run_inference.py
inference.output_prefix=test_outputs/protein_benchmark_4
inference.input_pdb=1aie.pdb
'contigmap.contigs=[A326-336/50-200/A337-346/50-200/A347
```

```
inference.num_designs=1

../scripts/run_inference.py
inference.output_prefix=test_outputs/protein_benchmark_5
inference.input_pdb=3gou.pdb
'contigmap.contigs=[A1-40/50-200/A41-80/50-200/A81-121]'
inference.num_designs=1

../scripts/run_inference.py
inference.output_prefix=test_outputs/protein_benchmark_6
inference.input_pdb=1np7.pdb
'contigmap.contigs=[A1-150/50-200/A151-250/50-200/A251-483]'
inference.num_designs=1

../scripts/run_inference.py
inference.output_prefix=test_outputs/protein_benchmark_7
inference.input_pdb=1goj.pdb
'contigmap.contigs=[A2-120/50-200/A121-250/50-200/A251-355]'
inference.num_designs=1

../scripts/run_inference.py
inference.output_prefix=test_outputs/protein_benchmark_8
inference.input_pdb=1fl5.pdb
'contigmap.contigs=[A1-70/50-200/A71-120/50-200/A121-212]'
inference.num_designs=1

../scripts/run_inference.py
inference.output_prefix=test_outputs/protein_benchmark_9
inference.input_pdb=1ova.pdb
'contigmap.contigs=[A23-150/50-200/A151-293/50-200/A294-391]'
inference.num_designs=1
```

Results & Discussions

By executing the iterative hallucination experiment obtained an error output and consequently a terminal process abortion related to the lack of available VRAM at the 1600 amino acids of length, thus we estimated the operational limit for protein hallucination with RFDiffusion in this same quantity. As the reader can see in figure 4 there is a linear tendency between the use of graphic memory (VRAM) and the quantity of aminoacids that compose the hallucinated protein (AA→ amino acids), however; we also must recongnize the appearance of two consecutive peaks in the consume of VRAM when the hallucinated protein length is between the 600 and 800 amino acids.

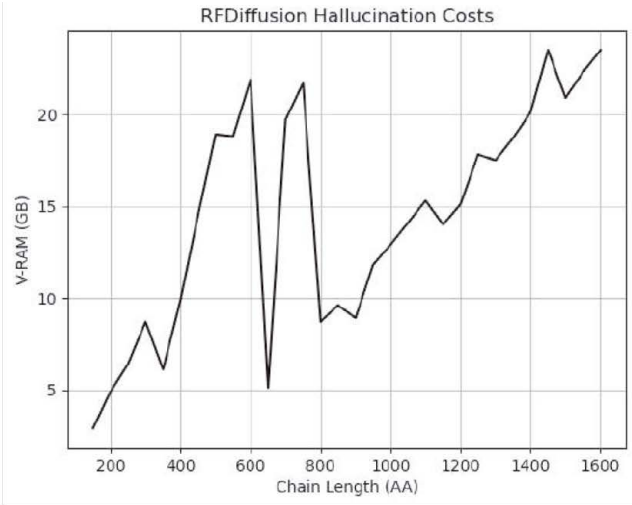

Fig. 2. Unconditionated benchmarking results graphic. As the reader can see the the consume of graphic memory (VRAM) is linear with respect to the length in amino acids of the protein to hallucinate with the exception of a pair consumption peaks that appear between the 600 and 800 aminoacids.

The explanation to these peaks in the computational costs of hallucinating proteins with these lengths could reside in the fact that RFDiffusion has an intrincate mechanism that evaluates the quality of the hallucinated proteins by using the Alphafold and ESMFold models, also the original RFDiffusion paper mentions that the complexity limit for this evaluation mechanism is around the 600 amino acids and therefore the first consumption peak is justified as internal works of the hallucination tool[18]. Meanwhile, the second consumption peak and its posterior drop in the VRAM usage could be seen as attempts by the tool to evaluate the quality of hallucinated proteins higher than 600 amino acids but given the increasing complexity of the generated protein the model decides to discard the quality evaluation mechanism in order to save computational resources and consequently begins to hallucinate proteins that have a much higher possibility of one: having no biological reliance and two: having no true functions at all.

With the purpose of display what kind of protein structres can hallucinate RFDiffusion when it is near its operational limit we show 16 different hallucinated test proteins in the figure 3.

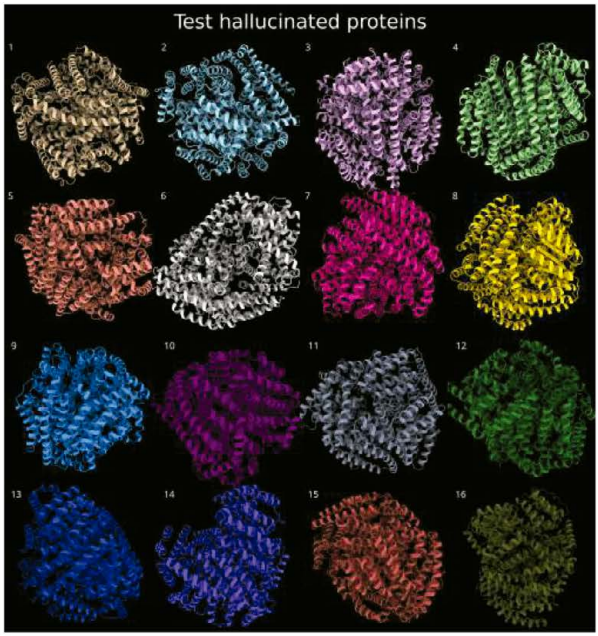

Fig. 3. Test hallucinated proteins. These proteins were hallucinated with a length 50 amino acids below the operational limit of RFDiffusion (1550 amino acids) with the purpose of display the results of a large protein hallucination.

At simple sight all of the hallucinated proteins have a sphere-like tridimensional structure with little variations between them, however; when we take a closer look into them we realize that all of these proteins are composed almost entirely of α-helices and a only some of them, such as the 8th and the 15th designs, have β-sheets that are well hidden in the interior of the hallucinated proteins.
This is rare given the fact that the α-helices trend to be composed by hydrophobic amino acids while the β-sheets trend to be composed of hydrophilic aminoacids[12, 4, 15, 1, 8, 10], so finding that the tridimensional structures of the hallucinated proteins are spheres of α-helices with no entrances to their interior nor any particular socket to accomodate other molecules to them reaffirms the possibility that when we hallucinate large proteins (>600 amino acids) with RFDiffusion the results have no biological sense.

The results of the scaffolded benchmarking of RFDiffusion can be seen in the figure 4. The graphic memory (VRAM) cost of these hallucinations reached almost of 14 GB when using the transport and sensorial proteins as templates but did not reach the 5 GB when the structural, signaling and regulatory pro-

teins were used as templates as well. To explain this situation we propose that the computational cost increase along with the complexity of the protein that is being used as template, given that the proteins that compose the chains selected from the downloaded structures are significantly different in length, 141 aminoacids in the case of chain A selected from the hemoglobin and 355 amino acids in the case of the chain A selected from the fast kinesin, the length of the hallucinated protein fragments made by RFDiffusion approached or already reached the upper limit of 200 amino acids and therefore with more the amino acids to generate the more computational resources were needed in spite of the template used.

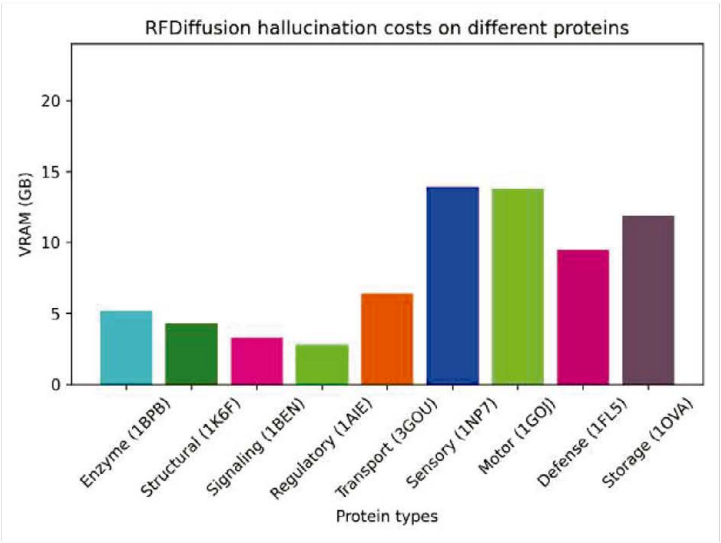

Fig. 4. Scaffolded benchmarking results graphic. The highest VRAM consumption peaks were registrated while hallucinating proteins using the cryptochrome (sensorial) and the kinesin (motor) proteins as templates.

In the figure 5 the resultant protein structures from this scaffolded benchmarking are shown. By comparing these structures with the ones generated in the unconditionated RFDiffusion benchmarking the reader can realise that there is more diversity in the secondary structures present in these scaffolded structures than just α-helices this is caused by two factors, being the first one that we are taking fragments of already known tridimensional structures that have particular secondary strcutures distributions and the scond one being that the length of the hallucinated protein fragments do not surpass the 600 amino acids treshold reported in the original paper of RFDiffusion and therefore the generated

segments can be evaluated and reshaped in order to have a biological sense with respect to the used template structures used in their design.

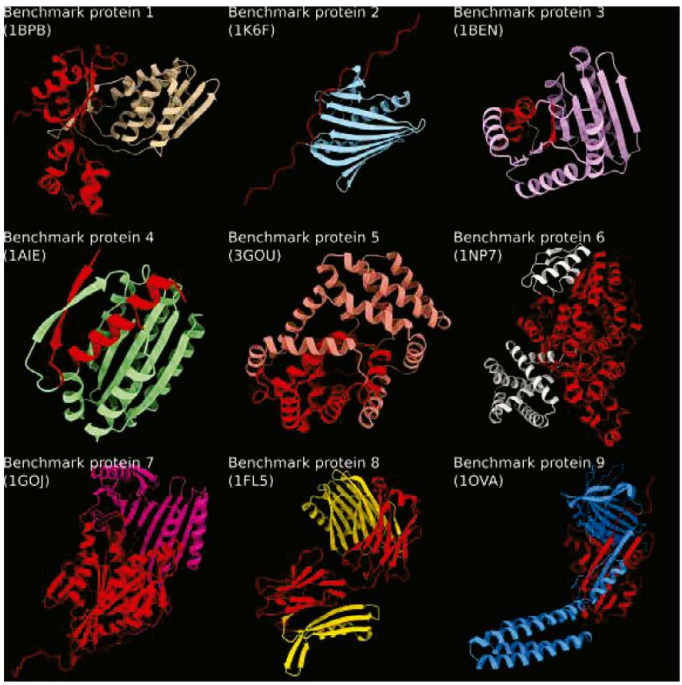

Fig. 5. Scaffolding benchmark test hallucinated proteins. The already known protein structures used in the hallucination of these proteins are signaled with red while the other colorful parts indicate the structures generated by RFDiffusion during the hallucination.

In order to confirm our suspicions about the results exposed in this work we would need to do a deep-dive into the souce code of RFDiffusion, which is a time and resource consuming task and also is not the scope of this work as we just wanted to know if the tool works in a local workstation computer.

Conclusions

The RFDiffusion model can work properly when installed into a workstation computer that has a high-end GPU and therefore making it the go-to tool to design de novo proteins in the absence of an online version. Even if it does not make this model accessible to the conventional computer users it allows to all

the scientists with access to this type of computers to design their desired proteins locally. The reader is adviced to take extreme caution when the proteins hallucinated with RFDiffusion exceed the 600 amino acids in length due to the fact that no quality evaluation mechanism is applied to these proteins due to the increasing complexity and computational costs that this task implies.
However, this problem can be overcome by adopting a protein complex design approach, we mean by this that the user can design small proteins that accomplish an specific step in a process and that do not exceed the 600 amino acids threshold and finally put these hallucinated pieces together to form a greater structure that fulfills the function of interest. With this the true potential of RFDiffusion can be unleashed as a model that can create proteins with a specific function that could not exist naturally while consuming a reasonable quantity of computational resources.

Acknowledgments. This paper was possible thanks to the biodigital innovation laboratory that allowed the access to the workstation computers where the experiments were made.

Disclosure of Interests. None of the authors of this paper have declared any kind of conflict of interests.

References

[1] Bruce Alberts et al. "Proteins". In: *Molecular Biology of The Cell*. 6th ed. Garland Science, pp. 109–173. ISBN: 978-0-8153-4432-2.

[2] Rita Berisio et al. "Crystal structure of the collagen triple helix model [(Pro-Pro-Gly)10]3". en. In: *Protein Science* 11.2 (2002). _eprint: https://onlinelibrary.wiley.com/doi/pdf/pp. 262–270. ISSN: 1469-896X. DOI: 10.1110/ps.32602. URL: https://onlinelibrary.wiley.com/doi/abs/10.1110/ps.32602 (visited on 02/16/2025).

[3] Ronald Brudler et al. "Identification of a new cryptochrome class. Structure, function, and evolution". eng. In: *Molecular Cell* 11.1 (Jan. 2003), pp. 59–67. ISSN: 1097-2765. DOI: 10.1016/s1097-2765(03)00008-x.

[4] David Eisenberg. "The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins". In: *Proceedings of the National Academy of Sciences* 100.20 (Sept. 2003). Publisher: Proceedings of the National Academy of Sciences, pp. 11207–11210. DOI: 10.1073/pnas.2034522100. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2034522100 (visited on 02/16/2025).

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. arXiv:2006.11239 [cs]. Dec. 2020. DOI: 10.48550/arXiv.2006.11239. URL: http://arxiv.org/abs/2006.11239 (visited on 02/16/2025).

[6] Ivan V. Korendovych and William F. DeGrado. "De novo protein design, a retrospective". en. In: *Quarterly Reviews of Biophysics* 53 (Jan. 2020), e3. ISSN: 0033-5835, 1469-8994. DOI: 10.1017/S0033583519000131. URL: https://www.cambridge.org/core/journals/quarterly-reviews-of-biophysics/article/de-novo-protein-design-a-retrospective/FF37903868E1651D7E61A8495FB00B50 (visited on 02/16/2025).

[7] Sidney Lyayuga Lisanza et al. "Multistate and functional protein design using RoseTTAFold sequence space diffusion". en. In: *Nature Biotechnology* (Sept. 2024). Publisher: Nature Publishing Group, pp. 1–11. ISSN: 1546-1696. DOI: 10.1038/s41587-024-02395-w. URL: https://www.nature.com/articles/s41587-024-02395-w (visited on 02/16/2025).

[8] Harvey Lodish et al. "Protein Structure and Function". In: *Molecular Cell Biology*. 8th ed. Macmillan Learning, pp. 67–129. ISBN: 978-1-4641-8339-3.

[9] P. R. Mittl, P. Chène, and M. G. Grütter. "Crystallization and structure solution of p53 (residues 326-356) by molecular replacement using an NMR model as template". eng. In: *Acta Crystallographica. Section D, Biological Crystallography* 54.Pt 1 (Jan. 1998), pp. 86–89. ISSN: 0907-4449. DOI: 10.1107/s0907444997006550.

[10] Carey L. Nesloney and Jeffery W. Kelly. "Progress towards understanding beta-sheet structure". In: *Bioorganic & Medicinal Chemistry* 4.6 (June 1996), pp. 739–766. ISSN: 0968-0896. DOI: 10.1016/0968-0896(96)00051-X. URL: https://www.sciencedirect.com/science/article/pii/096808969600051X (visited on 02/16/2025).

[11] M. R. Sawaya et al. "Crystal structure of rat DNA polymerase beta: evidence for a common polymerase mechanism". eng. In: *Science (New York, N.Y.)* 264.5167 (June 1994), pp. 1930–1935. ISSN: 0036-8075. DOI: 10.1126/science.7516581.

[12] J. Martin Scholtz and Robert L. Baldwin. "The Mechanism of alpha-Helix Formation by Peptides". en. In: *Annual Review of Biophysics and Biomolecular Structure* 21.1 (June 1992), pp. 95–118. ISSN: 1056-8700, 1545-4266. DOI: 10.1146/annurev.bb.21.060192.000523. URL: http://www.annualreviews.org/doi/10.1146/annurev.bb.21.060192.000523 (visited on 02/16/2025).

[13] G.David Smith, Ewa Ciszak, and Walter Pangborn. "A novel complex of a phenolic derivative with insulin: Structural features related to the T → R transition". en. In: *Protein Science* 5.8 (1996). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro. pp. 1502–1511. ISSN: 1469-896X. DOI: 10.1002/pro.5560050806. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.5560050806 (visited on 02/16/2025).

[14] Y.-H. Song et al. "Structure of a fast kinesin: implications for ATPase mechanism and interactions with microtubules". In: *The EMBO Journal* 20.22 (Nov. 2001), pp. 6213–6225. ISSN: 0261-4189. DOI: 10.1093/emboj/20.22.6213. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC125725/ (visited on 02/16/2025).

[15] Rajgopal Srinivasan and George D. Rose. "A physical basis for protein secondary structure". In: *Proceedings of the National Academy of Sciences* 96.25 (Dec. 1999). Publisher: Proceedings of the National Academy of Sciences, pp. 14258–14263. DOI: 10.1073/pnas.96.25.14258. URL: https://www.pnas.org/doi/abs/10.1073/pnas.96.25.14258 (visited on 02/16/2025).

[16] P. E. Stein et al. "Crystal structure of uncleaved ovalbumin at 1.95 A resolution". eng. In: *Journal of Molecular Biology* 221.3 (Oct. 1991), pp. 941–959. ISSN: 0022-2836. DOI: 10.1016/0022-2836(91)80185-w.

[17] S.S Sundaresan et al. *RCSB PDB - 3GOU: Crystal structure of dog (Canis familiaris) hemoglobin.* en-US. URL: https://www.rcsb.org/structure/3GOU (visited on 02/16/2025).

[18] Joseph L. Watson et al. "De novo design of protein structure and function with RFdiffusion". en. In: *Nature* 620.7976 (Aug. 2023). Publisher: Nature Publishing Group, pp. 1089–1100. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06415-8. URL: https://www.nature.com/articles/s41586-023-06415-8 (visited on 10/07/2024).

[19] Ling Yang et al. *Diffusion Models: A Comprehensive Survey of Methods and Applications.* arXiv:2209.00796 [cs]. Dec. 2024. DOI: 10.48550/arXiv.2209.00796. URL: http://arxiv.org/abs/2209.00796 (visited on 02/16/2025).

[20] J. Yin et al. "A comparative analysis of the immunological evolution of antibody 28B4". eng. In: *Biochemistry* 40.36 (Sept. 2001), pp. 10764–10773. ISSN: 0006-2960. DOI: 10.1021/bi010536c.