Generative artificial intelligence performs rudimentary structural biology modelling

Alexander M. Ille[1,2,3], Christopher Markosian[1,2,3], Stephen K. Burley[4,5,6,7], Michael B. Mathews[1,8], Renata Pasqualini[2,3 ‡ *], and Wadih Arap[2,9 ‡ *]

[1]School of Graduate Studies, Rutgers, The State University of New Jersey, Newark, New Jersey, USA
[2]Rutgers Cancer Institute of New Jersey, Newark, New Jersey, USA
[3]Division of Cancer Biology, Department of Radiation Oncology, Rutgers New Jersey Medical School, Newark, New Jersey, USA
[4]Research Collaboratory for Structural Bioinformatics Protein Data Bank, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA
[5]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA
[6]Rutgers Cancer Institute of New Jersey, Robert Wood Johnson Medical School, New Brunswick, New Jersey, USA
[7]Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California, San Diego, La Jolla, California, USA
[8]Division of Infectious Disease, Department of Medicine, Rutgers New Jersey Medical School, Newark, New Jersey, USA
[9]Division of Hematology/Oncology, Department of Medicine, Rutgers New Jersey Medical School, Newark, New Jersey, USA
‡ Equal contribution
** For correspondence: Renata Pasqualini (renata.pasqualini@rutgers.edu) or Wadih Arap (wadih.arap@rutgers.edu)*

Generative artificial intelligence (AI) is increasingly used by researchers in numerous fields, including the biological sciences. To date, the most commonly used tool grounded in this technology is Chat Generative Pre-trained Transformer (ChatGPT). While ChatGPT is typically applied for natural language text generation, other application modes include coding and mathematical problem-solving. We have recently reported the ability of ChatGPT to interpret the central dogma of molecular biology and the genetic code. Here we explored how ChatGPT might be able to perform rudimentary structural biology modelling. We prompted ChatGPT to model 3D structures for the 20 standard amino acids as well as an α-helical polypeptide chain, with the latter involving incorporation of the

Wolfram plugin for advanced mathematical computation. For amino acid modelling, distances and angles between atoms of the generated structures in most cases approximated to experimentally determined values. For α-helix modelling, the generated structures were comparable to that of an experimentally determined α-helical structure. However, both amino acid and α-helix modelling were sporadically error-prone and molecular complexity was not well tolerated. Despite current limitations, we show the capacity of generative AI to perform basic structural biology modelling with atomic-scale accuracy. These results provide precedent for the potential use of generative AI in structural biology as this technology continues to advance.

Introduction

Artificial intelligence (AI)-based capabilities and applications in scientific research have made remarkable progress over the past few years [1, 2]. Advances in the field of protein structure prediction have been particularly impactful: AI-based platform tools such as AlphaFold2 and RoseTTAFold are capable of processing amino acid residue sequences only as input and modelling their corresponding protein structures with accuracy comparable to lower-resolution experimentally determined structures [3-5]. AlphaFold2 and RoseTTAFold are trained on datasets of protein sequences and structures [6], and rely on neural network architectures specialized for modelling protein structures. Another category of AI-based tools for protein structure prediction are protein language models, which differ from AlphaFold2 and RoseTTAFold in that they are not trained on protein structures but rather on protein sequences [7-9]. Collectively, such protein structure prediction tools have been extensively used by researchers across various disciplines in the biological sciences and are expected to continue to add value alongside experimental structure determination [10-18].

More recently, generative AI language models, particularly Chat Generative Pre-trained Transformer (ChatGPT) [19-21], have garnered substantial interest with applications including performance of autonomous research [22], drug discovery [23-25], bioinformatic analysis [26, 27], the potential to aid in clinical diagnosis [28, 29], among others [30-32]. Unlike AlphaFold2, RoseTTAFold, and protein language models, ChatGPT is trained on natural language datasets and its neural network architecture is tailored towards understanding and generating natural language text rather than structural modelling. We have recently reported how ChatGPT interprets the central dogma of molecular biology and the genetic code [33]. While analogies can be made, the genetic code is not a natural language *per se*, yet ChatGPT seems to have at least some

inherent capability of processing it. In a similar vein, we were motivated to explore how ChatGPT performs rudimentary structural biology modelling and evaluate its capabilities and limitations in this domain. Surprisingly, we found that ChatGPT is capable of modelling the 20 standard amino acids and, with incorporation of the Wolfram plugin, a typical α-helical secondary structure element at the atomic level—albeit with some errors. To our knowledge, this is the first report to explore the capabilities of ChatGPT in rudimentary structural biology modeling.

Results

Modelling of individual amino acid structures

Amino acid residues are the components of proteins, and their atomic composition and geometric parameters have been well characterized [34-36], making them suitable candidates for rudimentary structure modelling. We therefore prompted ChatGPT-4 to model the 20 standard amino acids with minimal contextual information as input, including instructions for output in legacy Protein Data Bank (PDB) file format **(Table 1*A*, 2*A*, S1, and Fig. 1*A*)**. Multiple iterations (n=5 for each amino acid) were run by using the same input prompt to monitor consistency (see Methods). For each individual amino acid, ChatGPT-4 generated 3D structures with coordinate values for both backbone and sidechain atoms **(Fig. 1*B*)**. Generated structures contained all atoms specific to the amino acid prompted, except for a single iteration of cysteine which lacked the backbone O atom and a single iteration of methionine which lacked sidechain Cγ atom. Most amino acid structures (excluding achiral glycine) were modelled in L rather than D stereochemical configuration, while some were also modelled in planar configuration **(Fig. 1*C*)**. While the modelling favored the L-configuration, a more accurate distribution would be near exclusive L-configuration, given that D-amino acid residues are only rarely found in naturally occurring proteins [37].

Backbone bond lengths and angles of the modelled structures varied in accuracy, yet clustered in approximation to experimentally determined reference values [34, 35] **(Fig. 1, *D* and *E*)**. Moreover, all reference values fell within the standard deviations of backbone bond lengths and angles of the modelled structures. Finally, sidechain bond lengths and bond angles also varied in accuracy, yet >89% of calculated bond lengths were within 0.1 Å and >79% of calculated bond angles were within 10° of experimentally determined reference values [36, 38], again indicating remarkable precision **(Fig. 1*F-H*, Fig. S1-S4)**. Sidechain bond lengths and bond angles outside of these ranges generally occurred at random, but were notably more prevalent in the aromatic

rings of histidine and tryptophan, and the pyrrolidine component of proline. Although not entirely error-free, the ring structures of phenylalanine and tyrosine were more accurate, which may be due to the reduced complexity of their all-carbon ring composition. Amino acid modelling was also attempted with other generative AI models, including Google's Gemini Pro [39], Anthropic's Claude 2 [40], and OpenAI's ChatGPT-3.5 [19], but results were inaccurate and/or inconsistent (data not shown). Collectively, these findings demonstrate that ChatGPT-4 is capable of structurally modelling single amino acid residues in a manner that resembles their experimentally determined structures, though not without sporadic errors including incorrect stereochemistry and geometric distortion, which would require—at least presently—human operator curation or supervision to ensure fidelity.

Table 1. ChatGPT-4 dialog used for structural modelling. (*A*) Prompt used for modelling the structures of each of the 20 amino acids. The same prompt was used for each amino acid by replacing "[amino acid]" with the full individual amino acid name. **(*B*)** Prompt used for modelling the α-helical polypeptide structure with ChatGPT-4 running with the Wolfram plugin for enhanced mathematical computation.

(*A*) Amino acid structure modelling with ChatGPT-4	
Prompt: What are the typical distances and angles between the atoms of one [amino acid] residue in a protein? Based on these values, generate a structure in PDB file format for one [amino acid] residue. Ensure coordinate values have three decimal places and omit hydrogen atoms.	
(*B*) α-helix structure modelling with ChatGPT-4 running the Wolfram plugin	
Initial Prompt: What are the typical geometric attributes (including distances and angles) between the backbone atoms of a typical alpha-helical polypeptide chain? Based on this information, generate a structure in PDB file format for an alpha-helical polypeptide chain 10 residues in length including only alpha carbon atoms. Ensure coordinate values have 3 decimal places and use the Wolfram plug-in for coordinate calculations but not for PDB file formatting.	
Refinement Prompts	The generated structure does not resemble an alpha-helix. Try again.
	The diameter of the helix is too large.
	The diameter of the helix is too small.
	The pitch (residues per turn) of the helix is too large.
	The pitch (residues per turn) of the helix is too small.

Table 2. Exemplary coordinate output from ChatGPT-4. Within responses to modelling prompts **(Table 2, S1, and S2)** ChatGPT-4 provided coordinates for generated structures in PDB file format, as shown below with output examples for **(A)** the arginine amino acid structure and **(B)** the α-helix structure.

(A) Output for arginine structure PDB file format									**(B) Output for α-helix structure PDB file format**								
ATOM	1	N	ARG	A	1	0.000	0.000	0.000	ATOM	1	CA	ALA	A	1	2.300	0.000	0.000
ATOM	2	CA	ARG	A	1	1.460	0.000	0.000	ATOM	2	CA	ALA	A	2	-0.399	2.265	1.500
ATOM	3	C	ARG	A	1	2.160	1.370	0.000	ATOM	3	CA	ALA	A	3	-2.161	-0.787	3.000
ATOM	4	O	ARG	A	1	2.160	2.170	-1.080	ATOM	4	CA	ALA	A	4	1.150	-1.992	4.500
ATOM	5	CB	ARG	A	1	2.090	-1.180	-0.670	ATOM	5	CA	ALA	A	5	1.762	1.478	6.000
ATOM	6	CG	ARG	A	1	2.790	-2.440	0.110	ATOM	6	CA	ALA	A	6	-1.762	1.478	7.500
ATOM	7	CD	ARG	A	1	3.420	-3.620	-0.560	ATOM	7	CA	ALA	A	7	-1.150	-1.992	9.000
ATOM	8	NE	ARG	A	1	4.120	-4.880	0.220	ATOM	8	CA	ALA	A	8	2.161	-0.787	10.500
ATOM	9	CZ	ARG	A	1	4.750	-5.170	1.380	ATOM	9	CA	ALA	A	9	0.399	2.265	12.000
ATOM	10	NH1	ARG	A	1	4.750	-4.450	2.490	ATOM	10	CA	ALA	A	10	-2.300	0.000	13.500
ATOM	11	NH2	ARG	A	1	5.350	-6.340	1.550									

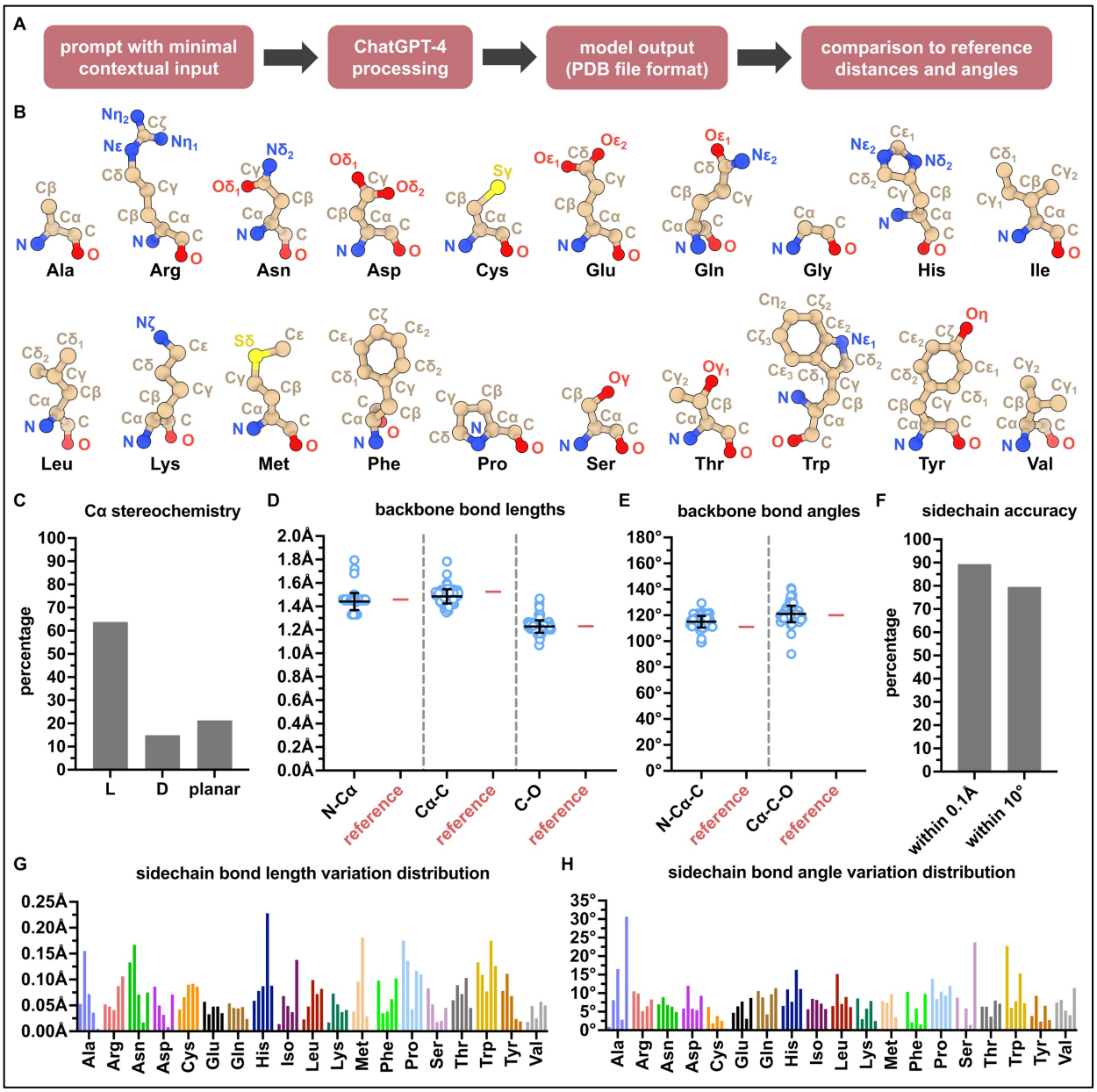

Figure 1. Modelling the 3D structures of the 20 standard amino acids with ChatGPT-4. (*A*) Procedure for structure modelling and analysis. **(*B*)** Exemplary 3D structures of each of the 20 amino acids modelled by ChatGPT-4. **(*C*)** Cα stereochemistry of modelled amino acids including L and D configurations as well as nonconforming planar, n=5 per amino acid excluding glycine (total n=94). **(*D*)** Modelled amino acid backbone bond lengths relative to experimentally determined reference values (red), n=5 per amino acid (total n=100 N-Cα; 100 Cα-C; 99 C-O). Data shown as means ± SD. **(*E*)** Modelled amino acid backbone bond angles relative to experimentally determined reference values (red), n=5 per amino acid (total n=100 N-Cα-C; 99 Cα-C-O). Data shown as means ± SD. **(*F*)** Sidechain accuracy of modelled amino acid structures

in terms of bond lengths (within 0.1 Å) and bond angles (within 10°) relative to experimentally determined reference values, n=5 per amino acid. See Methods for experimentally determined references. ***(G-H)*** Distributions of sidechain bond length and angle variation relative to experimentally determined reference values for each amino acid, excluding glycine. Bars represent the mean bond length or angle variation for each of the five iterations per amino acid. One of the methionine iterations was excluded due to lacking the sidechain Cγ atom.

Modelling of an α-helix structure

The α-helix is the most commonly occurring and extensively studied secondary structure element found in proteins [41-44]. Thus, we next prompted ChatGPT-4 to model an α-helical polypeptide chain, but were unable to obtain accurate structures despite multiple attempts with various prompts. We then incorporated the Wolfram plugin, a mathematical computation extension developed by Wolfram-Alpha for use with ChatGPT [45]. ChatGPT-4 used together with the Wolfram plugin was able to model a 10-residue α-helical structure and output the result in legacy PDB file format with minimal contextual information as input **(Table 1*B*, 2*B*, S2, and Fig. 2*A*)**. Multiple iterations were run by using the same input prompt to monitor consistency, and up to two prompt-based refinements after the first attempt were permitted per iteration for improved accuracy (see Methods). To reduce complexity, only Cα atoms were modelled.

ChatGPT-4 arbitrarily assigned all residues as alanine, which was likely done for the sake of simplicity, but nevertheless aligns well with the fact that alanine has the greatest α-helix propensity of all 20 standard amino acids [44]. Remarkably, accuracy of the modelled α-helix was comparable to an experimentally determined α-helical structure consisting of 10 consecutive alanine residues (PDB ID: 1L64) [46] **(Fig. 2, *B* and *C*)**. Forty-four percent of modelled structures had a root-mean-square deviation (RMSD) of < 0.5 Å relative to the reference experimental structure on the first attempt, and 88% had a RMSD of < 0.5 Å after two prompt-based refinements **(Fig. 2*D*)**. The structures generated by ChatGPT-4 were also compared to poly-alanine α-helix structures modelled by AlphaFold2, ChimeraX, and PyMOL, and the lowest RMSDs (*i.e.*, greatest structural similarity) were found to be between the AlphaFold2 and ChimeraX structures **(Fig. 2*B*)**. As with amino acid modelling, α-helix was also attempted with other generative AI models, including Google's Gemini Pro [39], Anthropic's Claude 2 [40], and OpenAI's ChatGPT-3.5 [19], but results were inaccurate and/or inconsistent (not shown), data indicating specifity for ChatGPT-4. Taken together, these results demonstrate the capability of ChatGPT-4, with incorporation of the Wolfram plugin, to predict the atomic level structure of an α-helix.

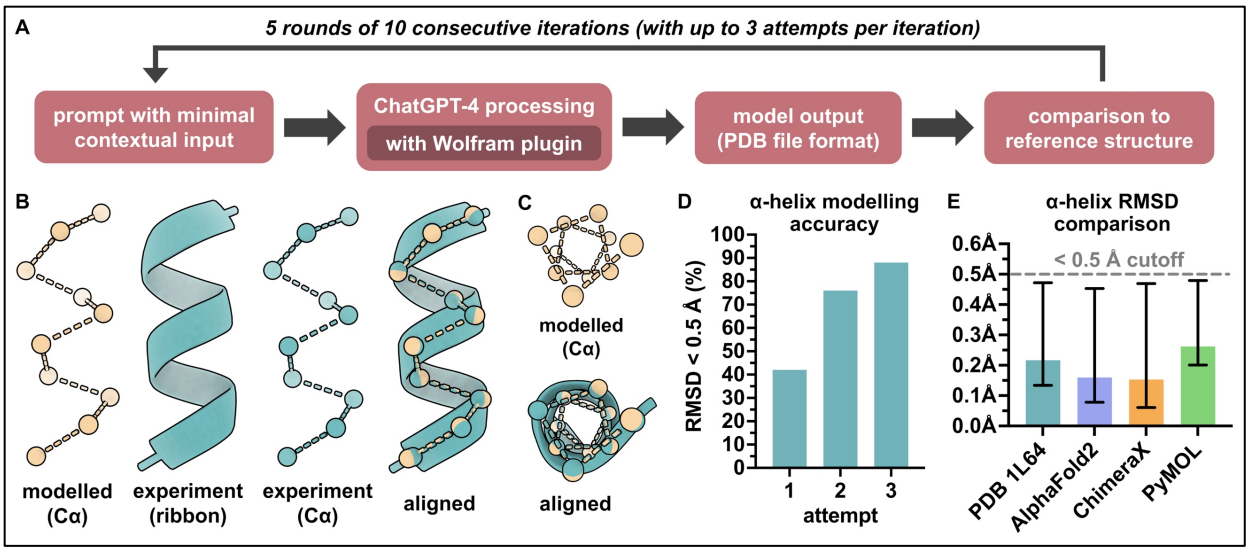

Figure 2. Modelling the 3D structure of an α-helical polypeptide structure with ChatGPT-4. **(*A*)** Procedure for structure modelling and analysis. **(*B*)** Exemplary 3D structure of a modelled α-helix (beige), an experimentally determined α-helix reference structure (PDB ID 1L64) (teal), and their alignment (RMSD = 0.147 Å). **(*C*)** Top-down view of modelled and experimental α-helices from panel B. **(*D*)** Accuracy of α-helix modelling as measured by number of attempts (including up to two refinements following the first attempt) required to generate a structure with RMSD < 0.5 Å relative to the experimentally determined reference structure, n = 5 rounds of 10 consecutive iterations (total n = 50 models). **(*E*)** Comparison of RMSDs between ChatGPT-4 α-helix structures and the experimentally determined α-helix structure, the AlphaFold2 α-helix structure, the ChimeraX α-helix structure, and the PyMOL α-helix structure. Only structures with RMSD < 0.5 Å (dashed grey line) relative to each reference structure are included (88% included in reference to PDB ID 1L64; 90% to AlphaFold; 90% to ChimeraX; 88% to PyMOL). Data shown as means ± range.

Discussion

The proof-of-concept findings reported here demonstrate the current capabilities and limitations of ChatGPT-4 for rudimentary structural biology modelling. Performance for modelling of the 20 standard amino acids was favorable in terms of atom composition, bond lengths, and bond angles. However, stereochemical configuration propensity and modelling of ring structures require improvement. Performance for α-helix modelling, with incorporation of advanced mathematical computation from the Wolfram plugin, was also favorable. While the requirement of prompt-based

refinements may be viewed as a limitation, they may also serve as a means by which the user can optimize and modify a structure. Nonetheless, improvements are required in the capacity to model more complex all-atom structures, not only Cα backbone atoms.

Of note, these structural modelling capabilities also raise the question of modelling methodology, especially since ChatGPT-4 was not explicitly developed for this purpose. It would be challenging to provide a precise answer for this, and several (proprietary or non-proprietary) computational methods may be involved. For instance, ChatGPT-4 may be utilizing pre-existing atomic coordinate information present in its broad training dataset, which includes "publicly available data (such as internet data) and data licensed from third-party providers" [20]. However, this reasoning does not adequately explain the geometric variability observed in the predicted structures, and why structural complexity appears to be a limiting factor. The modelling may also be performed *ab initio*, given that the generated responses often articulate geometric parameters (*e.g.*, specific bond lengths and angles, number of amino acid residues per α-helix turn, α-helix diameter, etc.) in addition to providing atomic coordinates **(Table S1 and S2)**. Alternatively, the modelling methodology may involve both the use of pre-existing coordinates plus *ab initio* computation.

The comparison of the α-helix model generated by ChatGPT-4 with those generated by other computational tools was quite revealing. AlphaFold2, as mentioned above, predicts structures based on training data consisting of protein sequences and 3D structures, and was developed specifically for modelling protein structures. In addition to their dedicated molecular analysis capabilities, ChimeraX and PyMOL may be used to model basic, idealized secondary structure elements in a manner which narrowly considers precise predefined geometries, thus providing accurate α-helix structures. Despite not being explicitly developed to model atomic coordinates for α-helical segments of protein chains, ChatGPT-4 generated an α-helix with accuracy comparable to the structures modeled by the above tools. The requirement of adding the Wolfram plugin likely suggests that mathematical computation is heavily relied upon by ChatGPT-4 for α-helix modelling. Yet, α-helix structural properties and self-instruction are generated by ChatGPT-4 prior to engaging the Wolfram plugin **(Table S2)**, suggesting that some degree of "reasoning" might perhaps be involved. So-called reasoning, in this regard, refers to the documented performance of benchmark reasoning exercises by ChatGPT [19, 20], and it should be noted that there is ongoing debate about what constitutes reasoning as it pertains to AI [47, 48].

Considering both strengths and weaknesses, the structural modelling capabilities of ChatGPT represent an intriguing aspect of the unprecedented advancement of generative AI, a transformative technology presumably still in its infancy. While this modelling is rudimentary and

currently of limited practical utility, it establishes an immediate and direct precedent for applying this natural language technology in structural biology as generative AI language models undergo continued development and specialization. In the interim, further research on the capabilities and limitations of generative AI is merited, not only in structural biology but also for other potential applications in the biological sciences.

Methods

Prompt-based modelling with ChatGPT

Modelling of individual amino acid structures was performed by challenging ChatGPT-4 [20, 21] with a single prompt **(Table 1*A*)**, one amino acid residue at a time. For each individual amino acid, the same prompt was used for five consecutive iterations with each iteration initiated in a new dialog. Modelling of α-helix structures was performed by challenging ChatGPT-4 running the Wolfram plugin [45] with an initial prompt followed by up to two refinement prompts in the same dialog, for a total of up to three attempts **(Table 1*B*)**. The same prompt was used for five rounds of ten consecutive iterations with each iteration initiated in a new dialog.

Analysis of generated structures

Structures were analyzed by using the UCSF ChimeraX version 1.6.1 [38]. For amino acid structures, the “distance” (for bond lengths) and “angle” (for bond angles) commands were used with atom specification tailored to each amino acid type. For α-helix structures, the matchmaker tool was used for alignment and RMSD determination across all atom pairs by using its default parameters (*i.e.*, best-aligning pair). Experimentally determined amino acid backbone bond lengths (N-Cα, 1.458Å; Cα-C, 1.525Å; C-O 1.231Å) [34], backbone bond angles (N-Cα-C, 111.0°; Cα-C-O, 120.1°) [35], and sidechain bond lengths and bond angles **(Table S3)** [36] served as references for evaluating predicted amino acid structures. Sidechain bond lengths and bond angles, obtained from a backbone-dependent rotamer library built into ChimeraX, represent those in which the backbone dihedral angles are ω=180° (typical of trans peptide bonds), φ=180°, and ψ=180° [36, 38]. An experimentally determined α-helical structure consisting of 10 consecutive alanine residues, detected within an engineered form of bacteriophage T4 lysozyme (PDB ID 1L64) [46], was used as the reference for evaluating the modelled α-helix structures. The AlphaFold2 α-helix structure was modelled using ColabFold [49] through ChimeraX run with

default parameters (*i.e.*, minimization and without templates usage) using an elongated polyalanine sequence as input **(Table S4 and Fig. S5*A*)**. The ChimeraX and PyMOL (version 2.5.7) α-helix structures were modelled using the build structure command and fab command, respectively, each by using a 10-residue alanine sequence as input and run with default α-helix parameters (*i.e.*, backbone dihedral angles set to φ=-57° and ψ=-47°) **(Fig. S5, *B* and *C*)**.

Author Contributions

A. M. I. conceptualization; A. M. I., C. M., and S. K. B. methodology; A. M. I. and C. M. investigation; A. M. I. data curation; A. M. I., C. M., S. K. B., M. B. M., R. P., W. A. formal analysis; A. M. I. and C. M. writing - original draft; S. K. B., M. B. M., R. P., and W. A. writing - review & editing; S. K. B., R. P. and W. A. funding acquisition; R. P. and W. A. overall project supervision.

Funding and Additional Information

This work was supported by core services from the Cancer Center Support Grant of the Rutgers Cancer Institute of New Jersey (P30CA072720), by the National Institutes of Health (R01CA226537 to R.P. and W.A.), and by the Levy-Longenbaugh Donor-Advised Fund (to R.P. and W.A.). RCSB Protein Data Bank is jointly funded by the National Science Foundation (DBI-1832184, PI: S.K. Burley), the US Department of Energy (DE-SC0019749, PI: S.K. Burley), and the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, and the National Institute of General Medical Sciences of the National Institutes of Health (R01GM133198, PI: S.K. Burley).

Conflict of Interest

R.P. and W.A. are founders and equity shareholders of PhageNova Bio. R.P. is Chief Scientific Officer and a paid consultant of PhageNova Bio. R.P. and W.A. are founders and equity shareholders of MBrace Therapeutics; R.P. and W.A. serve as paid consultants for MBrace Therapeutics. R.P. and W.A. have Sponsored Research Agreements (SRAs) in place with PhageNova Bio and with MBrace Therapeutics. These arrangements are managed in accordance with the established institutional conflict-of-interest policies of Rutgers, The State University of New Jersey. This study falls outside of the scope of these SRAs.

References

1. Jordan, M.I. and T.M. Mitchell, *Machine learning: Trends, perspectives, and prospects.* Science, 2015. **349**(6245): p. 255-60.
2. Wang, H., et al., *Scientific discovery in the age of artificial intelligence.* Nature, 2023. **620**(7972): p. 47-60.
3. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
4. Baek, M., et al., *Accurate prediction of protein structures and interactions using a three-track neural network.* Science, 2021. **373**(6557): p. 871-876.
5. Kryshtafovych, A., et al., *Critical assessment of methods of protein structure prediction (CASP)-Round XIV.* Proteins, 2021. **89**(12): p. 1607-1617.
6. Burley, S.K. and H.M. Berman, *Open-access data: A cornerstone for artificial intelligence approaches to protein structure prediction.* Structure, 2021. **29**(6): p. 515-520.
7. Elnaggar, A., et al., *ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning.* IEEE Trans Pattern Anal Mach Intell, 2022. **44**(10): p. 7112-7127.
8. Chowdhury, R., et al., *Single-sequence protein structure prediction using a language model and deep learning.* Nat Biotechnol, 2022. **40**(11): p. 1617-1623.
9. Lin, Z., et al., *Evolutionary-scale prediction of atomic-level protein structure with a language model.* Science, 2023. **379**(6637): p. 1123-1130.
10. Bordin, N., et al., *AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms.* Commun Biol, 2023. **6**(1): p. 160.
11. Bordin, N., et al., *Novel machine learning approaches revolutionize protein knowledge.* Trends Biochem Sci, 2023. **48**(4): p. 345-359.
12. Mosalaganti, S., et al., *AI-based structure prediction empowers integrative structural analysis of human nuclear pores.* Science, 2022. **376**(6598): p. eabm9506.
13. Fontana, P., et al., *Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold.* Science, 2022. **376**(6598): p. eabm9326.
14. Read, R.J., et al., *AlphaFold and the future of structural biology.* Acta Crystallogr F Struct Biol Commun, 2023. **79**(Pt 7): p. 166-168.
15. Edich, M., et al., *The impact of AlphaFold2 on experimental structure solution.* Faraday Discuss, 2022. **240**(0): p. 184-195.
16. Varadi, M., et al., *AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences.* Nucleic Acids Res, 2023.
17. Varadi, M. and S. Velankar, *The impact of AlphaFold Protein Structure Database on the fields of life sciences.* Proteomics, 2023. **23**(17): p. e2200128.
18. Burley, S.K., W. Arap, and R. Pasqualini, *Predicting Proteome-Scale Protein Structure with Artificial Intelligence.* N Engl J Med, 2021. **385**(23): p. 2191-2194.
19. Brown, T.B., et al., *Language Models are Few-Shot Learners.* arXiv, 2020.
20. OpenAI, *GPT-4 Technical Report.* arXiv, 2023.
21. OpenAI. *Introducing ChatGPT*. 2022 [cited 2023 June 15, 2023]; Available from: https://openai.com/blog/chatgpt.
22. Boiko, D.A., et al., *Autonomous chemical research with large language models.* Nature, 2023. **624**(7992): p. 570-578.
23. Gurwitz, D. and N. Shomron, *Artificial intelligence utility for drug development: ChatGPT and beyond.* Drug Dev Res, 2023.
24. Wang, R., H. Feng, and G.W. Wei, *ChatGPT in Drug Discovery: A Case Study on Anticocaine Addiction Drug Development with Chatbots.* J Chem Inf Model, 2023. **63**(22): p. 7189-7209.

25. Chakraborty, C., M. Bhattacharya, and S.S. Lee, *Artificial intelligence enabled ChatGPT and large language models in drug target discovery, drug discovery, and development.* Mol Ther Nucleic Acids, 2023. **33**: p. 866-868.
26. Lubiana, T., et al., *Ten quick tips for harnessing the power of ChatGPT in computational biology.* PLoS Comput Biol, 2023. **19**(8): p. e1011319.
27. Shue, E., et al., *Empowering beginners in bioinformatics with ChatGPT.* Quant Biol, 2023. **11**(2): p. 105-108.
28. Eriksen, A.V., S. Möller, and J. Ryg, *Use of GPT-4 to Diagnose Complex Clinical Cases.* NEJM AI, 2023. **1**(1).
29. Hirosawa, T., et al., *ChatGPT-Generated Differential Diagnosis Lists for Complex Case-Derived Clinical Vignettes: Diagnostic Accuracy Evaluation.* JMIR Med Inform, 2023. **11**: p. e48808.
30. van Dis, E.A.M., et al., *ChatGPT: five priorities for research.* Nature, 2023. **614**(7947): p. 224-226.
31. Graf, A. and R.E. Bernardi, *ChatGPT in Research: Balancing Ethics, Transparency and Advancement.* Neuroscience, 2023. **515**: p. 71-73.
32. Thapa, S. and S. Adhikari, *ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls.* Ann Biomed Eng, 2023. **51**(12): p. 2647-2651.
33. Ille, A.M. and M.B. Mathews, *AI interprets the Central Dogma and Genetic Code.* Trends Biochem Sci, 2023. **48**(12): p. 1014-1018.
34. Engh, R.A. and R. Huber, *Accurate bond and angle parameters for X-ray protein structure refinement.* Acta Crystallographica Section A, 1991. **47**(4): p. 392-400.
35. Berkholz, D.S., et al., *Conformation dependence of backbone geometry in proteins.* Structure, 2009. **17**(10): p. 1316-25.
36. Shapovalov, M.V. and R.L. Dunbrack, Jr., *A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions.* Structure, 2011. **19**(6): p. 844-58.
37. Fujii, N. and T. Saito, *Homochirality and life.* Chem Rec, 2004. **4**(5): p. 267-78.
38. Meng, E.C., et al., *UCSF ChimeraX: Tools for structure building and analysis.* Protein Sci, 2023. **32**(11): p. e4792.
39. Google, *Gemini: A Family of Highly Capable Multimodal Models.* arXiv, 2023.
40. Antrhopic, *Model Card and Evaluations for Claude Models.* 2023.
41. Doig, A.J., et al., *Structure, stability and folding of the alpha-helix.* Biochem Soc Symp, 2001(68): p. 95-110.
42. Eisenberg, D., *The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins.* Proc Natl Acad Sci U S A, 2003. **100**(20): p. 11207-10.
43. Pauling, L., R.B. Corey, and H.R. Branson, *The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain.* Proc Natl Acad Sci U S A, 1951. **37**(4): p. 205-11.
44. Pace, C.N. and J.M. Scholtz, *A helix propensity scale based on experimental studies of peptides and proteins.* Biophys J, 1998. **75**(1): p. 422-7.
45. Wolfram, S., *ChatGPT Gets Its "Wolfram Superpowers"!* 2023.
46. Heinz, D.W., W.A. Baase, and B.W. Matthews, *Folding and function of a T4 lysozyme containing 10 consecutive alanines illustrate the redundancy of information in an amino acid sequence.* Proc Natl Acad Sci U S A, 1992. **89**(9): p. 3751-5.
47. Eisenstein, M., *A test of artificial intelligence.* Nature, 2023.
48. Biever, C., *ChatGPT broke the Turing test - the race is on for new ways to assess AI.* Nature, 2023. **619**(7971): p. 686-689.
49. Mirdita, M., et al., *ColabFold: making protein folding accessible to all.* Nat Methods, 2022. **19**(6): p. 679-682.