NanoGPT2: De novo nanobody sequence generation using a fine-tuned protein language model

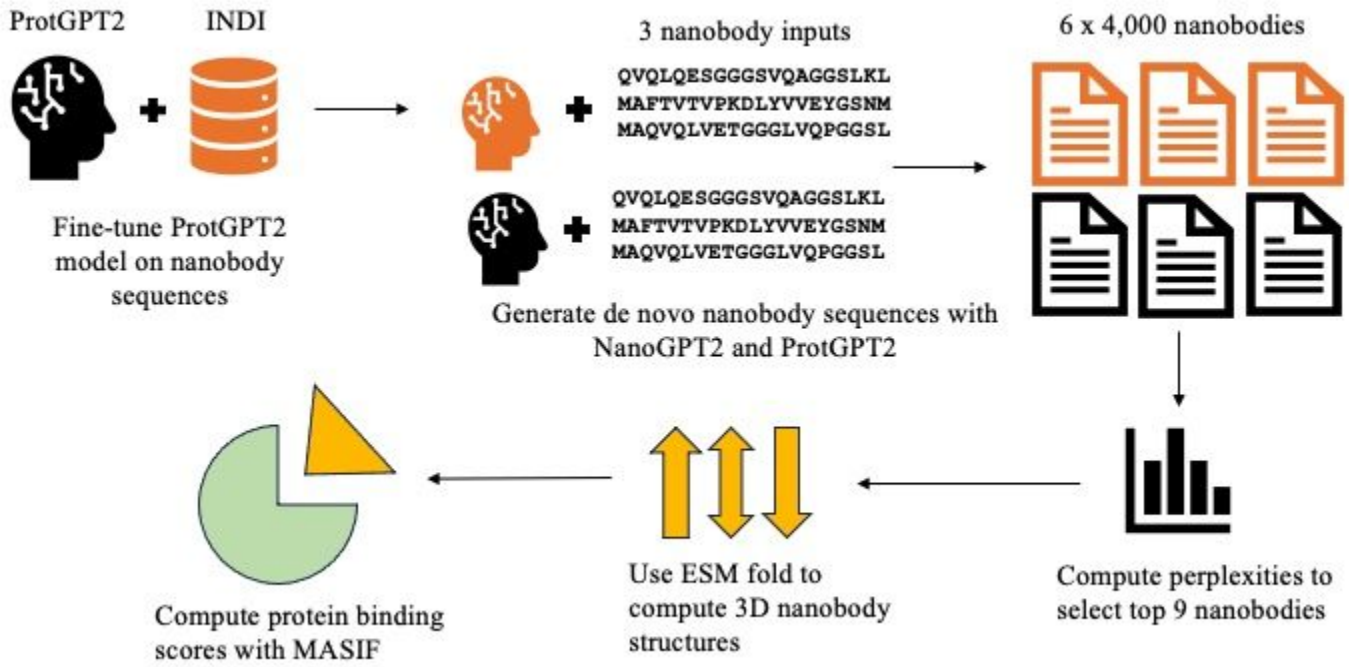
By: Will Lounsbery-Scaife and Nick Hayes

COMS 4762, Final Video Presentation

Introduction & Purpose

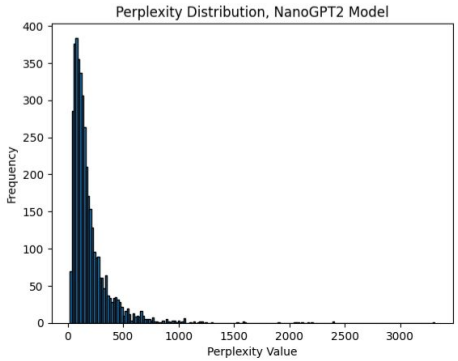
- Within the past few years, **deep learning** has been applied to **de novo protein generation tasks** with much success.
- Last year, **ProtGPT2**, a large-language model (LLM) built on the GPT2 architecture, was able to produce de novo protein sequences with **striking similarity to those found in nature**
- **Nanobodies**, a class of proteins that are **similar to antibodies but more simple**, are known for having many **therapeutic effects** due to their size and properties. However, effective **de novo synthesis of nanobodies remains challenging**.
- In this project, we successfully **fine-tuned ProtGPT2 to create NanoGPT2, a new LLM intended to facilitate de novo nanobody synthesis**.
- Additionally, we assess the **structural integrity** and **predicted protein-protein interaction affinities** of our **best nanobody sequences** using **ESMFold** and **MaSIF** to demonstrate the potential downstream **clinical relevance** of our approach.

Methods & Data

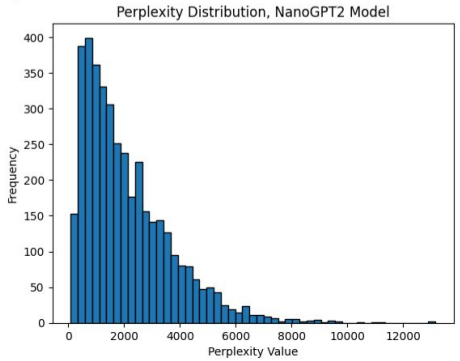


Results: Fine-tuned vs. Base Model

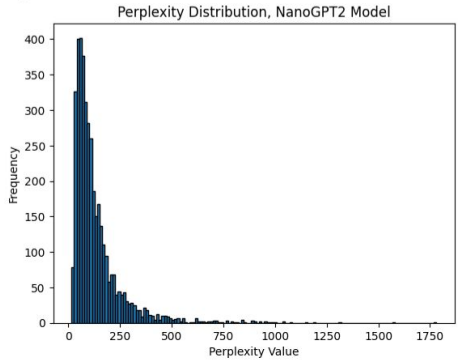
a)



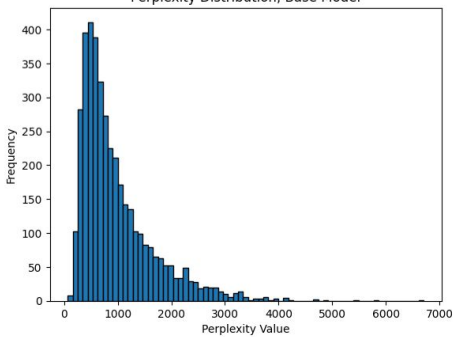
b)



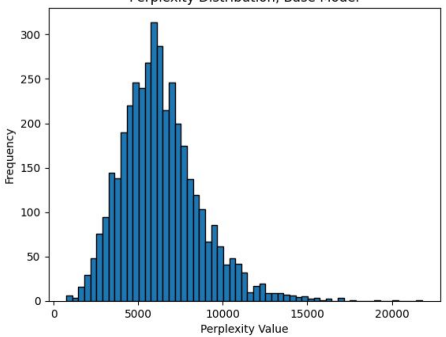
c)



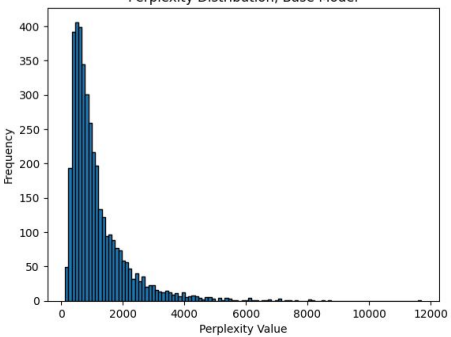
a)



b)

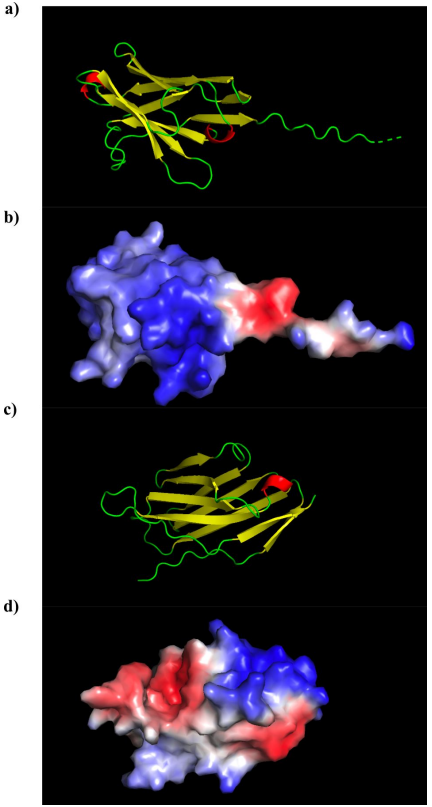


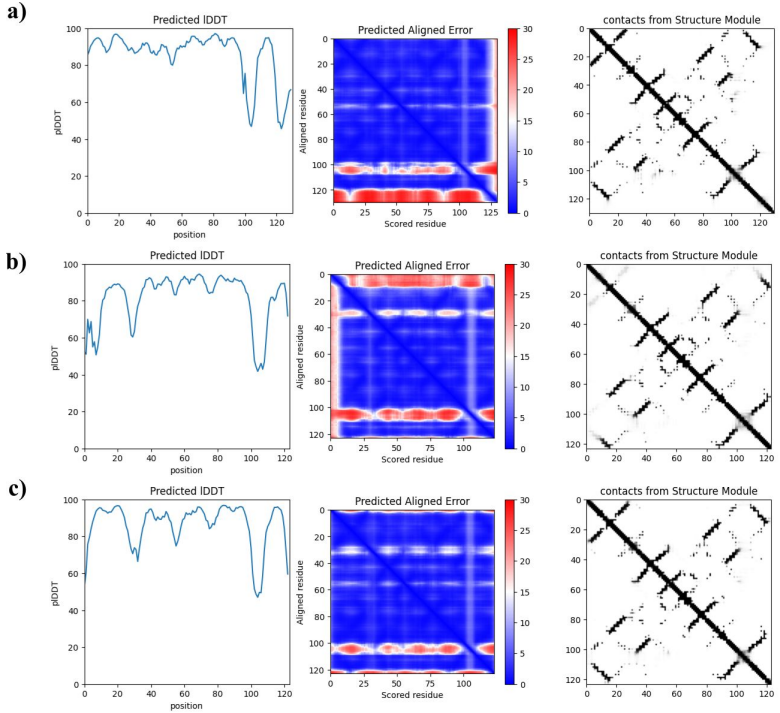
c)



Nanobody ID	Perplexity Score
7OM4, 1	18.7
7OM4, 2	19.6
7OM4, 3	20.4
5JDS, 1	82.5
5JDS, 2	100.2
5JDS, 3	100.6
5DWX, 1	14.6
5DWX, 2	15.0
5DWX, 3	15.6

Results: ESM and MaSIF predictions





Nanobody ID	Top Score	Protein ID	Chemical Function
7OM4, 1	0.9830	3F8H	Unknown
7OM4, 2	0.9630	2J6X	Oxidoreductase
7OM4, 3	0.8446	3EQX	DNA binding
5JDS, 1	0.9868	1NJJ	Lyase
5JDS, 2	0.9969	2EJ0	Transferase
5JDS, 3	0.9896	1FL6	Immune system
5DWX, 1	0.9587	1VJL	Unknown
5DWX, 2	0.9841	2FU5	Signaling
5DWX, 3	0.9816	2QUY	Hydrolase

Discussion: NanoGPT2 was successful

- Based on our results, we concluded that **NanoGPT2** was **far superior** at the de **novo nanobody generation tasks** compared to the baseline model, **ProtGPT2**.
- For each of the three reference nanobody groups selected, we noted **statistically significant perplexity scores against the baseline (p-value=0.00)**, with a substantial separation in means.
- We found a notable **clinically relevant partner protein** was associated with **African Sleeping Sickness**
- Furthermore, our **top scoring sequences** were found to be structurally viable
- **Future work** could **involve attempting to express these these nanobodies in a wetlab setting**

References

1. Ferruz, N., Schmidt, S., & Höcker, B. (2022b). ProtGPT2 is a deep unsupervised language model for protein design. Nature Communications, 13(1). https://doi.org/10.1038/s41467-022-32007-7
2. Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., & Naik, N. (2023). Large language models generate functional protein sequences across diverse families. Nature Biotechnology, 41(8), 1099–1106. https://doi.org/10.1038/s41587-022-01618-2
3. Leem, J., Mitchell, L., Farmery, J. H., Barton, J., & Galson, J. D. (2022). Deciphering the language of antibodies using self-supervised learning. Patterns, 3(7), 100513. https://doi.org/10.1016/j.patter.2022.100513
4. Liu, M., Li, L., Jin, D., & Liu, Y. (2021). Nanobody—A versatile tool for cancer diagnosis and therapeutics. WIREs Nanomedicine and Nanobiotechnology, 13(4). https://doi.org/10.1002/wnan.1697
5. Jin, B., Odongo, S., Radwanska, M., & Magez, S. (2023). Nanobodies: A Review of generation, Diagnostics and Therapeutics. International Journal of Molecular Sciences, 24(6), 5994. https://doi.org/10.3390/ijms24065994
6. Arras, P., Yoo, H. S., Pekar, L., Clarke, T. L., Friedrich, L., Schröter, C., Schanz, J., Tonillo, J., Siegmund, V., Doerner, A., Krah, S., Guarnera, E., Zielonka, S., & Evers, A. (2023). AI/ML combined with next-generation sequencing of VHH immune repertoires enables the rapid identification of de novo humanized and sequence-optimized single domain antibodies: a prospective case study. Frontiers in Molecular Biosciences, 10. https://doi.org/10.3389/fmolb.2023.1249247
7. Deszyński, Piotr, et al. "INDI—integrated nanobody database for immunoinformatics." Nucleic Acids Research 50.D1 (2022): D1273-D1281. https://doi.org/10.1093/nar/gkab1021
8. Rives, Alexander, et al. "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences." Proceedings of the National Academy of Sciences 118.15 (2021): e2016239118. https://doi.org/10.1073/pnas.2016239118
9. Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M. M., & Correia, B. E. (2019). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nature Methods, 17(2), 184–192. https://doi.org/10.1038/s41592-019-0666-6
10. Jackson, L. N., Goldsmith, E. J., & Phillips, M. A. (2003). X-ray Structure Determination of Trypanosoma brucei Ornithine Decarboxylase Bound to d-Ornithine and to G418. Journal of Biological Chemistry, 278(24), 22037–22043. https://doi.org/10.1074/jbc.m300188200
11. Remmert, M., Biegert, A., Hauser, A. et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9, 173–175 (2012). https://doi.org/10.1038/nmeth.1818
12. Fu, Limin, et al. "CD-HIT: accelerated for clustering the next-generation sequencing data." Bioinformatics 28.23 (2012): 3150-3152. https://doi.org/10.1093/bioinformatics/bts565