CASNOVO: Advanced Gene Editing Utilizing Machine Learning, Large Language Models (LLM), and Natural Language Processing (NLP)

Sanvi Asaigoli
William G. Enloe High School, Raleigh, NC

ABSTRACT

Designing Cas9 proteins that possess the stability and specificity required for CRISPR-Cas9 poses a tedious and resource-expensive challenge for genetic medicine development. Sub-par fidelity of Cas9 proteins results in off-target effects, compromising the safety of gene therapies. Conventional laboratory protein engineering has improved; however, the rise in computational design of endonuclease proteins and peptide-based binders with machine learning has presented a more efficient avenue for developing novel therapeutics using protein-based large language models (LLMs). Methods shown in ProtBERT[(6)], PepMLM[(7)], and ProLLaMA[(5)] show transformer-based language model architectures for the process of generating de novo proteins and predicting protein structure and function with high novelty, stability, and low perplexity. In this work, we present CASNOVO, a novel, fine-tuned LLM tool for the generation and validation of de novo Cas9 proteins.

The proposed tool fine-tunes a transformer large language model via training on a dataset of over 100,000 existing Cas9 proteins manually curated from the UniProt Database. The full-length Cas9 sequences provide us with information about their properties. ProtGPT2 is a transformer neural network trained on millions of naturally occurring amino acid sequences. Its algorithm was leveraged in the model's training, while generation involved using learned information by the model to form viable proteins. Validation efforts involved the usage of AlphaFold and ESMFold, bioinformatics tools specialized in protein engineering. Generated sequences were pushed through these algorithms and received various metrics measuring the viability of the novel proteins. This includes the predicted local difference distance test (pLDDT), Sequence Coverage, *and* post-translational modification scores (PTM).

INTRODUCTION

By allowing genetic material to be added, removed, or altered in specific locations in the genome, it holds great promise in showing progress in a wide range of research in clinical trials to cure genetic disorders such as hemophilia, cystic fibrosis, and, recently, an FDA-approved cure for sickle cell disease. CRISPR (clustered regularly interspaced short palindromic repeats) technology is a groundbreaking advancement in biotechnology that is being used to improve medicine, agriculture, and synthetic biology. Cas9 is a form of a Cas protein which is a type of enzyme with the ability to cleave DNA, originally found in naturally occurring bacterial immune systems, and a vital component of the CRISPR-Cas9 gene editing system. The sgRNA (single-guide RNA) is used to guide the Cas9 protein to the specific target DNA sequence to be edited and form complementary base pairs by binding to the target sequence as well as the Cas9 protein. The Cas9 then searches for a Protospacer Adjacent Motif (PAM) sequence near the target site, which allows the Cas9 to bind and cleave the DNA. Once the target site and PAM are located, the Cas9 induces a double-stranded break (DSB) in the DNA, for which the cell can repair the DNA break via a Non-Homologous End Joining (NHEJ), which leads to random insertions/deletions of DNA sequences, or Homology-Directed Repair (HDR), which enables precise edits if a repair template is provided.

Current methodologies often rely on iterative cycles of computational modeling, experimental validation, and in vitro assays; however, these processes are not only time-consuming and resource-intensive but also demand costs and potential for consequences. Moreover, the cost associated with protein synthesis, testing, and validation remains a substantial barrier, limiting the scalability of CRISPR-Cas applications, particularly in resource-limited settings. Off-target effects can also occur when Cas9 binds and breaks happen in unintended DNA sequences. Off-target effects may be caused by sgRNA mismatches, when similar sequences in the genome can confuse the Cas9 protein, leading to erroneous cuts or variable PAM binding efficiency. Off-target effects can cause gene disruption, oncogenesis, autoimmune disease, and fatal health risks. Cost-efficient fixes matter toward these complications for many reasons, such as clinical safety and scalability. In

agriculture or industrial biotech, cost-efficient CRISPR tools reduce production costs for engineered crops or microorganisms.

Engineering Cas9 proteins tailored for specific applications is an exceptionally expensive and tedious process. Cas9 proteins are large, intricate proteins with multiple functional domains, each contributing to DNA binding, sgRNA recognition, and DNA cleavage. Any modification to improve one aspect by a small margin, such as reducing off-target effects, increasing specificity, or customizing each protein to each specific application, requires iterative and expensive fine-tuning. Many labs rely on directed evolution, which involves introducing mutations into Cas9, screening thousands of variants for desired traits, and validating the best-performing mutants through repeated testing cycles. The cost is further exacerbated by the need for high-throughput screening platforms, sophisticated bioinformatic pipelines, and specialized laboratory equipment for both computational and experimental stages. These challenges collectively create significant financial and temporal barriers, limiting access to smaller labs and organizations with fewer resources.

METHODOLOGY

CASNOVO leverages ProtGPT2[(2)], a transformer large language model trained on around 50 million naturally occurring amino acid sequences. The base algorithm was then trained on a dataset of over 100,000 Cas9 proteins downloaded and saved in a FASTA file format and curated from the UniProt Database[(1)]. The fine-tuned architecture learned protein sequence patterns and motifs to generate 300 novel proteins in silico. In training, the pre-existing protein sequences are tokenized with digits, and the function reads and processes the FASTA files imported in a process that can be named Environmental Configuration. In Data Preprocessing, the sequences are then padded with zeros to make them all the same length, through a process known as normalization, ensuring uniformity in the dataset. ProtGPT2 is leveraged and loaded by Hugging Face Transformers, and its algorithm was leveraged in training.

The necessary libraries and dependencies are important in the algorithm. In a second form of Environmental Configuration, the model optimizes compatibility, specifically for tokenization and memory management. It is then changed from a figurative evaluation mode to an inference mode, where the model uses learned features rather than continually learning and adjusting. To generate novel Cas9 protein sequences, the model begins with an empty prompt. Based on the evaluation, it creates new sequences on tokens, which are then decoded into amino acid characters.

Generated sequences were pushed through a deep-learning validation pipeline employing AlphaFold[3] and ESMFold[4], bioinformatics tools capable of contextualizing protein stability and quality of proteins, determining protein structure, stability, and perplexity in transformer-based large language model tools. Validated proteins were assessed by the predicted local difference test (pLDDT), post-translational modification score (PTM), sequence coverage, tertiary structural visualizations, and the model-given accuracy metrics.

RESULTS & DISCUSSION

Generated protein sequences that were pushed through the ESMFold and AlphaFold algorithms were assessed by the predicted local distance difference test (pLDDT), post-translational modification score (PTM), sequence coverage, and tertiary structural visualizations. pLDDT scores provide a measure of confidence in the predicted 3D structure of a protein according to the organization of the amino acids (protein positions). The functionality of a Cas9 protein highly depends on its 3D conformation; therefore, having a high confidence in predicting structure is crucial for assessing their viability. Additionally, when comparing different synthetic or naturally occurring Cas9 variants, pLDDT scores provide insight that helps evaluate which variant is more likely to adopt a functional conformation. As pLDDT scores are a robust metric for evaluating the structural integrity and potential functionality of synthetic Cas9 proteins, CASNOVO relies on this

numerical metric to measure the viability of the generated proteins. The x-axis represents the positions of the proteins, and the y-axis represents the pLDDT scores 0-100.

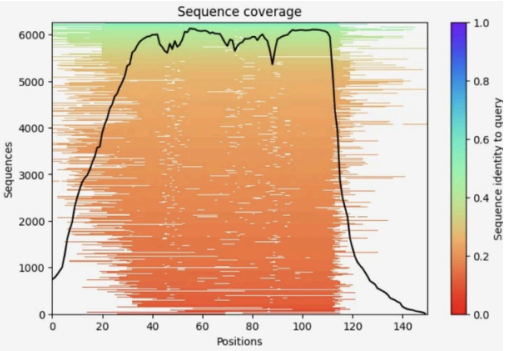


Figure E

Sequence coverage measures the extent to which each position in a reference sequence is represented across multiple aligned sequences; it demonstrates how thoroughly a particular region of a protein sequence is covered by the dataset. The color gradients indicate sequence identity in reference to the protein query, where red represents low identity, and blue/green represents high identity. Figure E contains novel CASNOVO-generated sequences, and Figure F contains scores of existing Cas9 proteins. These figures show similar graphs and this is promising because they hold similar protein positions and properties as those in the dataset. Both graphs show a sharp drop-off in coverage after their main alignment regions (~120E, ~110F). This similarity suggests that the generated Cas9 sequences maintain structural and functional integrity across key conserved domains.

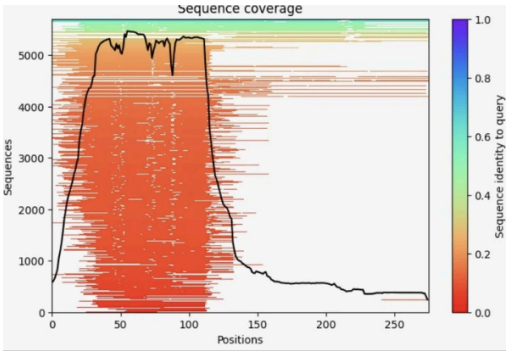
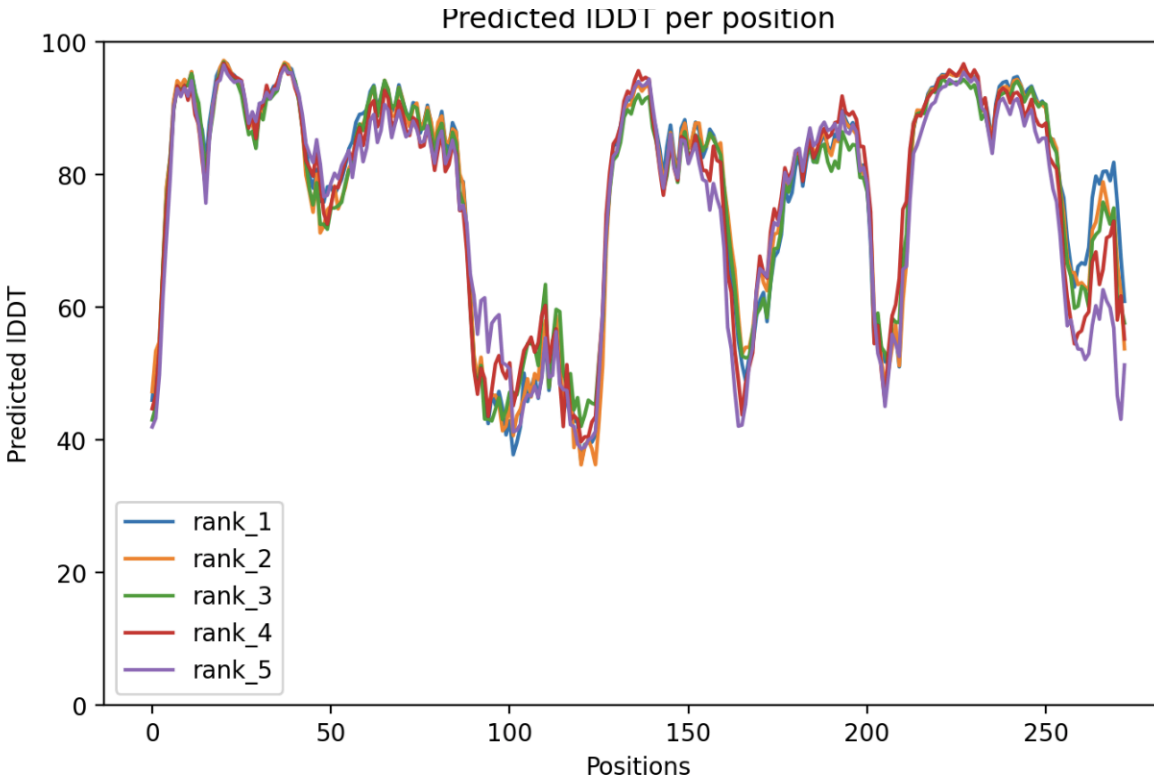


Figure F

The Cas9 proteins from the UniProt dataset contained pLDDT scores between 60 and 80. CASNOVO's fine-tuned algorithm has presented an unprecedented mean pLDDT score of 76.6, with further validation of the generated proteins pending, a close to two-fold increase from the base generation of ProtGPT2. The pLDDT score is a confidence metric used in protein structure prediction models to measure the reliability of predicted atomic positions in protein structures. As the mean pLDDT score of all the generated proteins falls between 75-80, the predicted models for Cas9 are structurally robust, and most residues in the proteins have well-defined positions. There are noticeable dips in the pLDDT scores and these likely respond to flexible loops, disordered

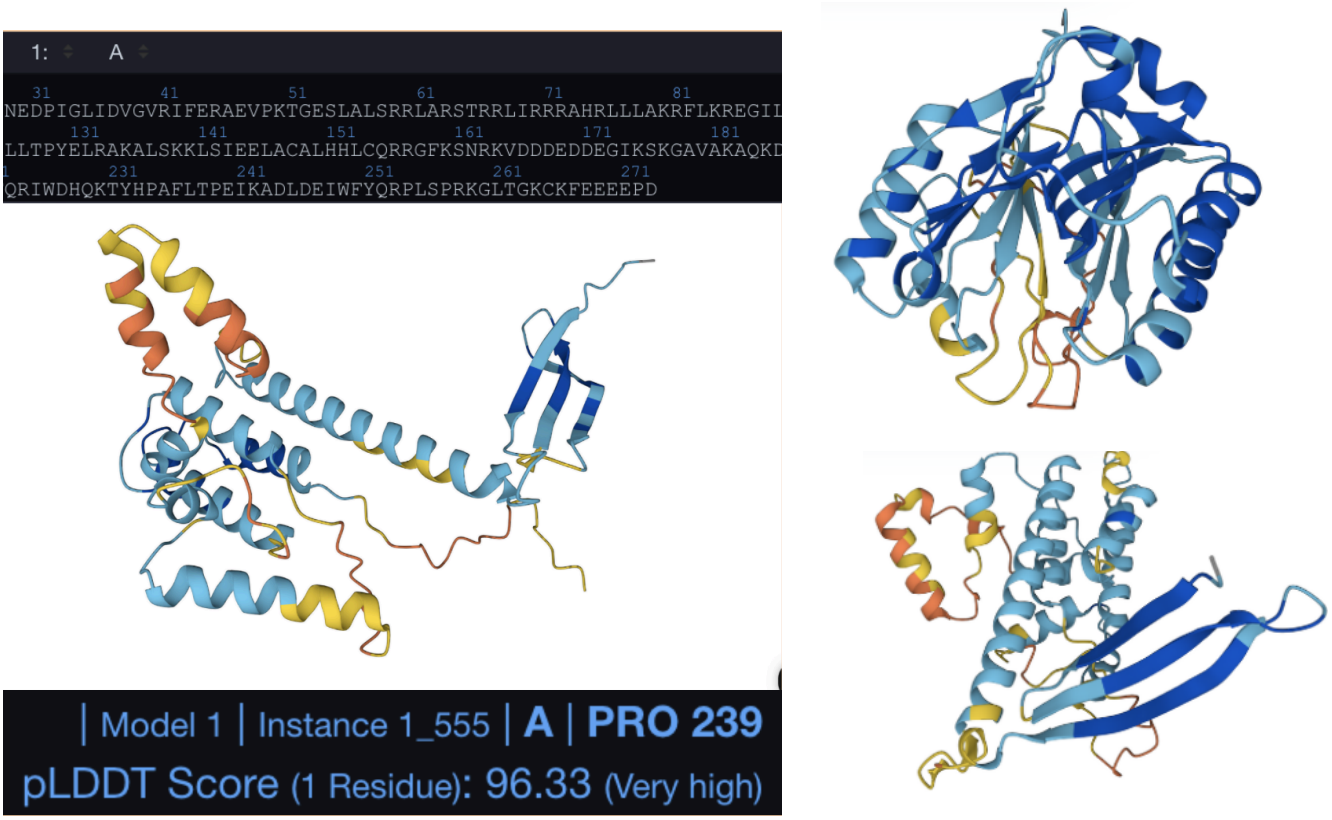


regions, or domains where structural prediction is inherently challenging due to high variability, lack of structural templates, or accuracy of validation algorithms. Across all five models (Rank_1 to Rank_5), the overall trend remains consistent, with minimal variation between ranks. This suggests prediction is robust and produces results across multiple iterations. The consistency and structural

integrity these metrics provide insights that are crucial for applications requiring precise structural information, such as national protein design, gene-editing optimizations, or drug targeting.

These results indicate the promise of the Cas9 generation pipeline. The model-given accuracy metric measures how similar the traits of the pre-existing proteins from the inputted dataset are to the novel-generated proteins. An accuracy metric greater than or equal to 80% is considered excellent; CASNOVO's accuracy metric resulted in ~84.03%. After rigorous training and validation, it can be concluded that a transformer-based deep learning model has the potential to generate viable Cas9 proteins. Data and numerical metrics show that there is great potential in generating functional Cas9 proteins. However, there are great limitations in the study thus far. Validation efforts have been limited to in-silico, whereas eventual validation in-vitro would tell

more about the Cas9 protein functionality in a biological setting. While metrics such as pLDDT scores provide insight into the general integrity of the protein, in-vitro validation will put CASNOVO to the ultimate test, confirming its true functionality and its potential application in genetic therapies. While pLDDT scores are an effective measure of structural confidence, it does not necessarily account for all aspects of a protein's functionality. For example, they do not directly measure the ability of Cas9 proteins to bind or cleave to sgRNA. This limitation suggests that additional validation metrics and methods are necessary to fully assess the potential of the generated proteins.



AlphaFold and ESMFold have the capability of visualizing the tertiary structure of novel proteins given their amino acid sequences. By inputting them into the combined algorithm, we can visualize the confidence metric of the predicted residue. The sequence shown is the primary structure of the protein, consisting of a linear chain of amino acids represented by their single-letter codes. These tools provide not only a 3D spatial arrangement of the protein but also the pLDDT confidence score for each predicted residue, displayed as a color gradient. Blue (pLDDT 70-90) and

dark blue (pLDDT >90) represent high confidence and warmer tones, such as yellow or orange (50-70) represent lower confidence of predicting the residues visually. Large sections of the proteins are in dark blue and blue, indicating high confidence in structural predictions, representing stable secondary structures. Certain regions, particularly loops and terminal ends, are shown in yellow/orange/red, indicating lower prediction confidence. These regions may play roles in protein dynamics, ligand binding, or regulation, and are inherently more flexible. Each position corresponds to a residue that is later folded into a 3D structure in the visual models. The specific residue PRO 239 is highlighted with an exceptionally high pLDDT score of 96.33, indicating high confidence in its structural prediction.

These results demonstrate that the novel Cas9 proteins generated by leveraging Machine Learning and Natural Language Processing (NLP) exhibit strong sequence coverage, structural integrity, and high confidence in their predicted tertiary structures. Sequence alignment shows consistency with natural Cas9 proteins, while pLDDT confidence scores reveal well-folded and reliable core regions, with flexible regions corresponding to expected dynamic domains. These success metrics highlight the transformative potential of computational protein engineering in generating optimized Cas9 variants tailored for specific applications. These advancements pave the way for more efficient, cost-effective, and scalable gene-editing tools, addressing critical challenges in synthetic biology, agriculture, and medicine. CASNOVO stands as a testament to the incredible power of innovation, where computational technology and biology converge to rewrite the rules of what is possible. By generating novel Cas9 proteins with precision and purpose, we are not only transforming the landscape of gene editing but also unlocking new avenues to cure diseases, feed the world, and address the most pressing challenges of our world.

RESOURCES

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., & Linial, M. (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, *38*(8), 2102–2110. https://doi.org/10.1093/bioinformatics/btac020

Chen, T., Dumas, M., Watson, R., Vincoff, S., Peng, C., Zhao, L., Hong, L., Pertsemlidis, S., Mayumi Shaepers-Cheu, Wang, T. Z., Divya Srijay, Monticello, C., Pranay Vure, Rishab Pulugurta, Kseniia Kholina, Goel, S., DeLisa, M. P., Truant, R., Aguilar, H. C., & Chatterjee, P. (2024). PepMLM: Target Sequence-Conditioned Generation of Therapeutic Peptide Binders via Span Masked Language Modeling. *ArXiv*, arXiv:2310.03842v3. https://pmc.ncbi.nlm.nih.gov/articles/PMC10593082/

Ferruz, N., Schmidt, S., & Höcker, B. (2022). ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-32007-7

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., & Back, T. (2021). Highly Accurate Protein Structure Prediction with Alphafold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574

Lv, L., Lin, Z., Li, H., Liu, Y., Cui, J., Chen, C. Y.-C., Yuan, L., & Tian, Y. (2024, February 26). *ProLLaMA: A Protein Large Language Model for Multi-Task Protein Language Processing*. ArXiv.org. https://doi.org/10.48550/arXiv.2402.16445

UniProt: the universal protein knowledgebase. (2016). *Nucleic Acids Research*, *45*(D1), D158–D169. https://doi.org/10.1093/nar/gkw1099