Machine Learning in Biochemical Research: A comprehensive overview of machine learning use in protein structure prediction and design

Author: Samuel Herman

Abstract

Machine learning (ML) is a computational method that utilizes training off of datasets to enable algorithms to complete desired tasks. Recently, numerous advances in ML have been seen in Biochemical research for determining and modifying protein structure using programs such as AlphaFold. However, utilization of these techniques to effectively complete protein research requires both knowledge of the existing algorithms and choosing the correct methodology to achieve a desired effect. Here we show the unique qualities of differing ML-based methods for both determining the structure of sequenced proteins and creating proteins with desired characteristics. AlphaFold was demonstrated to be useful for finding the structures of published protein sequences, while RoseTTA-fold's ability to be utilized on a personal computer makes it useful for researchers determining the structure of their own modified or novel proteins. Three methods of ML assisted protein design - modifying existing functions, adding new functions to existing enzymes, and *de novo* design of proteins - were also analyzed. ML can be used to significantly decrease the time required for the first two techniques while enabling the third to be completed in the first place. Modifying and adding functions to existing enzymes was most useful when a preexisting protein of similar or related function exists, while *de novo* design was best used when no related protein could be identified. These findings can assist researchers when choosing a desired ML method for completing protein based research and highlight areas of future improvements for ML use in biochemistry.

Background

Machine learning (ML) is a method of computation that has been in use for years, though it has only recently gained more attention due to sweeping advancements made in its use by both the public and in research. ML in its simplest form is a computational technique that utilizes the recognition of patterns in large data samples in order to better perform a task without explicitly being programmed to do so (Sarker, 2021). ML has found applications in a variety of fields, with biochemical research being no exception. Of particular emphasis in recent years has been the use of ML models to determine the structures of proteins and also assist in the design of new proteins that are capable of performing desired functions. The 2024 Nobel Prize in Chemistry was awarded for the development of two different machine learning models, one being AlphaFold which has so far predicted the structure of millions of proteins, and the other a machine learning model capable of designing new protein structures to achieve specific tasks (Abriata, 2024).

With the recent increase in ML use in biochemistry, researchers are left with the difficult task of discerning how to actually choose and utilize these systems to accomplish research tasks. This can prove difficult with many systems existing that perform the same function utilizing different techniques, such as Alpha-Fold and RoseTTAFold which are both ML based algorithms that can predict protein structure. Additionally, even if one does find an appropriate model, they are then left with the problem of finding how these *in silico* results can be translated into *in vitro* or *in vivo* results.

The purpose of this paper is to review numerous ML models that are currently used in biochemical research to both predict protein structure and design proteins. We intend to

identify the differences in how the algorithms function, identify their applications within research, and also identify strengths and weaknesses in different models. Additionally, we will identify current examples of these techniques and discuss how ML aided in the completion of these projects.

Protein Structure Prediction

A wide variety of different machine learning algorithms have been developed in an attempt to predict the structure of the most fundamental organic macromolecule - proteins. Determining the structure of proteins was previously a highly labor intensive process requiring protein purification, crystallization, X-ray crystallography studies, and most importantly - time. (McPherson & Gavira, 2013). In some instances, determining protein through wet-lab based techniques is simply not possible, especially in intrinsically disordered proteins or flexible regions of protein structure (Bai et al., 2023). *In silico* methods were seen as a promising way to reduce the significant amount of time required to determine protein structure. Conventional computational methods have been used to attempt to refine structures for decades, typically by utilizing physics based predictions for interactions between amino acid side-chains, though these structures often proved inaccurate in comparison to experimentally determined structures (Zhang, 2020) . However, recent utilization of machine learning to predict protein structure based solely on known protein sequence has proved incredibly useful. Of particular recent interest is AlphaFold and RoseTTAFold, two AI-Based-Prediction models that have been used to great effect to predict protein structure.

Both of these programs work through relatively similar methods, training off of the large amount of 3D protein structure data found within the Protein Data Bank. Both programs predict

the structures of known proteins, and then compare the predicted protein structure to experimentally determined structures to train their prediction models. However, the methods utilized by each program to actually model a protein based on a given sequence is significantly different. AlphaFold predicts structures by using two different computational tracks, one of which uses evolutionary relationships found between amino acid residues and another predicting the geometric relationships between residue pairs (Fig. 1). AlphaFold first scans the PDB until it finds proteins of similar sequence compared to the query sequence. These similar protein sequences are then compared between species to find specific amino acid residues that have been preserved across evolution, and then pairs these preserved amino acids based on the likelihood that they interact, such as pairing a positively and negatively charged residue. Based on these evolutionary relationships, the geometric algorithm determines the distances between each residue and maps the predicted distances between residues. This process repeats until a distance map is created that best aligns with both the evolutionary data and geometric data. This allows the program to generate a 3D atomic structure based purely on prediction, and that structure is then fed through the evolutionary and geometric algorithms multiple times to generate the final predicted 3D structure of the protein. Note that the AlphaFold model relies solely on its deep-learning in order to actually produce 3D structure rather than relying on physical constraints. This often results in early predictions having nonsensical structures that notably do not preserve the integrity of the protein backbone, with domains seemingly leaping across the protein before the final prediction forms a physically feasible structure (Jumper et al., 2021). Based purely on these ML predictions, a finalized protein structure is generated.

RoseTTAFold, on the other hand, utilizes a three-track neural network that processes 1-dimensional, 2-dimensional, and 3-dimensional data to produce a final series of protein backbone coordinates. RoseTTAFold utilizes machine learning to predict the distance between the individual amino acid residues within the protein (the 2-dimensional data) based on the amino query amino acid sequence (the 1-dimensional data). This 2-dimensional data is then utilized to generate 3-dimensional coordinates after sufficient comparison between all three algorithms to create the agreed upon final protein structure. Further processing by traditional physics based modeling is required to place the side chains and refine structure. Unlike AlphaFold, use of evolutionary data is limited when generating the model, instead relying more heavily on geometric prediction to generate the final structure (Baeck et al., 2021).

These fundamental differences create unique strengths and weaknesses for these algorithms. AlphaFold's emphasis on evolutionary data found within the PDB in addition to its geometric based prediction allows it to produce highly accurate protein structures that often are a near perfect match to experimental data. AlphaFold, however, relies on access to supercomputing in order to predict structures in a timely manner. RoseTTAFold's emphasis on geometric prediction results in overall lower accuracy compared to AlphaFold, though this also allows the program to be utilized on individual computers containing a single GPU in order to generate a protein structure. Fortunately, the use of AlphaFold to predict the structure of previously sequenced proteins is essentially unnecessary. This is because AlphaFold has already been utilized to predict the protein structure of nearly every protein known in nature, with 200 million protein structures already being published to the publicly accessible PDB (AlphaFold and beyond, 2023). The evidence of AlphaFold's accuracy in prediction is high, and the ability for this

model to predict nearly every protein structure has proven incredibly useful for researchers. (Abramson et al., 2024). Accuracy of predictions is often computed with external validations such as calculating the RMSD of structures determined experimentally vs those predicted. This allows researchers to compare known residue position to predicted residue position to determine the accuracy of predicted structures. Additionally, confidence of results for individual residues is measured through the predicted local distance difference test (pLDDT) (Mariani et al., 2013). This test uses the stereochemical plausibility of atoms within protein structure to determine regions where the experimental results likely match predicted results. Note that while AlphaFold is often superior in instances where large amounts of data already exists for a given protein, RoseTTAFold often performs better at prediction when less data is available due to its emphasis on geometric data over existing evolutionary data (Wang et al., 2024).

It is important to note that there are instances where prediction via ML simply is not capable of predicting protein structure regardless of the confidence value corresponding to the prediction. This is especially common with intrinsically disordered regions of proteins which are often represented as having defined folded structures (Li et al., 2025). Another issue is predicting interaction with other macromolecules such as DNA or ligands. However, newer advances such as those made in AlphaFold 3 have enabled more accurate prediction of structure based on molecular interactions outside of the protein (Abramson et al., 2024).

It is the author's opinion that both of these algorithms' combined strengths and weaknesses make them useful in different scenarios. AlphaFold has already proved its use in determining highly accurate structures for previously sequenced proteins, though it relies on resource heavy computational methods that can only be achieved through the use of

supercomputing. RoseTTAFold is more accessible to be run on lower end hardware possessed by individual labs, however it is not as accurate in prediction and therefore is better suited for use in finding the structure of proteins that have not already been published to the PDB, such as novel or mutated proteins. If needed, further analysis of these structures can later be performed using more resource intensive methods such as x-ray crystallography or even prediction by AlphaFold if deemed necessary.

Protein Design and Modification

In order to actually achieve machine learning, one must first have a database to actually train models on and set goals as to how to create a protein for a certain task. Regarding a database, the hard work has already been done by researchers. Decades of sequence data, structural data, and experimental results from previous studies can be utilized to train ML models, and many algorithms are capable of utilizing their own generated data in order to further refine generated structures and sequences. Regarding how to create a desired molecule, three methods of protein design are typically used: a *de novo* design where a new protein is produced from scratch, modification of existing proteins to function better or worse than they normally do, or modification of existing proteins to achieve a new task entirely (Notin et al., 2024). All of these methods have seen progress in machine learning, with advantages and disadvantages for each.

<u>Modification of Existing Function</u>

The most immediate and obvious modification of a protein is making it better at doing the job it is already designed to perform. This is a technique that could prove especially useful for enzyme modification, as enzyme kinetics can already be readily measured using standard

laboratory techniques, and increasing enzyme efficiency could have applications in industry or areas where a known protein is already used readily. One study not only utilized a ML model to modify two enzymes to increase their activity , but also managed to limit human interaction in this process through the use of automated laboratory equipment interfacing with their model (Singh et al., 2025). The chosen enzymes were *A*tHMT and *Ym*Phytase which are responsible for biocatalytic alkylation and phosphate hydrolyzation at low pH respectively. While the goal was to improve activity for both of these enzymes, the researchers had additional modification goals for each individual enzyme. For *At*HMT, researchers wanted to improve the ethyltransferase activity of the enzyme to be selective for ethyl iodide over methyl iodide, while the goal for *Ym*Phytase was to broaden its pH range for activity. These enzymes were chosen due to a lack of previous study using ML tools, potential industrial applications, and their suitability for automation-based research.

Researchers first needed a way to communicate with their ML algorithm, which they decided to achieve through the modification of an existing large language model. A user could input a text based query, such as “help improve the activity of enzyme X”, and the language model would respond with executing tasks that were preprogrammed by researchers to assist users with limited coding skills. The automated equipment that was interfaced was the Illinois Biological Foundry for Advanced Biomanufacturing (iBioFAB), which consists of a robotic arm capable of interacting with standard biochemical equipment and performing experimentation such as plasmid making and incubation of bacterial samples (Fig. 2). Note that while the iBioFAB completes most of its work autonomously, it still requires human interfacing and supervision when designing an experiment.

The iBioFAB created the dataset needed for ML-based analysis by randomly inducing mutation in both enzymes and performing an assay of enzyme activity, thus creating a "protein library" that demonstrated which mutations increased and decreased activity. ML was then used to analyze these mutations, and mutagenesis was repeated under guidance of the ML algorithm. This process was repeated three more times with minimal human intervention over the course of four weeks. It was found that activities of the ML engineered enzymes were 16-fold and 26-fold higher for *At*HMT and *Ym*Phytase respectively. Additionally, *At*HMT was found to have a 90-fold preference for ethyl iodide over methyl iodide, and *Ym*Phytase performed significantly better at approximately neutral pH values. The applications of this technology are promising, with further development possibly allowing for large-scale automation of protein modification in times significantly faster than can be achieved by human researchers.

New Function in Existing Proteins

In cases where no protein exists to perform a desired function, one must design a protein. One method of doing so is the modification of an existing protein to perform a new task. This technique is referred to as protein evolution, and traditionally relies on random mutation of a large quantity of proteins to achieve a desired effect (Vidal et al., 2023). However, with the use of predictive ML based models, researchers have been able to instead "direct" protein evolution rather than relying entirely on random chance to achieve the desired function. (Wu et al., 2019). This is useful because a combination of mutations is often needed to achieve successful evolution, and without a way to actually predict which mutations are affected, hundreds of thousands of mutations would need to be tested. By using *in silico* methods to

simulate mutations, one can significantly increase the efficiency in identifying desirable mutations and decrease time significantly. The ML algorithm was trained on the screening data from a library of protein variants containing random mutations that were analyzed for desired properties. ML then directs simulation to determine the fitness of every possible sequence to identify those mutations with the highest fitness for the desired effect. These mutations can then be tested *in vitro* to determine successful mutations. While this process still takes time, it is significantly faster than screening all possible mutations randomly.

This technique was tested with the enzyme *Rma* NOD to catalyze a carbon-silicon bond forming reaction that is not naturally occurring, specifically between phenyldimethyl silane and ethyl 2-diazopropanoate. This enzyme was chosen due to its stability in high temperatures which should support destabilizing mutations and because it is not enantioselective. The goal was to make the enzyme selective for the formation of either the (R) and (S) product through separate mutations which was successfully achieved for both stereoisomers.

Similarly, researchers were able to modify green fluorescent protein (GFP) to instead emit a longer wavelength corresponding to yellow fluorescence through ML guided directed mutation (Saito et al., 2018). The utilization of ML allowed researchers to generate a significantly larger library of candidate mutants to achieve the desired task compared to traditional means, resulting in a final protein library of ~80 candidate yellow fluorescent proteins. Overall, by directing mutation using ML to generate possible protein structures, one can achieve a desired result in significantly less time while also accounting for synergistic mutations that are often not considered through more traditional protein modification methods (Xu et al., 2025).

De Novo Design

One can also design a protein from scratch in order to achieve a desired effect, a technique known as *de novo* design. As proteins are often extremely complex and one cannot predict the structure or its interactions with molecules manually, ML has proved vital in *de novo* design. Of particular importance is the protein language model (PLM), ML algorithms that are essentially trained to read the sequence of a protein and understand the relationship between the sequence and the function of a protein without the need to generate its structure. These PLMs can then be used to build a novel protein one sequence at a time with this training (Zhang et al., 2024). A common analogy is comparing a PLM to naturally spoken languages (Fig. 3). PLMs treat amino acid residues as letters, primary structure and as words, secondary and tertiary structures as sentences, and the final quaternary structure as an entire paragraph carrying meaning. Therefore, a PLM should be capable of creating a protein with new structure by simply predicting the next residue in the primary structure of a protein.

An application of protein language models for designing novel proteins has been demonstrated through the use of machine learning in designing antibiotics. Researchers utilize ApexAmphion, an ML program that utilizes a protein language model to generate hundreds of suitable sequences for antimicrobial peptides (Cao et al., 2025). Researchers first modified ProGen2-xlarge, a protein language model containing 6.4 billion parameters, to train their model in identifying antibiotic motifs in peptides. AphexAmphion was then trained using both antimicrobial peptide sequences known to be effective against at least one pathogen and ineffective sequences to optimize its ability to generate antibiotic sequences. Finally, researchers rewarded the model for generating peptides with desired characteristics such as

hydrophobicity, size, and molecular weight. Researchers chose 100 generated peptides and tested their effectiveness against 16 different bacterial strains, including Gram-negative, Gram-positive, antibiotic resistant, and antibiotic susceptible bacterial strains. Of these 100 peptides, all were predicted to be effective in low concentration, and 99 of the 100 compounds were shown to successfully kill at least two different bacterial strains through *in vitro* testing. Additionally, analysis of each peptide showed diversity among the generated sequences, meaning each antimicrobial sequence was not just a small alteration to a known antibacterial agent or another generated sequence. This presents significant potential in solving issues such as antibiotic resistance, and such techniques can be used to design novel antibodies, ligands, and other important proteins to achieve desired effects *in vitro* based solely on *in silica* studies.

Regarding the actual program required to achieve *de novo* design, a popular choice is RFdiffusion which utilizes a modified RoseTTAFold program to generate novel protein designs (Watson et al., 2023). The program is also capable of overcoming previous issues with *de novo* based programs such as RF Hallucination and Rosetta, now including the ability to design proteins with desired symmetry, increased protein complexity, and decreased the time for protein design. Note that a current limitation is the size of the proteins that can be generated which is capped at 600 amino acids.

Review of 3 Methods

Each of these methods have fundamental differences in experimental design that carry unique benefits and also downsides based on factors including time, budget, and ease of use. Modifying existing proteins has potential when a protein that performs the desired function already exists, significantly reducing workload or complexity of design for an experiment when

designing a protein, which decreases even further when considering ML integration. Note that the use of iBioFab has a high cost of entry which could be negated by performing experimentation manually. However this will increase the time requirement for the experiment significantly even under the guidance of an ML algorithm for analyzing data and directing mutations.

Adding new functions to existing proteins through protein evolution has been used for years in order to design proteins to achieve a desired function, however the process previously took a significantly longer period due to its reliance on trial and error for single mutations. Through ML guided protein evolution, this process becomes significantly easier with little modification to the process and is simpler to achieve due to only relying on mutagenesis rather than constructing a protein from the ground up. However, one must first identify a protein that has high potential for modification in order to achieve a desired result which can potentially limit the new functions a protein can create.

Despite the inherent complexity of *de novo* techniques, they seem most likely to achieve the greatest amount of innovation when combined with ML in biochemical studies. The ability to generate a protein from scratch that can complete any desired task is incredibly useful for situations where no proteins with potential for modification can be easily identified, or one wants to create a wide variety of proteins that achieve the same function. This is a technique that has only recently become more accessible due to the significant improvements in both predicting protein structure and generating protein language models through ML, and is likely only going to become more accessible as research into improving these models increases.

Conclusion

Machine learning is an extremely powerful computational technique for biochemical research. It has made previously difficult and time consuming tasks, specifically determining protein structure and creating new proteins, significantly easier and faster to achieve. As such, applications of ML include predicting protein structure, designing novel proteins, modifying existing proteins, assisting in data interpretation, and helping with experimental design. Though models can differ significantly in how they perform their tasks, the ability for ML to pull meaning and find patterns in difficult to interpret data has made it a vital tool for the modern biochemist. Utilizing ML to complete older biochemistry techniques such as protein evolution has made these processes significantly easier and less time consuming. Integration of ML models with large learning models can make it possible for a researcher even with limited coding experience to utilize ML in their research (Jin et al., 2025).

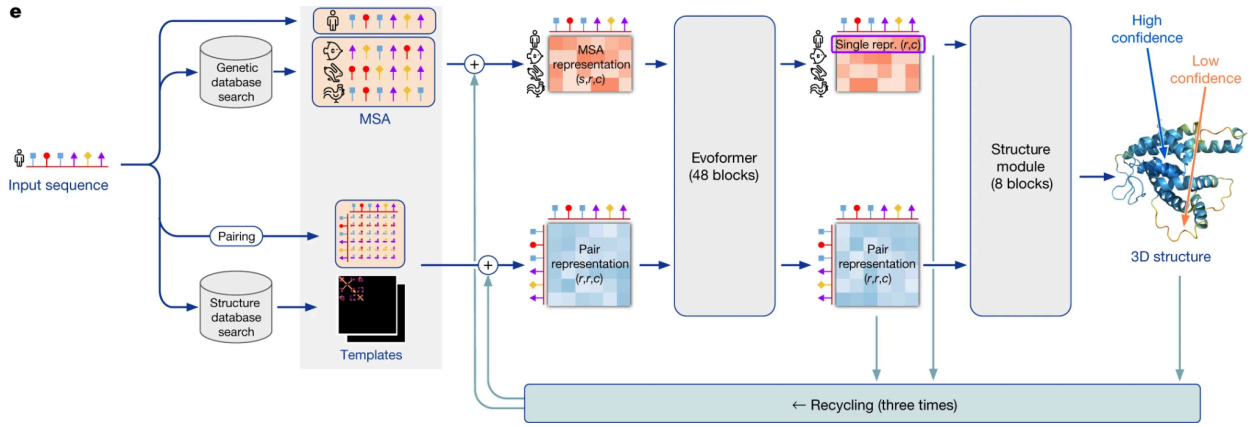
There are immediate concerns with overreliance on ML use in research that is often echoed in public centers with the utilization of available ML based algorithms, such as using large language models, replacing human generated experimentation and data. Concerns related to information transparency are often raised due to the massive amount of information required for many of these algorithms to function and the complexity involved in the generation of protein structures. These concerns can be combated through continued publication of how these algorithms function and through peer review ensuring that necessary data is included within research. Additional concerns about not checking work, replacing human researchers with automated research machines such as iBioFab, and overreliance on ML generated data such as structures made via AlphaFold are commonly cited as reasons to use caution when

approaching ML use (Terwilliger et al., 2023). However, it is important to note that replacing human researchers is still an expensive endeavor, and researcher oversight and interfacing is still required for the techniques mentioned in this paper. Proteins that will be utilized in highly regulated industries such as pharmaceuticals will still likely receive thorough review before approval.

Overall, ML use in biochemistry has high potential to make significant leaps in research, with developments such as AlphaFold already proving ML's potential by finding the structures of almost every known naturally occurring protein in a fraction of the time it has taken researchers to find the structures of only a small fraction of proteins. Future improvements in ML algorithm development will only further streamline biochemical research, such as further simplifying researcher interaction with ML to reduce the barrier of entry into utilizing ML for researchers unfamiliar with coding, or creating more accurate algorithms that can be run on local hardware without requiring supercomputer access to make predictions or interpret data. Additionally, by developing cheaper automated laboratory equipment powered by ML, labs can significantly reduce the amount of time spent performing labor intensive and repetitive tasks such as protein assays and gene amplification, therefore increasing productivity significantly.

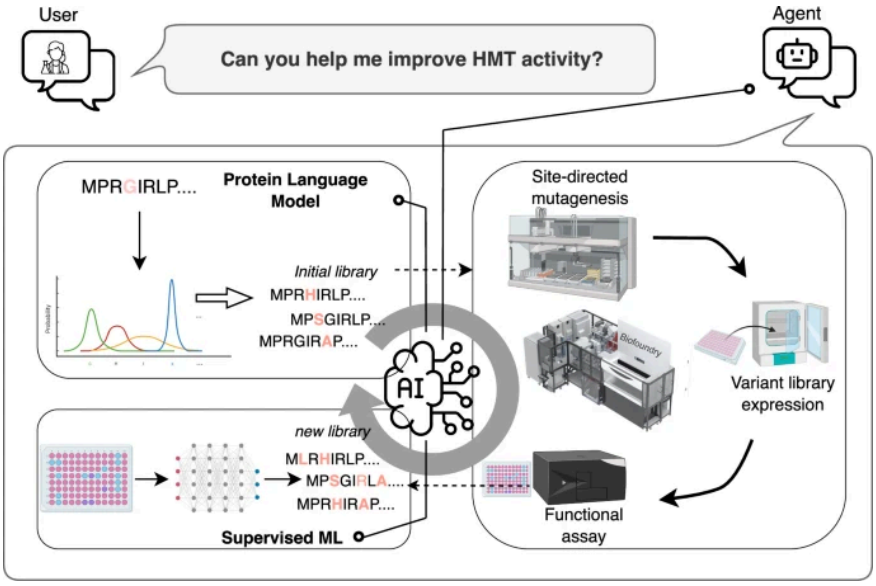
Figures

Fig. 1: *Protein modeling pathway for AlphaFold* (Jumper et al., 2021)



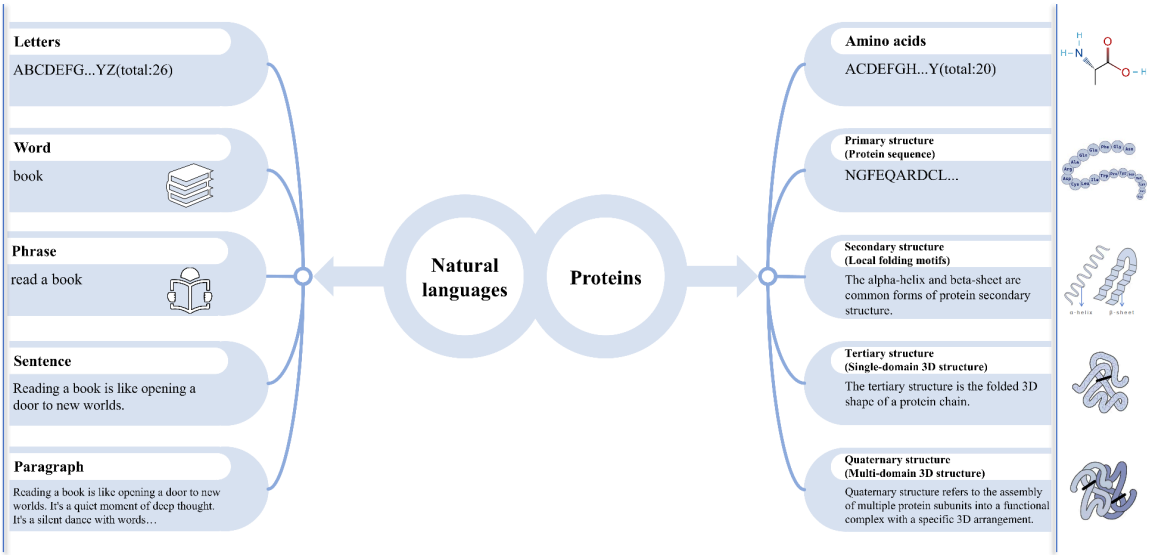
AlphaFold generates a protein structure by both comparing evolutionary data and geometric prediction of amino acid pairs using machine learning.

Fig. 2: *Overview of the generalized platform for autonomous protein engineering* (Singh et al., 2025)



Model for the use of ML in combination with iBioFab to produce desired changes in enzyme from a text based prompt by researchers.

Fig. 3: Comparison of protein language model to natural language (Wang et al., 2025)



A protein language model understands how to read the sequence of an amino acid much as one uses letters to understand the meaning of words.

Works Cited

Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630* (8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w.

Abriata, L. A. The Nobel Prize in Chemistry: past, present, and future of AI in biology. Communications Biology 2024, 7 (1), 1409. https://doi.org/10.1038/s42003-024-07113-5.

AlphaFold and beyond. *Nature Methods* **2023**, *20* (2), 163. https://doi.org/10.1038/s41592-023-01790-6.

Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; Van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. Science 2021, 373 (6557), 871–876. https://doi.org/10.1126/science.abj8754.

Bai, X.-C.; Gonen, T.; Gronenborn, A. M.; Perrakis, A.; Thorn, A.; Yang, J. Challenges and opportunities in macromolecular structure determination. *Nature Reviews Molecular Cell Biology* **2023**, *25* (1), 7–12. https://doi.org/10.1038/s41580-023-00659-y.

Cao, H.; Torres, M. D. T.; Zhang, J.; Gao, Z.; Wu, F.; Gu, C.; Leskovec, J.; Choi, Y.; Fuente-Nunez, C. de la; Chen, G.; Heng, P.-A. *A deep reinforcement learning platform for antibiotic discovery*. Arxiv. https://arxiv.org/abs/2509.18153 (accessed 2025-09-29).

Jin, S.; Wu, Q.; Fu, G.; Lu, D.; Wang, F.; Deng, L.; Nie, K. Breaking Evolution's ceiling: AI-Powered Protein Engineering. *Catalysts* **2025**, *15* (9), 842. https://doi.org/10.3390/catal15090842.

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.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; W, A., Senior; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Li, M.Q.C., Wang, S., Lin, SR. et al. Advantages and Limitations of AlphaFold in Structural Biology: Insights from Recent Studies. Protein J (2025). https://doi.org/10.1007/s10930-025-10310-8

Mariani, V.; Biasini, M.; Barbato, A.; Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **2013**, *29* (21), 2722–2728. https://doi.org/10.1093/bioinformatics/btt473.

McPherson, A.; Gavira, J. A. Introduction to protein crystallization. *Acta Crystallographica Section F Structural Biology Communications* **2013**, *70* (1), 2–20. https://doi.org/10.1107/s2053230x13033141.

Saito, Y.; Oikawa, M.; Nakazawa, H.; Niide, T.; Kameda, T.; Tsuda, K.; Umetsu, M. Machine-Learning-Guided mutagenesis for directed evolution of fluorescent proteins. *ACS Synthetic Biology* **2018**, *7* (9), 2014–2022. https://doi.org/10.1021/acssynbio.8b00155.

Sarker, I. H. Machine learning: algorithms, Real-World applications and research directions. *SN Computer Science* **2021**, *2* (3), 160. https://doi.org/10.1007/s42979-021-00592-x.

Singh, N.; Lane, S.; Yu, T.; Lu, J.; Ramos, A.; Cui, H.; Zhao, H. A generalized platform for artificial intelligence-powered autonomous enzyme engineering. *Nature Communications* **2025**, *16* (1). https://doi.org/10.1038/s41467-025-61209-y.

Terwilliger, T. C.; Liebschner, D.; Croll, T. I.; Williams, C. J.; McCoy, A. J.; Poon, B. K.; Afonine, P. V.; Oeffner, R. D.; Richardson, J. S.; Read, R. J.; Adams, P. D. AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. Nature Methods 2023, 21 (1), 110–116. https://doi.org/10.1038/s41592-023-02087-4.

Vidal, L. S.; Isalan, M.; Heap, J. T.; Ledesma-Amaro, R. A primer to directed evolution: current methodologies and future directions. RSC Chemical Biology 2023, 4 (4), 271–291. https://doi.org/10.1039/d2cb00231k.

Wang, L.; Li, X.; Zhang, H.; Wang, J.; Jiang, D.; Xue, Z.; Wang, Y. A comprehensive review of protein language models. arXiv (Cornell University) 2025. https://doi.org/10.48550/arxiv.2502.06881.

Wang, J.; Watson, J. L.; Lisanza, S. L. Protein design using Structure-Prediction networks: AlphaFold and RoseTTAFold as protein structure foundation models. Cold Spring Harbor Perspectives in Biology 2024, 16 (7), a041472. https://doi.org/10.1101/cshperspect.a041472.

Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Hanikel, N.; Pellock, S. J.; Courbet, A.; Sheffler, W.; Wang, J.; Venkatesh, P.; Sappington, I.; Torres, S. V.; Lauko, A.; De Bortoli, V.; Mathieu, E.; Ovchinnikov, S.; Barzilay, R.; Jaakkola, T. S.; DiMaio, F.; Baek, M.; Baker, D. De novo design of protein structure and function with RFdiffusion. Nature 2023, 620 (7976), 1089–1100. https://doi.org/10.1038/s41586-023-06415-8.

Wu, Z.; Kan, S. B. J.; Lewis, R. D.; Wittmann, B. J.; Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences* **2019**, *116* (18), 8852–8858. https://doi.org/10.1073/pnas.1901979116.

Xu, W.; Li, A.; Zhao, Y.; Peng, Y. Decoding the effects of mutation on protein interactions using machine learning. *Biophysics Reviews* **2025**, *6* (1), 011307. https://doi.org/10.1063/5.0249920.

Zhang, Y. Toward the solution of the protein-structure prediction problem. The FASEB Journal 2020, 34 (S1), 1. https://doi.org/10.1096/fasebj.2020.34.s1.00169.

Zhang, Z.; Wayment-Steele, H. K.; Brixi, G.; Wang, H.; Kern, D.; Ovchinnikov, S. Protein language models learn evolutionary statistics of interacting sequence motifs. Proceedings of the National Academy of Sciences 2024, 121 (45), e2406285121. https://doi.org/10.1073/pnas.2406285121.