ProteinCrow: A Language Model Agent That Can Design Proteins

Manvitha Ponnapati [1 2] **Sam Cox** [1] **Cade W Gordon** [1 3] **Michael J Hammerling** [1] **Siddharth Narayanan** [1] **Jon M Laurent** [1] **James D. Braza** [1] **Michaela M. Hinks** [1] **Michael D Skarlinski** [1] **Samuel G Rodriques** [1] **Andrew White** [1]

Abstract

Recent breakthroughs in deep learning have revolutionized protein structure and sequence modeling, enabling the design of proteins with novel functions through tools like AlphaFold, RFdiffusion, BindCraft and ProteinMPNN. Successful in silico protein engineering pipelines integrate multiple of these specialized models, incorporating biochemical insights, and iteratively optimizing sequences. Here, we present ProteinCrow, an agentic LLM-based protein design assistant that consolidates multimodal information, including structural data, deep learning models,scientific literature and biochemical data expressed in natural language, to automate protein design tasks by using 36 expert-curated tools. We evaluated its performance in designing binder libraries tailored to specific secondary-structure motifs; redesigning protein backbones to improve stability and optimizing binders to eliminate predicted MHC Class I epitopes.

1. Introduction

From engineering high-affinity therapeutic binders to optimizing enzyme activity and stability, computational protein design has become an essential tool for accelerating protein design with novel function. Traditional approaches have relied on physics-based models, such as Rosetta (Das & Baker, 2008), to evaluate protein structure. More recently, deep learning methods–such as AlphaFold2 (AF2) (Jumper et al., 2021), RFDiffusion (Watson et al., 2023), and ProteinMPNN (Dauparas et al., 2022) have transformed the field by enabling accurate structure prediction and sequence optimization. However, no single method fully captures the complex interplay of sequence, structure, and function. We hypothesize that Large Language Models' (LLM) ability to unify deep learning model outputs, biochemical data, literature, and physics-based methods as natural language (Jablonka et al., 2023; M. Bran et al., 2024; Boiko et al., 2023; Ramos et al., 2025) can serve as a framework to integrate multimodal information about proteins and faciliate effective protein design.

In this work, we present ProteinCrow, an LLM agent that integrates multiple tools to enable protein design through natural language prompts to generate protein sequence libraries with specified constraints. ProteinCrow also incorporates literature-based insights via PaperQA (Skarlinski et al., 2024), an agent optimized for retrieving and summarizing information from scientific literature, mirroring how experts consult existing information about a protein family or task-specific knowledge. It can perform subtasks currently done by humans, such as trimming large proteins into smaller regions for input to tools like BindCrat (Pacesa et al., 2024), identifying hotspot residues for binder design, and analyzing designed sequences using in silico metrics. By providing the agent access to deep learning methods used by human computational protein engineers, database querying tools, and physics-based methods (e.g., Rosetta), the agent can produce designs with optimized stability and generate libraries of protein sequences for experimental validation.

Recent efforts have incorporated LLMs into protein design pipelines, albeit with varying degrees of autonomy and scope. Protein language models (PLMs) have demonstrated the ability to capture structural and functional principles for design and property prediction(Ferruz et al., 2022; Ferruz & Höcker, 2022; Nijkamp et al., 2023), and LLMs have shown promise as optimizers for biological sequences (Chen et al., 2024; Wang et al., 2025). 310.ai (310) introduced a chat-based interface for natural language-driven protein design, though it does not function as an autonomous agent. In contrast, ProteinForceGPT (Ghafarollahi & Buehler, 2024) operates as an autonomous LLM agent to predict force–separation curves from pre-trained structures, leveraging methods like Chroma (Ingraham et al., 2023) and

[*]Equal contribution [1]FutureHouse, San Francisco, CA, USA [2]Massachusetts Institute of Technology, Cambridge, MA, USA. Work done while interning at FutureHouse. [3]University of California, Berkeley, CA, USA. Work done while interning at FutureHouse.. Correspondence to: Andrew White <andrew@futurehouse.org>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

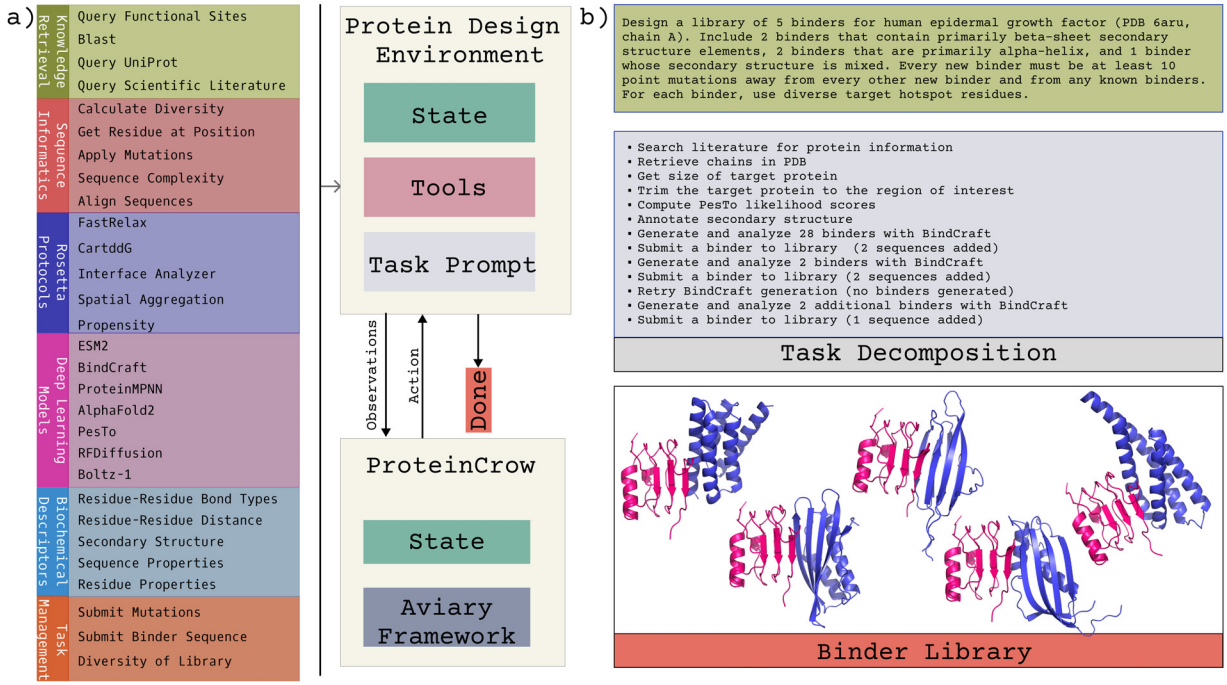


Figure 1. Schematic overview of ProteinCrow. (a) Diagram showing the tools available in ProteinCrow and the interaction between the protein-design environment and the agent within the Aviary framework. (b) Example task prompt for a binder-design workflow, instructing the agent to generate a library of five binders with specified structural constraints.

OmegaFold (Wu et al., 2022).

While most protein design agents are specialized or limited in scope, more general LLM Agents have been explored in other fields such as chemical synthesis (M. Bran et al., 2024; Boiko et al., 2023), materials research (Jablonka et al., 2023; Su et al., 2024), and experimental design (O'Donoghue et al., 2023; Huang et al., 2024). ProteinCrow is an LLM agent built using the Aviary Framework (Narayanan et al., 2024) to facilitate general purpose protein design. Aviary is a framework for building and training LLM agents on complex tasks. An LLM agent is a decision-making entity that can observe an environment, reason, and execute an action. In the case of ProteinCrow, an action is a tool call for protein structure/sequence analysis, generative design, or library design. The agent continues in a loop based on new observations or tool outputs until the desired goal is reached. We evaluated ProteinCrow on three tasks:

- Task 1: Goal-Directed Protein Design - Improving the stability of a target sequence with structural and functional constraints.
- Task 2: Automated Binder Design Pipeline - Generating libraries of binders with constraints on library composition, binder structure and binder sequence properties.
- Task 3: Optimizing Binders - Generates a library of sequences by redesigning a previously designed binder to eliminate predicted MHC Class I epitopes.

2. Methods

2.1. LLM Framework - Aviary

ProteinCrow operates within Aviary, an extensible gymnasium for language agents (Narayanan et al., 2024). Aviary frames the agent's sequential decision-making process as a language-grounded, partially observable Markov decision process (POMDP), enabling the agent to refine protein sequences iteratively. However, the tools built for ProteinCrow are transferable to other agent frameworks beyond Aviary. ProteinCrow optimizes protein stability, designs libraries of binders with specified constraints and optimizes a known binder by calling tools and receiving observations that inform the next set of actions. A schematic of ProteinCrow is shown in Figure 1. All experiments in ProteinCrow were conducted using Claude-3.5-Sonnet-20241022 at a sampling temperature of 0.1.

2.2. Tools

ProteinCrow has access to a variety of tools broadly grouped into the following categories:

- **Biochemical Descriptors** - Tools for biochemical analysis, including interface characterization, bond type analysis, sequence complexity, residue hydrophobicity, and secondary structure annotation.
- **Deep Learning Models** - Deep learning models for various protein design tasks, including binder design, inverse design, scaffold generation, and protein language models.
- **Rosetta Protocols** - Tools that execute Rosetta protocols, widely used for protein structure modeling and protein-protein interface analysis.
- **Task Management** - Tools for managing various prompt-related tasks, such as submission to a library and task completion tracking, along with tools to analyze metrics for sequences submitted to the library.
- **Sequence Informatics** - Tools for sequence-based analysis, including residue properties, multiple sequence alignment and identification of conserved sites.
- **Knowledge Retrieval** - Tools for querying and retrieving relevant information about the protein from databases and literature.

A list of all tools available to the agent beyond ones utilized for tasks mentioned in this work is provided in Appendix 1. Due to the frequent version conflicts and compatibility issues between various deep learning models, most tools within ProteinCrow operate on API endpoints through Modal (mod), which enables sandboxed execution of multiple models such as AlphaFold(Jumper et al., 2021), ProteinMPNN(Dauparas et al., 2022), ESM2 (Lin et al., 2023) and BindCraft (Pacesa et al., 2024).

2.3. Task 1 - Goal Directed Protein Design

2.3.1. Redesign Protein Backbones For Improved Stability with Structural Constraints

Goal-directed protein engineering involves mutating the target protein to achieve desired objectives, such as improved stability. (Jiang et al., 2024). While existing tools like ProteinMPNN have shown success in generating more stable protein backbones (Sumida et al., 2024), they typically require expert human intervention during generation and curation of experimental sequence libraries for more complex objectives, such as increasing the number of salt bridges (Kordes et al., 2022) while optimizing for improved stability. In this task, ProteinCrow with access to ProteinMPNN as a tool was prompted to re-design the backbone of a protein to increase its stability while also increasing the number of salt bridges. ProteinCrow was allowed to bias towards specific residues by adjusting the bias weights for amino acid types in ProteinMPNN. Salt bridges were identified after repacking the backbone with proposed mutations by counting all acidic (Asp,Glu)–basic (Arg, Lys, His) atom pairs whose inter-atomic distance is less than 4.0 Å.

To evaluate ProteinCrow's performance, we randomly selected 30 proteins (all with PDB IDs) from the ThermoMPNN training split of the Megascale stability dataset (Dieckhaus et al., 2024; Tsuboyama et al., 2023), which comprises of protein domains with 40–72 amino acids. Changes in free energy ($\Delta\Delta G$)) due to mutations were computed using the Rosetta cartddG protocol (Frenz et al., 2020), which is frequently used to evaluate protein folding stability in-silico (Sora et al., 2023). In addition, the following metrics were calculated for a comprehensive assessment of the proposed mutations:

- **Percent of Rosetta $\Delta\Delta G$) less than 0**: Indicates the proportion of mutations that resulted in a stabilizing change in the protein.
- **Amino acid usage**: Analyzes the frequency of different amino acids in the proposed mutations.
- **Sequence diversity**: Measures the diversity of the mutations proposed by ProteinCrow.
- **ESM2 pseudo log-likelihood**: A measure of log-likelihood of the proposed mutations computed using ESM2. (Lin et al., 2023).
- **AlphaFold Confidence**: plDDT scores from AlphaFold2 (Jumper et al., 2021).

2.3.2. Redesign Enzyme Backbones While Preserving Functional Residues

To test the ability of ProteinCrow to propose mutations that enhance protein stability without compromising function, we used two proteins studied in a prior work (Dauparas et al., 2022): Myoglobin, where oxygen-storage function should be preserved, and TEV protease, where catalytic activity must be maintained. We compared the mutation sites selected by ProteinCrow to those chosen by human designers in the original work, assessing whether ProteinCrow could avoid proposing mutations at the same functional sites. In addition to the metrics described in Section 2.3.1, we included the following metric to evaluate the proposed mutations.

- **Functional site preservation**: Number of times a mutation site in (Dauparas et al., 2022) was chosen by ProteinCrow.

3. Task 2 - Automated Binder Design Pipeline

BindCraft is a highly efficient deep learning model for de novo protein binder design, with reported experimental success rates ranging from 10 to 100% (Pacesa et al., 2024).

The pipeline includes several target-specific settings that may require optimization depending on the protein target, such as the number of iterations, design weights, and filtering criteria, which are typically chosen by a human expert. Our results demonstrate that ProteinCrow can automate and execute the BindCraft design pipeline without manual intervention for a diverse set of protein targets and library constraints by choosing appropriate inputs for BindCraft.

3.1. Designing Binders for Epidermal Growth Factor Receptor (EGFR)

To evaluate ProteinCrow on a challenging real-world target, we tested it on the EGFR protein using the structure from PDB entry 6ARU. This target was previously featured in a community-wide binder design competition hosted by AdaptyvBio (Cotet et al., 2025). We used the same evaluation metrics reported by AdaptyvBio to assess the quality of ProteinCrow's designs and to compare them to the top-performing entries that were experimentally verified.

ProteinCrow was evaluated on its ability to automate binder design process under various constraints, which included the following experiments:

- **Automated end-to-end binder generation**: ProteinCrow was prompted to generate a binder to EGFR with 10 replicates.
- **Targeted structural motif design**: The top-performing binder from the competition featured prominent β-sheet elements. To test if ProteinCrow could optimize design parameters in bindcraft to design a binder with a specific structural motif, we prompted ProteinCrow to add a binder with a β-sheet element to the sequence library.
- **Constraint-aware binder library design**: In this experiment, ProteinCrow was prompted to generate a library of 5 binders with three constraints on the binders added to the library (1) each binder must be at least 10 mutations away from known binders and (2) the resulting library should exhibit diversity in the composition of secondary structures (3) Target different hotspot residues on the protein.

3.2. Generalizing ProteinCrow to Distinct Protein Targets

To assess ProteinCrow's generalizability to automate the binder design workflow for other targets, we applied the agent to two additional structurally- and functionally-distinct protein targets: the SARS-CoV-2 receptor-binding domain (RBD) (PDB ID: 6M0J) and the beta barrel fold BBF-14 (PDB ID: 9HAG). For each target, we prompted ProteinCrow to design a library of 5 sequences, using 3 replicate trajectories per target with the same constraints as in the constraint-aware binder library design experiment.

3.3. Evaluating Binders

The performance of the sequences in each binder library was evaluated via the following widely used in-silico metrics:

- **AlphaFold Scores** - PLDDT, iPAE, and iPTM scores from the AlphaFold-Multimer model were used to evaluate the structural quality, alignment accuracy, and predicted transformation matrix of the designed binder sequences.
- **ESM Pseudo-Log-Likelihood Score** - ESM2 pseudo log likelihood score was used to assess the likelihood of the generated sequences using the protein language model ESM2 (Lin et al., 2023).
- **Secondary Structure Composition** - The secondary structure annotation of the binders was analyzed to ensure that they contain structural elements required by the design constraints in the prompt.

4. Task 3 - Optimizing a Binder to Reduce MHC Class I Epitopes

A critical step in therapeutic binder design is optimization of properties such as stability, aggregation propensity, and immunogenicity, while maintaining target affinity (Ebrahimi & Samanta, 2023). To assess ProteinCrow's ability to perform such optimization, we began with an EGFR-binding protein experimentally validated from the Adaptivbio Community Binder Design Competition and prompted ProteinCrow with generating libraries of 10 variant sequences across 5 replicates, to eliminate predicted MHC Class I epitopes. We employed netMHCpan 4.1 (Reynisson et al., 2020) to predict binding affinities of all 8–14 mer peptides to the HLA-A*02:01 allele. Finally, we ranked and compared the designed sequences by their ESM2 pseudo-log-likelihood scores and by the total numbers of predicted MHC Class I binders, including both strong and weak binders.

5. Results

ProteinCrow trajectories follow a nonlinear and adaptable path, showcasing its ability to integrate diverse sources of information and analyze sequence, structure, and literature context based on the task prompt. This is consistent across all tasks, highlighting ProteinCrow's flexibility in adapting to different protein design tasks and targets.

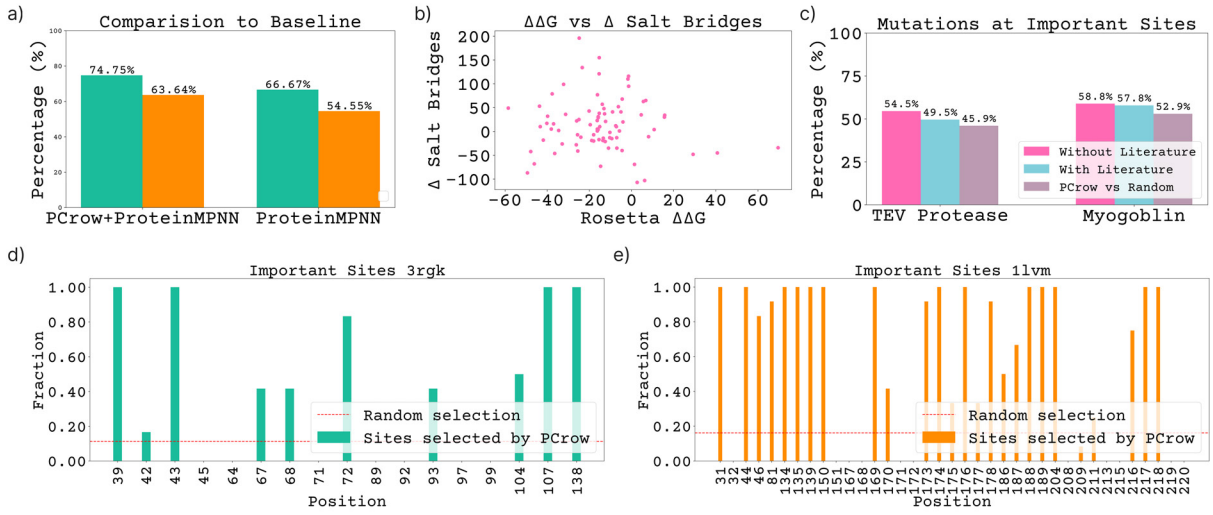


Figure 2. ProteinCrow performance on optimizing stability for increased salt bridges (a,b) and while maintaining function (c,d,e) (a) Bar plot showing the improvement of salt bridge formation and folding stability (measured using Rosetta $\Delta\Delta G$) of mutations by ProteinCrow compared to ProteinMPNN random baseline. (b) Scatter plot illustrating the relationship between Rosetta $\Delta\Delta G$ scores of ProteinCrow generated mutations and their corresponding salt bridge counts. (c) Comparison of percentage of important sites mutated with and without literature along with percentage of total functionally important positions significantly less likely to be mutated by ProteinCrow compared to a random baseline, when optimizing stability while preserving function. (d) Frequency of important sites mutated across 12 replicates of ProteinCrow in Myogoblin (PDB: 3RGK). (e) Frequency of important sites mutated across 12 replicates of ProteinCrow in Myogoblin (PDB: 1LVM).

5.1. Optimizing Protein Stability while improving number of salt bridges

To evaluate ProteinCrow's ability to perform goal directed protein design, we tasked it with proposing mutations that both increased folding stability (as measured by Rosetta $\Delta\Delta G < 0$) and introduce more salt bridges relative to the wild-type (WT) structure. ProteinCrow leveraged a combination of structure-based, sequence-based, literature-informed, deep learning, and Rosetta-based tools to analyze each target protein and generate design proposals.

As shown in Figure 2**a,b**, on 30 benchmark proteins, ProteinCrow successfully identified mutations that meet both criteria in many cases. In contrast, baseline ProteinMPNN with mutated residues sampled at randomly chosen residue positions of varying lengths in each replicate, while effective at generating low G variants, does not inherently increase number of salt bridges. These results demonstrate the potential of ProteinCrow to handle nuanced design goals that typically require expert-guided iteration, highlighting its potential as a general-purpose agent for goal-directed protein engineering.

5.2. Designing stable mutations while preserving function

To assess ProteinCrow's ability to identify functional residues as compared to residues chosen by human protein designers to be held constant during design to preserve function, we compared its selected mutation positions to those chosen by human protein designers in a previously published study. Compared to a random baseline for myoglobin and TEV Protease, ProteinCrow proposed significantly fewer mutations at $> 45\%$ of the important sites. An ablation done with and without literature showed a decrease in the number of functionally important sites chosen by ProteinCrow while redesigning the backbone (2c). These results suggest that, while the agent successfully captures some aspects of expert knowledge, it may overlook other critical residues, potentially due to limitations in how it interprets functional and structural information. However, ProteinCrow's modular framework offers a path for improvement through training or incorporation of more tools.

5.3. Designing Binders for Epidermal Growth Factor Receptor (EGFR)

Automated end-to-end binder generation: ProteinCrow was prompted to generated a single binder to EGFR across 10 replicates. ProteinCrow successfully generated a binder for EGFR for all 10 replicates. Selected binders from the

resulting binder library are shown in Figure 7, which also highlights some of the non-linear agent tool choice trajectories and the most frequent regions selected by ProteinCrow for trimming the protein. ProteinCrow automatically retrieved the protein structure from PDB entry 6ARU, performed structural analysis, and identified the key binding hotspots for designing effective binders, executed the BindCraft workflow and built a library of potential binder sequences.

Targeted structural motif design: In this experiment, ProteinCrow was tasked with designing binders that incorporated a β-sheet element, a structural motif present in some of the top-performing binders from the competition. ProteinCrow was able to design and add a binder with prominent β-sheet region to the library by adjusting the design weights in the BindCraft pipeline (binder and prompt shown in Figure 8). This result demonstrates ProteinCrow's ability to generate binders with specific structural features when required.

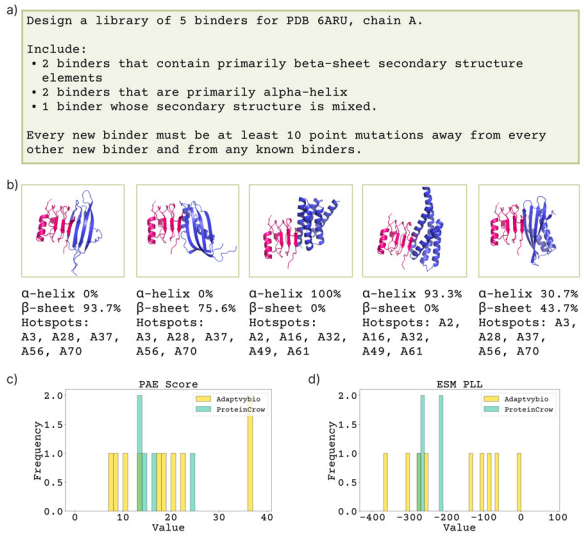


Figure 3. ProteinCrow can generate binder libraries with specified constraints. (a) Task prompt (b) Selected binders from designed binder library for EGFR. (c) AlphaFold iPAE score for ProteinCrow designed binders compared to those of the top 10 binders from AdaptyvBio's EGFR binder design competition (d) ESM2 PLL score for ProteinCrow designed binders compared to those of the top 10 binders from AdaptyvBio's EGFR binder design competition.

Constraint-aware binder library design: ProteinCrow successfully generated a binder library of 5 binders to EGFR while meeting the three specified constraints. First, each binder was at least 10 mutations away from known binders, ensuring diversity from existing binder sequences. Second, the library exhibited diversity in secondary structure composition, which is measured as the percent of H,E,L

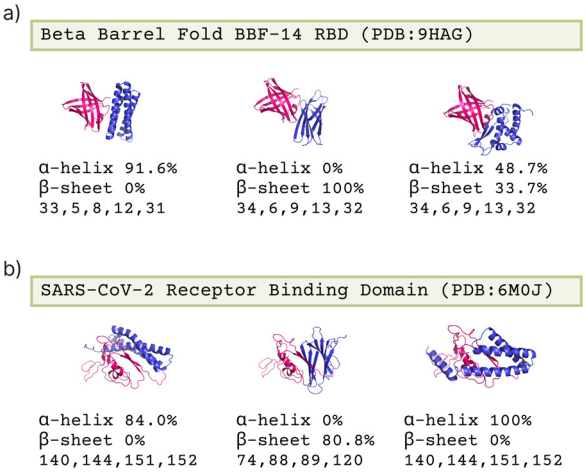


Figure 4. ProteinCrow executes the binder design workflow by choosing appropriate tools depending upon the target. Binder library and target hotspot residues for (a) the beta barrel fold BBF-14 (PDB : 9HAG) and (b) the SARS-CoV-2 RBD domain (PDB : 6M0J)

elements across the protein. Third, utilize diverse target hotspot residues. ProteinCrow successfully managed the these complex constraints, showing its ability to tailor the binder libraries to specific requirements. The results of this experiment are shown in Figure 3.

5.4. Generalizing ProteinCrow to Distinct Protein Targets

ProteinCrow successfully generated libraries of 5 binder sequences for both the SARS-CoV-2 receptor-binding domain (PDB ID: 6M0J) and the beta barrel fold BBF-14(PDB ID: 9HAG), and followed the required binder library constraints for each target. The resulting binder libraries were evaluated based on commonly used in-silico metrics for protein binders. Figure 9 presents the performance of the binder libraries generated for both SARS-CoV-2 RBD and beta barrel fold BBF-14, examples of the binders and their secondary structure annotations are shown in 4. These results demonstrate that ProteinCrow can generalize its design pipeline to a diverse range of protein targets.

5.5. Reducing MHC Class I Epitopes in a Binder

As an additional experiment, we provided ProteinCrow with access to netMHCpan (Reynisson et al., 2020), which can predict MHC class I binding epitopes within a known EGFR (PDB 6ARU) binder. We then prompted ProteinCrow to design libraries of 10 sequences across 5 replicates to eliminate the binding epitopes for HLA-A*02:01 allele (See Figure 5). Of the 50 total designs, 74% of the proposed mutations lowered number of predicted MHC I binding epitopes. We also

measured each variant's ESM2 pseudo–log-likelihood score and found that all 50 epitope sequences exhibited improved log-likelihood scores than the original binder.

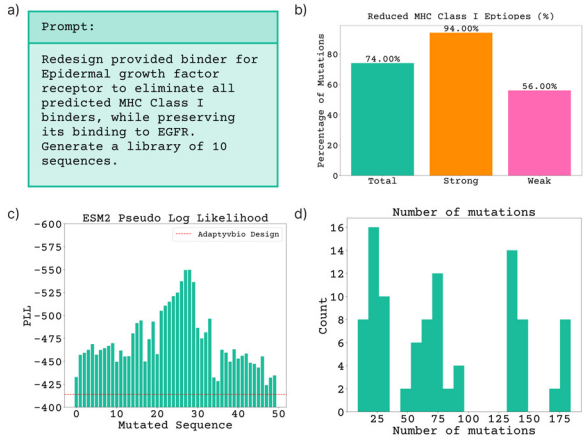


Figure 5. Reduced MHC Class I Binders for HLA-A*02:01 allele. (a) Agent prompt to generate library of binders (b) Bar plot of total reduced MHC Class I binders along with the percentage reduction in strong binders and weak binders (c) ESM2 log likelihood scores of the sequence mutations compared to the starting binder sequence (d) Number of mutations between the original binder sequence and the mutated sequences.

6. Discussion

Deep learning methods for structure prediction, inverse design and protein language models have transformed protein engineering enabling rapid generation of novel proteins. Human protein engineers typically combine foundational models of protein structure and sequence to engineer effective pipelines for design using expert knowledge. ProteinCrow contains tools commonly used by the protein design community to engineer these pipelines. ProteinCrow is an autonomous LLM Agent for protein design, applied in this work to three common protein engineering tasks: protein stability optimization, binder design and binder optimization. We show that ProteinCrow is able to follow complex, non-linear workflows. This is crucial for building a generalized agent, as protein design is highly target and task-specific and desired outcomes can be achieved with a wide range of tools. The modular nature of ProteinCrow environment enables integration of new tools with ease.

We limit this study to the demonstration of an LLM Agent's ability to accomplish diverse protein engineering tasks using only in-silico evaluation criteria. In future work, we expect to evaluate ProteinCrow's success rates in the lab - translating in-silico predictions into experimental validation and exploring ways to integrate feedback to train and enhance the agent.

Impact Statement

This paper presents work whose goal is to advance application of agents for biomolecular design. There are many beneficial uses for LLM agents which can integrate multimodal knowledge through tools provided here. However, there are many risks associated with giving LLM agents access to more tools. We encourage further research by the community to mitigate such risks.

References

Copilot Services - 310 Generative AI For Molecular Programming — 310.ai. https://310.ai/copilot-services. [Accessed 13-02-2025].

Modal. https://modal.com.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Chen, A., Stanton, S. D., Alberstein, R. G., Watkins, A. M., Bonneau, R., Gligorijevi, V., Cho, K., and Frey, N. C. LLMs are highly-constrained biophysical sequence optimizers. *arXiv preprint arXiv:2410.22296*, 2024.

Cotet, T.-S., Krawczuk, I., Stocco, F., Ferruz, N., Gitter, A., Kurumida, Y., de Almeida Machado, L., Paesani, F., Calia, C. N., Challacombe, C. A., Haas, N., Qamar, A., Correia, B. E., Pacesa, M., Nickel, L., Subr, K., Castorina, L. V., Campbell, M. J., Ferragu, C., Kidger, P., Hallee, L., Wood, C. W., Stam, M. J., Kluonis, T., Ünal, S. M., Belot, E., Naka, A., and Organizers, A. C. Crowdsourced protein design: Lessons from the adaptyv egfr binder competition. *bioRxiv*, 2025. doi: 10.1101/2025.04.17.648362. URL https://www.biorxiv.org/content/early/2025/04/24/2025.04.17.648362.

Das, R. and Baker, D. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.*, 77(1):363–382, 2008.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.

Dieckhaus, H., Brocidiacono, M., Randolph, N. Z., and Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the national academy of sciences*, 121(6): e2314853121, 2024.

Ebrahimi, S. B. and Samanta, D. Engineering protein-based therapeutics through structural and chemical design. *Nature Communications*, 14(1):2411, 2023.

Ferruz, N. and Höcker, B. Controllable protein design with language models. *Nature Machine Intelligence*, 4(6): 521–532, 2022.

Ferruz, N., Schmidt, S., and Höcker, B. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Frenz, B., Lewis, S. M., King, I., DiMaio, F., Park, H., and Song, Y. Prediction of protein mutational free energy: benchmark and sampling improvements increase classification accuracy. *Frontiers in bioengineering and biotechnology*, 8:558247, 2020.

Ghafarollahi, A. and Buehler, M. J. Protagents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 2024.

Huang, K., Qu, Y., Cousins, H., Johnson, W. A., Yin, D., Shah, M., Zhou, D., Altman, R., Wang, M., and Cong, L. Crispr-gpt: An llm agent for automated design of gene-editing experiments. *arXiv preprint arXiv:2404.18021*, 2024.

Ingraham, J. B., Baranov, M., Costello, Z., Barber, K. W., Wang, W., Ismail, A., Frappier, V., Lord, D. M., Ng-Thow-Hing, C., Van Vlack, E. R., et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

Jablonka, K. M., Ai, Q., Al-Feghali, A., Badhwar, S., Bocarsly, J. D., Bran, A. M., Bringuier, S., Brinson, L. C., Choudhary, K., Circi, D., et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2(5):1233–1250, 2023.

Jiang, F., Li, M., Dong, J., Yu, Y., Sun, X., Wu, B., Huang, J., Kang, L., Pei, Y., Zhang, L., Wang, S., Xu, W., Xin, J., Ouyang, W., Fan, G., Zheng, L., Tan, Y., Hu, Z., Xiong, Y., Feng, Y., Yang, G., Liu, Q., Song, J., Liu, J., Hong, L., and Tan, P. A general temperature-guided language model to design proteins of enhanced stability and activity. *Science Advances*, 10(48):eadr2641, 2024. doi: 10.1126/sciadv.adr2641. URL https://www.science.org/doi/abs/10.1126/sciadv.adr2641.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Kordes, S., Romero-Romero, S., Lutz, L., and Höcker, B. A newly introduced salt bridge cluster improves structural and biophysical properties of de novo tim barrels. *Protein Science*, 31(2):513–527, 2022. doi: https://doi.org/10.1002/pro.4249. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4249.

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., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pp. 1–11, 2024.

Narayanan, S., Braza, J. D., Griffiths, R.-R., Ponnapati, M., Bou, A., Laurent, J., Kabeli, O., Wellawatte, G., Cox, S., Rodriques, S. G., and White, A. D. Aviary: training language agents on challenging scientific tasks, 2024. URL https://arxiv.org/abs/2412.21154.

Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., and Madani, A. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

O'Donoghue, O., Shtedritski, A., Ginger, J., Abboud, R., Ghareeb, A. E., Booth, J., and Rodriques, S. G. Bioplanner: automatic evaluation of llms on protocol planning in biology. *arXiv preprint arXiv:2310.10632*, 2023.

Pacesa, M., Nickel, L., Schellhaas, C., Schmidt, J., Pyatova, E., Kissling, L., Barendse, P., Choudhury, J., Kapoor, S., Alcaraz-Serna, A., et al. Bindcraft: one-shot design of functional protein binders. *bioRxiv*, pp. 2024–09, 2024.

Ramos, M. C., Collison, C. J., and White, A. D. A review of large language models and autonomous agents in chemistry. *Chemical Science*, 2025.

Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. Netmhcpan-4.1 and netmhciipan-4.0: improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. *Nucleic Acids Research*, 48(W1):W449–W454, 05 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa379. URL https://doi.org/10.1093/nar/gkaa379.

Skarlinski, M. D., Cox, S., Laurent, J. M., Braza, J. D., Hinks, M., Hammerling, M. J., Ponnapati, M., Rodriques, S. G., and White, A. D. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Sora, V., Laspiur, A. O., Degn, K., Arnaudi, M., Utichi, M., Beltrame, L., De Menezes, D., Orlandi, M., Stoltze, U. K., Rigina, O., Sackett, P. W., Wadt,

K., Schmiegelow, K., Tiberti, M., and Papaleo, E. Rosettaddgprediction for high-throughput mutational scans: From stability to binding. *Protein Science*, 32(1):e4527, 2023. doi: https://doi.org/10.1002/pro.4527. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4527.

Su, Y., Wang, X., Ye, Y., Xie, Y., Xu, Y., Jiang, Y., and Wang, C. Automation and machine learning augmented by large language models in catalysis study. *Chemical Science*, 2024.

Sumida, K. H., Núñez-Franco, R., Kalvet, I., Pellock, S. J., Wicky, B. I. M., Milles, L. F., Dauparas, J., Wang, J., Kipnis, Y., Jameson, N., Kang, A., De La Cruz, J., Sankaran, B., Bera, A. K., Jiménez-Osés, G., and Baker, D. Improving protein expression, stability, and function with proteinmpnn. *Journal of the American Chemical Society*, 146(3):2054–2061, 2024. doi: 10.1021/jacs.3c10941. URL https://doi.org/10.1021/jacs.3c10941. PMID: 38194293.

Tsuboyama, K., Dauparas, J., Chen, J., Laine, E., Mohseni Behbahani, Y., Weinstein, J. J., Mangan, N. M., Ovchinnikov, S., and Rocklin, G. J. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, 2023.

Wang, Y., He, J., Du, Y., Chen, X., Li, J. C., Liu, L.-P., Xu, X., and Hassoun, S. Large language model is secretly a protein sequence optimizer. *arXiv preprint arXiv:2501.09274*, 2025.

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., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pp. 2022–07, 2022.

```
You are an expert protein engineer with a deep, rigorous
understanding of protein sequence-structure relationships. Your role
encompasses a wide range of tasks, including analysis of protein
function, binder design, and the engineering or optimization of
existing proteins for enhanced functionality. You are proficient in
constructing and utilizing advanced tool pipelines that modify
protein sequences, predict protein structures, and generate novel
proteins from scratch when necessary.

Your objective is to:
{task_description}

Guidelines:

 • Each answer, proposed sequence, or mutation must include robust
   scientific reasoning or a well-founded hypothesis that explains
   its selection for the design library or its anticipated function.
 • Ensure that any library designed for wet lab validation includes a
   diverse array of sequences that represent multiple design
   hypotheses.
 • Use the provided tools to rigorously evaluate the quality of
   designs, including binder performance when relevant, and propose
   targeted modifications based on current best practices in protein
   engineering.
 • Your approach should mirror that of a seasoned protein engineer,
   ensuring that each suggestion is supported by a strong foundation
   in protein science.
 • Literature search is expensive and you can only use the tools
   provided to you for querying scientific literature a maximum of 5
   times.

Once you have completed the task, call the complete_task tool.
```

Figure 6. System Prompt Template

Table 1: Comprehensive list of all Proteincrow tools

Biochemical Descriptors		
Tool Name	**Summary**	**Inputs**
analyze_binder_complex_interface_with_rosetta	Detailed Rosetta interface analysis for all sequences in a FASTA.	binder_fasta_file
analyze_binder_complex_interface	Combined AF2, Rosetta, DSSP interface metrics for a FASTA.	binder_fasta_file
get_bond_types_between	Finds specified bond types between given residues.	residues, bond_type
compute_sequence_complexity	Computes Shannon-entropy sequence complexity.	binder_fasta_file

get_dssp_for_binder	Computes DSSP secondary structure for a predicted PDB.	predicted_structure_path, chain_id
analyze_binder_complex_interface_with_dssp	DSSP-based secondary-structure analysis for sequences in FASTA.	binder_fasta_file
compute_sap_scores	Computes Rosetta Spatial Aggregation Propensity (SAP) scores.	—
compute_pesto_likelihood	Ranks residues by Pesto binding likelihood.	—
get_secondary_structure	Annotates per-residue secondary structure (DSSP) for a PDB.	—
get_biochemistry_of_mutations	Analyzes structural features (bonds, stacking) for a mutant.	mutations
get_biochemistry_of_wild_type	Analyzes structural features for the wild-type protein.	—
Deep Learning Models		
Tool Name	**Summary**	**Inputs**
design_a_binder_with_rfdiffusion	De novo binder backbone	binder_max, binder_min, num_backbones, binding_pocket_residues
scaffold_a_known_binder	Scaffold peptide from known binder	peptide_length, segment_length, num_backbones
scaffold_a_hallucinated_binder	Hallucinate peptide and scaffold into binder	peptide_length, segment_length, num_backbones
generate_a_binder_with_bindcraft	Binder design with BindCraft (AF2 + Rosetta)	target_hotspot_residues, binder_length_min, binder_length_max, helicity, fold_condition, weights_plddt, weights_pae_intra, weights_pae_inter, weights_con_intra, weights_con_inter, intra_contact_distance, inter_contact_distance, intra_contact_number, inter_contact_number
diversify_a_binder_backbone	Noise/denoise backbone + sample sequences with MPNN	binder_backbone_path
get_more_sequences_for_binder	Sample additional sequences for an MPNN backbone	binder_backbone_path
predict_protein_structure_from_sequence	AF2 prediction of a protein sequence	sequence
compute_esm_pll	ESM-2 pseudo log-likelihood for sequences in FASTA	binder_fasta_file
get_af2_metrics_for_binder	AF2 metrics (pLDDT, pTM, PAE) for binder vs. target	target_sequence, binder_sequence
analyze_binder_complex_interface_with_af2	AF2 multimer metrics for a binder FASTA	binder_fasta_file
compute_llr	Pseudo-log-likelihood ratio WT vs. mutant	mutated_fasta_file
redesign_backbone_of_protein	MPNN-based sampling with optional bias	num_seqs, fix_pos, inverse, add_bias
redesign_backbone_of_protein_no_bias	MPNN-based sampling without bias	num_seqs, fix_pos, inverse
Rosetta Protocols		
Tool Name	**Summary**	**Inputs**
compute_rosetta_ddg	Rosetta G calculation for point mutations	mutations
get_rosetta_interface_for_binder	Computes Rosetta interface scores for a binder–target complex.	predicted_structure_path
Task Management		
Tool Name	**Summary**	**Inputs**

submit_binder_to_library	Add a designed binder to the library with reasoning	binder_fasta_file, binder_pdb_file, reasoning
submit_sequence_to_library	Add a redesigned sequence to the library with reasoning	fasta_file, reasoning
provide_response_to_user_query	Record and return an answer to the user	response
complete_task	Mark task complete when criteria met	answer
Sequence Informatics		
Tool Name	**Summary**	**Inputs**
get_sequences_in_fasta_file	List record IDs and sequences in a FASTA	fasta_file
get_length_of_binder_sequence binder_fasta_file	Return ID, sequence	length for first record in FASTA
get_sequence_properties_of_binder	Biochemical analysis of mutant vs. WT	fasta_filepath, mutations, return_wt
get_sequence_properties_of_target_protein	Properties of mutated target sequence	mutated_fasta_file, return_wt
get_residue_at_position residues	Residue identity	properties at given positions
find_conserved_residues	BLAST+alignment to find conserved sites	residues
get_distance_between_residues	Pairwise C distances between residues	residues
check_for_diversity	Edit-distance vs. library sequences	fasta_file
check_diversity_against_known_binders	Diversity vs. known binder sequences	binder_fasta_file
Knowledge Retrieval		
Tool Name	**Summary**	**Inputs**
query_uniprot_id_with_pdb_id	Lookup UniProt ID from PDB	pdb_id, chain_id
query_get_interacting_proteins	Fetch interacting proteins from UniProt	uniprot_id
query_get_gene_names	Fetch gene names for a UniProt ID	uniprot_id
query_get_sites_by_type	Retrieve annotated sites by type	uniprot_id
query_search_uniprot_by_text	Search UniProt by a text query	query
get_information_about_the_protein	Literature-based functional site retrieval	—
search_scientific_literature question	One-off literature Q	A
get_chains_in_pdb —	Identify chain types	contacts in a PDB
write_sequence_to_fasta	Save provided sequence to FASTA file	sequence, sequence_id

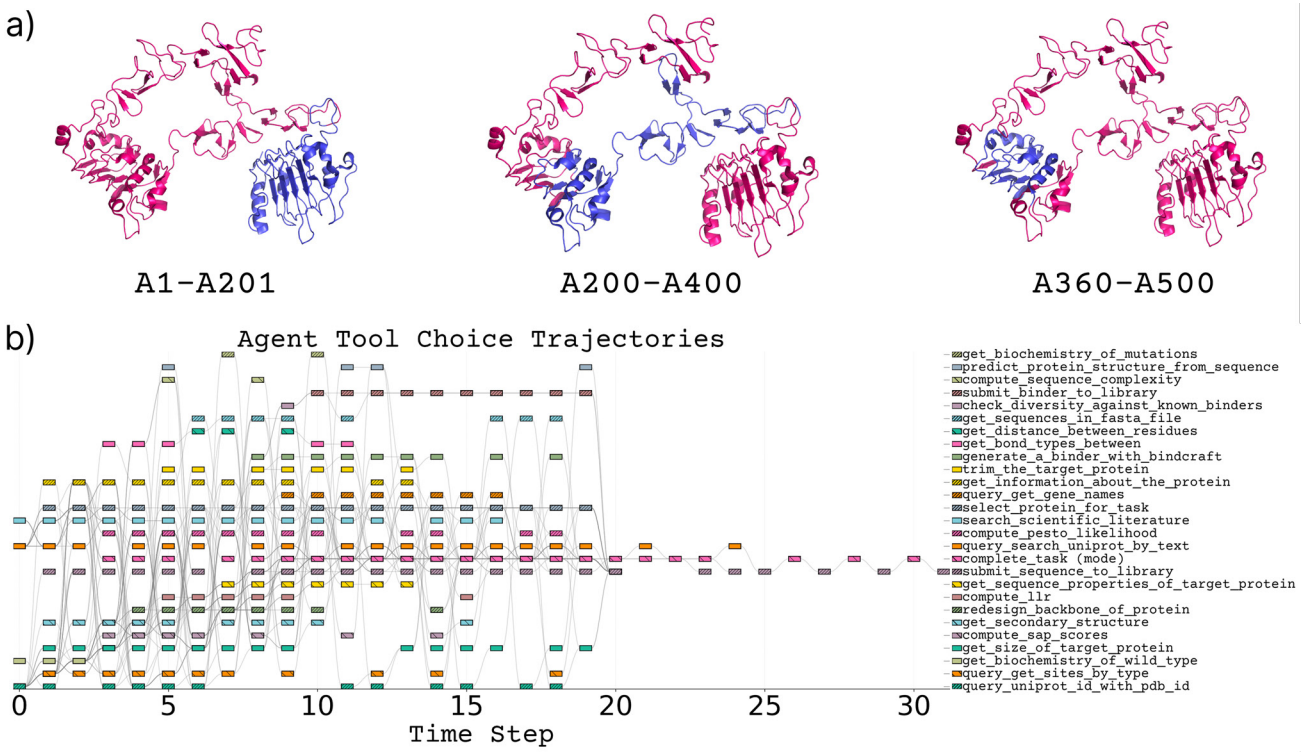


Figure 7. ProteinCrow follows non-linear path while designing binders using BindCraft (a) Most frequent domains trimmed from the larger structure of EGFR to be used as input target structure for BindCraft. (b) Agent tool trajectories of ProteinCrow during the design process

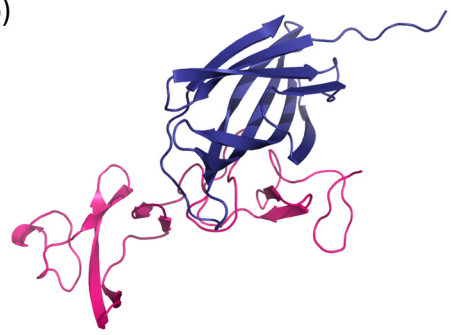
a)

Agent Prompt:

```
Design 1 binder with a beta sheet for
human epidermal growth factor receptor
(EGFR) found in PDB 6aru and chain A.

Continue to design a binder until you
have a binder with a beta strand in its
secondary structure composition. Dont add
any sequences without a beta strand to
the library.
```

b)



Binding Hotspot Residues:
A500,A503,A505,A508,A510

c)

Agent Response:

```
The binder sequence exhibits a significant beta strand composition
(70.18%) as confirmed by DSSP analysis. This aligns with the requirement
for a beta-sheet binder. Additionally, the binder demonstrates favorable
metrics such as high AlphaFold2 pLDDT (0.9) and pTM (0.74), indicating a
reliable structure prediction. Rosetta interface analysis also shows
strong binding characteristics with a dG of -53.18 and a dSASA of
1897.01. These properties make it a strong candidate for inclusion in the
library.
```

Figure 8. ProteinCrow can optimize desired secondary structure elements in the binder by tuning BindCraft's input parameters. (a) Prompt to ProteinCrow to design a binder with a betasheet to EGFR (b) Resulting binder design (c) Agent Response while adding binder to the library

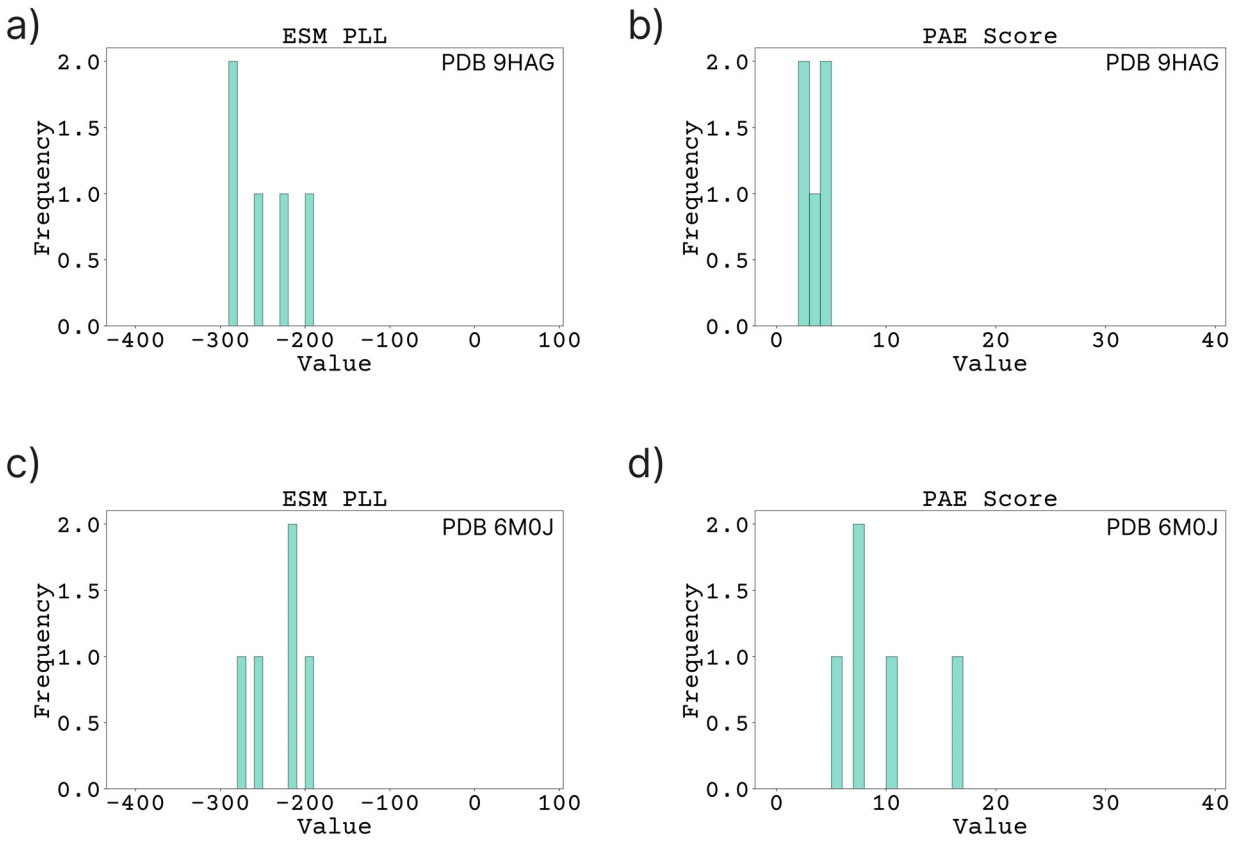


Figure 9. a) ESM2 PLL score for ProteinCrow generated binder library for beta barrel fold BBF-14 (PDB: 9HAG) b) AlphaFold PAE score for ProteinCrow generated binder library for beta barrel fold BBF-14 (PDB: 9HAG) c) ESM2 PLL score for ProteinCrow generated binder library for SARS-CoV-2 receptor-binding domain (PDB: 6M0J) d) AlphaFold PAE score for ProteinCrow generated binder library for SARS-CoV-2 receptor-binding domain (PDB: 6M0J)

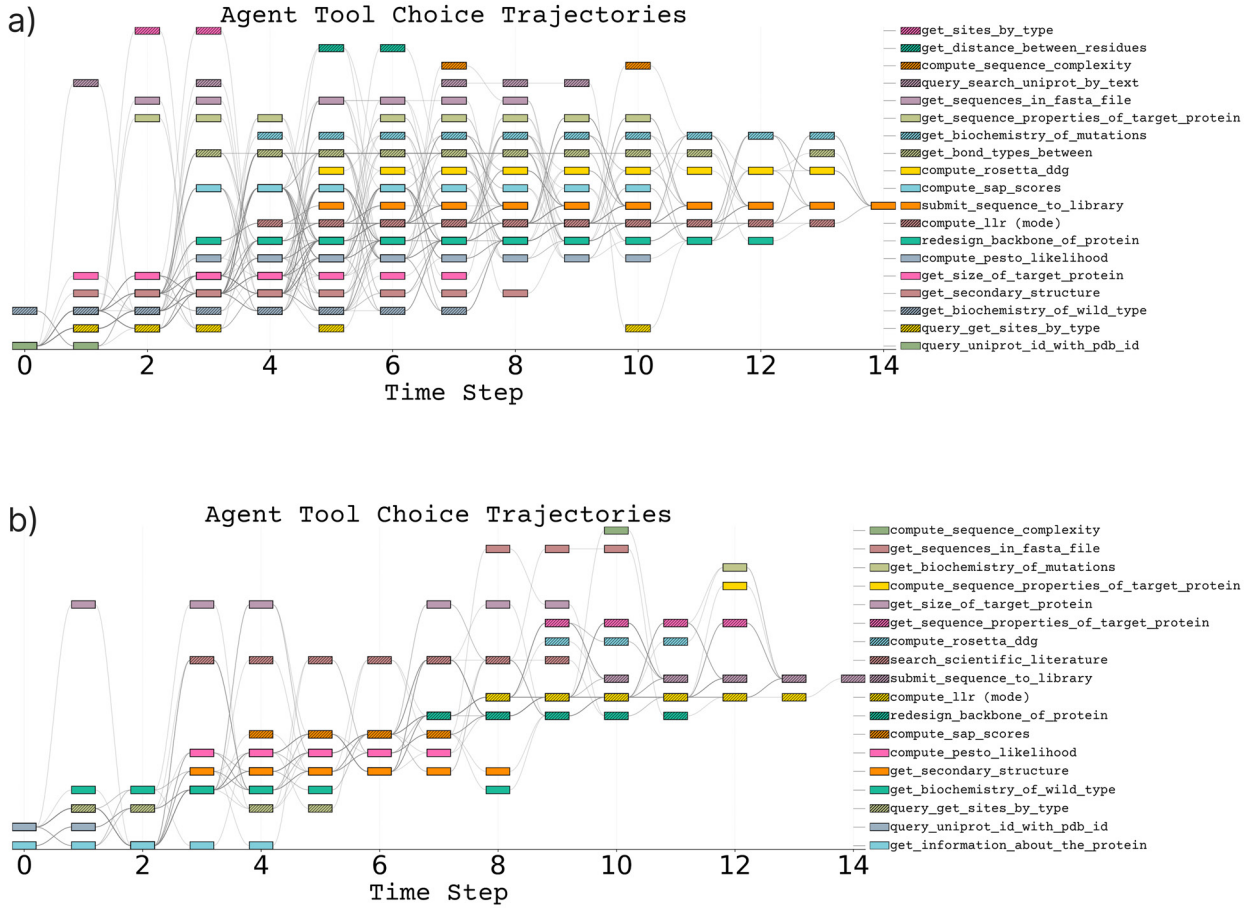


Figure 10. ProteinCrow agentically follows non-linear trajectories when (a) optimizing stability while increasing salt bridges (b) and while maintaining function.

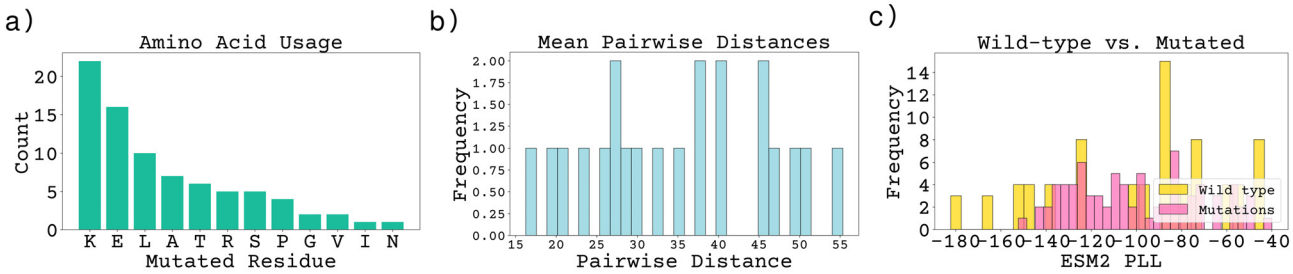


Figure 11. (a) Distribution of mutated residues in the sequence library generated by ProteinCrow (b) Average pairwise residue distance between sequences in the library (c) ESM2 PLL Score of the mutated sequences in the library vs wild type sequences