LLM-Assisted versus Agentic Approaches to De Novo Minibinder Design for a KRAS G12D Neoantigen

Anonymous Authors[1]

Abstract

We compare two modes of large language model (LLM) integration into de novo protein minibinder design: (1) human designs with LLM in the loop, where a researcher makes every scientific decision while using an LLM to accelerate code generation and analysis; and (2) LLM-agent designs with human in the loop, where the researcher specifies only the therapeutic target and the agent handles pipeline construction and in silico evaluation autonomously. Both target the KRAS G12D neoantigen peptide VVGADGVGK on HLA-A*11:01 starting with an existing framework (Liu et al., 2025). Across two rounds of design screens, the human-led campaign produced 15 high-confidence hits and the LLM agent presented top 41 hits. Evaluated by our new multi-layered evaluation suite of structure/sequence quality, Rosetta-based energy calculations, and docking verification, we conclude that the agent-led campaign is useful for exploring new design ideas and running large parallel campaigns, but domain expertise remained essential for correct binding orientation, calibrated specificity thresholds, and viable candidate selection.

1. Introduction

1.1. Minibinders Present Emerging Opportunities For Therapeutics Development

De novo protein minibinders are small (< 100 residues), hyperstable proteins designed to bind a specific surface patch with nanomolar-to-picomolar affinity (Cao et al., 2022; Watson et al., 2023; Bennett et al., 2023). The field has validated binders to more than 200 protein targets, with wide applications in medicine, nanotechnology, and various industrial sectors (Yang et al., 2026; Listgarten & Jiang, 2026). De novo minibinders tend to show high selectivity against human extracellular proteins, likely because their rigid, helix-rich scaffolds limit the induced-fit promiscuity seen in binders with flexible loops. Deep learning models have enabled rapid structure-based minibinder designs for therapeutics (Glögl et al., 2024; Ragotte et al., 2025; Sappington et al., 2024), while key challenges remain, including low binding affinity, off-target toxicity, and unclear functional activity (Bennett et al., 2023; Yu et al., 2026; van den Berg et al., 2015; Zhang et al., 2025; Son et al., 2024; Kosonocky et al., 2026).

1.2. Peptide–MHC Class I Complexes Are Important But Challenging Design Targets

Major Histocompatibility Complex Class I (MHC-I) molecules present 8–11-residue peptides derived from intracellular proteins on the cell surface for immune surveillance. In cancer cells, disease-driving mutations generate neoantigens and are presented as short peptides by MHC, which can be targeted by engineered proteins for therapeutics and diagnostics. Compared with conventional protein targets, pMHC-I complexes pose acute specificity challenges: the dominant surface is MHC, conserved across all peptides in a given allele, while the peptide exposes only a narrow, partially solvent-accessible strip. A binder must make extensive peptide contacts while discriminating the target from thousands of self-peptides on the same HLA—demands that make pMHC-I one of the hardest binder design challenges.

Three groups (Liu et al., 2025; Householder et al., 2025; Johansen et al., 2025; Hickok & Stromnes, 2025) published de novo pMHC-I minibinders using the same core toolkit: RFdiffusion (Watson et al., 2023) for backbone generation above the peptide-binding groove, ProteinMPNN (Dauparas et al., 2022) for sequence design, and AlphaFold2 (AF2) (Jumper et al., 2021; Bennett et al., 2023; Motmaen et al., 2023) for structural filtering and specificity screening. Collectively, they demonstrated that, for certain targets, it is possible to rapidly design de novo minibinders with high binding affinity, high specificity, and highly functional activity as part of re-engineered T cell receptors or T cell engagers. AlphaFold2, contact-based filtering (contact molecular score

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

(Cao et al., 2022)), and inverse-folding models (ProteinMPNN (Dauparas et al., 2022)) led to a desirable number hits for experimental validation within 2-3 design rounds. Other approaches (Pacesa et al., 2025; Stark et al., 2025; Cho et al., 2025) operate without iterative rounds but have not been benchmarked on pMHC-I.

A persistent challenge across all de novo binder campaigns is the gap between in silico confidence metrics and experimentally measured binding affinity. Large-scale studies using quantitative affinity data, spanning tens of thousands of antibodies and minibinder interactions, have shown that AlphaFold-derived metrics such as interaction prediction score from aligned errors (ipSAE) and interface predicted template modeling (ipTM) enrich for binders but produce high false-positive rates and fail to rank affinity reliably (Bio, 2026; Overath et al., 2025; Cotet et al., 2025). Performance varies markedly by target, point mutation campaigns are a particular blind spot, and providing structural templates during prediction improves but does not solve classification (Bio, 2026; Pak et al., 2023). These findings explain why large campaigns generating thousands of computationally screened designs typically yield only a handful of experimentally validated hits, and why developability remains unsystematically addressed.

1.3. LLM Integration in Protein Design

Three modes of LLM integration can be identified along a spectrum of human control:

Human designs with LLM in the loop The researcher makes every scientific decision and uses an LLM as an interactive programming assistant that generates code, debugs scripts, and summarizes the literature. The LLM is easy to adopt because it accelerates specific tasks but does not require significant changes to the research workflow.

LLM-agent designs with human in the loop Platforms such as Latent Labs (Latent Labs Team et al., 2025; 2026), Boltz, and Tamarind Bio accept a target specification and broad objectives, with an agent handling literature search, pipeline assembly, backbone generation, sequence design, and in silico evaluation (Marshall, 2026). The researcher reviews candidate outputs rather than intermediate decisions. These platforms have claimed nanobody and scFv binders to challenging targets including GPCRs and KRAS (G12D).

No human in the loop Fully autonomous AI-driven science, where agents independently form hypotheses, execute wet-lab experiments, and iterate without human intervention, is an aspirational frontier that several companies are actively pursuing. However, no peer-reviewed papers, preprints, or experimental benchmarks have been published to support these claims. For the purposes of this paper, fully autonomous protein design without human oversight remains an unvalidated aspiration rather than a demonstrated capability, and we do not attempt to reproduce it here.

2. Problem Statement

KRAS G12D is present in many cancers, yet current approaches have not generated any de novo binder for KRAS presented at the cell surface by MHC I with high specificity. The neoantigen peptide VVGADGVGK (HLA-A*11:01) differs from wildtype VVGAGGVGK by a single glycine-to-aspartate substitution at position 5. We use the Liu et al. (Liu et al., 2025) framework as a starting point and ask: (1) how effectively can a researcher with LLM assistance produce viable candidates for this novel, challenging target? (2) what can an LLM agent contribute? and (3) where does domain expertise remain irreplaceable?

3. Successful binder designs inform new design evaluation metrics

3.1. Analysis of Published Designs

Prior to launching our campaigns, we analyzed 18 published binder sequences (Liu et al., 2025) using AlphaFold3 (Abramson et al., 2024). When we omitted the β-2 microglobulin chain, two designs (prame-9 and sars-6) showed back-face MHC docking in blind AF3 predictions, i.e., the binder engaged the MHC α-helix groove rather than the peptide-presenting face. Running AlphaFold3 with default MSA setting (including MSAs for the designed minibinder in many cases) led to reduced contact between binder and peptide for 4 designs, as indicated by significantly decreased ipSAE value, though the change can be overlooked due to low overall binder root mean squared deviation (RMSD) (differences < 0.2). Therefore, we ran AlphaFold3 with the full pMHC-binder complex and disabled MSA generation for the minibinder.

Structure and sequence analysis of the 18 designs revealed consistent features: all adopt 3–4-helix topologies arcing above the peptide groove; 17 designs contact the peptide with α helices and 1 design uses a loop as the primary peptide-contacting element. Net charge is consistently negative (-2.8 to -10.8), no cysteine, and the amino acid composition is dominated by Asp, Glu, Ala, Leu, and Arg. Hotspot Rosetta contact molecular surface (CMS) scores are high (including polar residues), mean binder predicted local distance difference test (pLDDT) > 85, and ipSAE > 0.8 across validated designs (Figure A.1,Table 1). Compared to criteria defined by BindCraft on other targets (Pacesa et al., 2025), the designs show higher average interface hydrophobicity and more unsatisfied hydrogen bonds (Figure A.2).

3.2. Customized evaluation metrics for pMHC binders

We provide the following additional evaluation metrics in addition to AlphaFold2 initial guess (iPAE, binder pLDDT and binder RMSD), and AlphaFold2 monomer pLDDT:

Improved AlphaFold2 initial-guess scoring with ipSAE Each backbone–sequence pair is predicted using AlphaFold2-initial-guess protocol (Bennett et al., 2023). A key addition is to use **ipSAE** computed exclusively over binder–peptide residue pairs rather than standard interface predicted aligned error (iPAE), which is dominated by the much larger binder–MHC interface and is therefore a poor proxy for peptide-specific binding. Designs passing ipSAE ≥ 0.8 proceed. We examined iPAE and ipSAE values of the same designs, and observed that certain designs with iPAE>10 but ipSAE<0.8 still produced promising results upon further partial diffusion and sequence redesign.

Additional sequence quality evaluation To reduce non-specific binding, the designs are evaluated for net charge, surface hydrophobicity, and the presence of cysteine (can form disulfide bonds with unintended targets). The thresholds for net charge and surface hydrophobicity were calibrated based on authors' successful designs.

AlphaFold3-based evaluation on structural confidence and Rosetta energy calculations Template-based AlphaFold2 refolding decreases false negative rate, but introduces the issue of circular evaluation. AlphaFold2 models are also limited in cases where a de novo protein is present in the complex, meaning that no MSA can be computed effectively. Therefore, much of the downstream structure-based evaluations are based on template-free AlphaFold3 generated structures on a handful of high-confidence designs. Biophysical properties using Rosetta are commonly used in AI-based protein designs (Overath et al., 2025; Pacesa et al., 2025; Listgarten & Jiang, 2026), and were therefore incorporated into our evaluation framework.

It remains to be seen whether these additional metrics correlate with experimental outcomes.

4. Human Designs with LLM in the Loop

4.1. Overview of Pipeline and Evaluation Methods

The pipeline (Figure 1) follows an existing pipeline (Liu et al., 2025); the additions below constitute our primary methodological contribution. All implementations and result visualizations were written in collaboration with the Claude Code (Sonnet 4.6) with user-specified goals and feedback for each step. Each suggestion and script written by Claude were evaluated by the human researcher. The pipeline was executed on an academic computing cluster with CPU and GPU access.

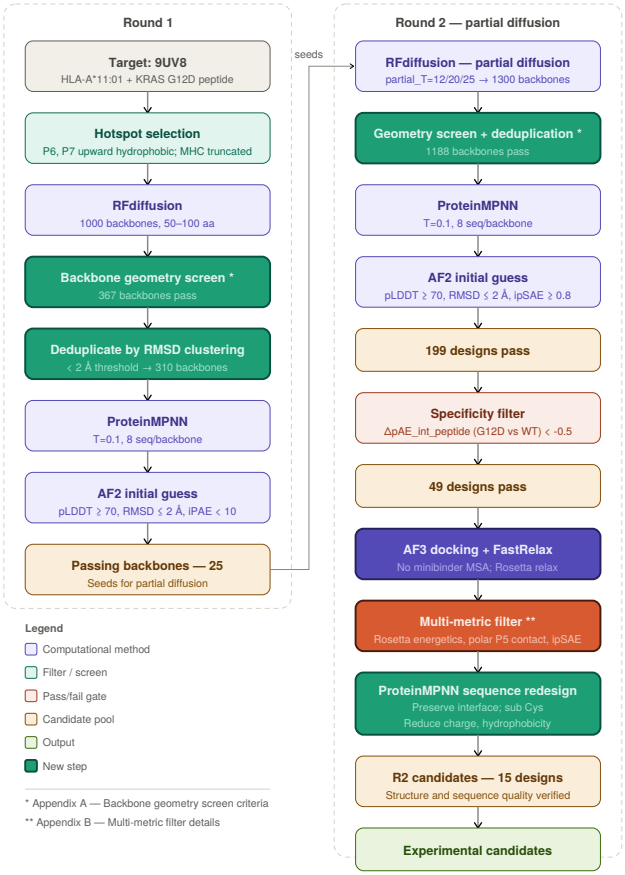

Figure 1. **Human-directed binder design pipeline.** Our framework generated 15 high-confidence de novo minibinders to the KRAS G12D pMHC complex within 2 rounds. To expedite the design process, we created a backbone geometry screen customized for this target and calibrated with experimentally validated designs from Liu et al. Additionally, we used AlphaFold3 to dock the designed binders without any templating to validate successful binding to the hotspot P5 G12D, binding energetics, contact scores, and structural confidence. A subset of the sequences passing the AlphaFold3 based filters underwent redesigns to improve downstream therapeutic developability properties.

To circumvent tedious manual inspections and wasting computing resources on poor backbones, careful research was conducted to filter backbones by geometric alignment with the peptide target and similarity with the authors' successful designs (secondary structure and radius of gyration) (More details in Appendix B). Among these, Claude suggested using the fraction of binder residues in proximity to the MHC over those in proximity to the peptide. While this makes intuitive sense, it practically removed promising backbones. Early human intervention avoided filtering out positive designs.

Round 1 produced only 2 designs with iPAE < 5 from 310 geometrically filtered backbones. Most designs also have

poly-Alanines, indicating low confidence in the backbones. This motivated Round 2 partial diffusion (partial_t 12–25) from 25 surviving backbones with a less stringent threshold (iPAE $<$ 10). This generated 199 designs passing all AlphaFold2 initial guess filters, of which 49 passed the primary specificity filter of Δ pAE_int_peptide $<$ -0.5, iPAE of each design against G12D (on-target) and G12 WT (off-target) peptides; we note this threshold has not been independently calibrated for this specific peptide pair and experimental validation is required to assess its discriminative power.

4.2. Final Design Candidates

Upon evaluation with AlphaFold3-based metrics (Appendix C), 15 designs have appropriate energetics, specific polar interactions and overall high CMS score with the peptide P5 G12D, despite higher number of buried unsaturated hydrogen bonds and higher fraction of exposed hydrophobicity, as expected (Section 3.1, Figure A.1, Figure A.2).

These passing designs cluster around the binder_255 scaffold lineages, with highly promising candidates from binder_34 and binder_252, containing 91 or 95 residues, helix dominant topologies (3 helices) arcing above the peptide groove, and net charge -10–+2 (Table 1). Contact map analysis shows that they all carry 1–2 Arg or Lys making a hydrogen bond or salt bridge to the p5 Asp of KRAS G12D, indicating good specificity for G12D over wild-type KRAS peptide (Figure 3).

FastRelax of the these designs yields dG_separated -52.9 to -67.6 REU (more negative indicates lower Rosetta interface energy), packstat 0.53–0.65 (higher indicates better interface packing quality), dSASA_int 2014–2858 Å² (higher indicates larger total buried interface area), delta_unsatHbonds 2–16 (lower indicates fewer unsatisfied buried hydrogen bonds, which is better), and CMS hotspot score 28–62 Å² (higher values indicates tighter geometric complementarity to the key specificity residue). All except for net charges and presence of 1 cysteine in some designs compare favorably with validated successful designs (Table 1,Figure A.1,Figure A.2). The issues with net charges and cysteine presence can be easily fixed with ProteinMPNN redesign while keeping the non-Cys interface residues fixed.

4.3. Results

Scale of screening Across 2 rounds, 10,500 backbone trajectories and 84,000 backbone–sequence pairs were evaluated to yield 15 specific candidates—a ratio of 5,600 computational designs per passing candidate, consistent with published campaigns (Overath et al., 2025; Liu et al., 2025). Due to target-specific research and benchmarking, it took several weeks and many CPU and GPU resources to complete the designs.

LLM-assisted programming accelerates the design process but lacks scalability Environmental setup used to be a major bottleneck in pre-LLM period, given missing YAML files and differences in computing systems. With LLM assistance, human researchers can quickly gather relevant information to install new tools and patch environmental issues. Researchers' domain expertise helps spot hallucination and improper result interpretations by the LLM. However, given the large number of decisions and multiple screening rounds of many candidates, it can get overwhelming to remember why each decision is made, especially when designing binders across very different targets.

5. Agentic Design

5.1. Overview of Pipeline

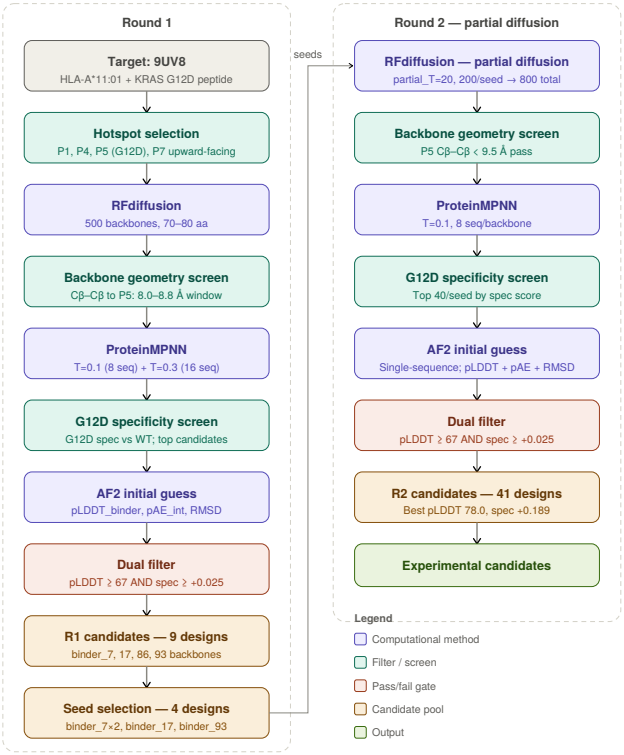

Figure 2. **Agentic binder design pipeline.** Summary of the design pipeline written by the agent. While the major steps match the paper provided in the prompt (Liu et al., 2025), the choice of evaluations and all executive decisions came from the agent.

5.1.1. Round 1

An LLM agent (Claude Sonnet 4.6, Anthropic) in the Claude Code harness (v2.1.132) was provided a detailed prompt instructing it to design a minibinder to pMHC presenting the KRAS G12D peptide (Appendix D.1). The agent was also provided a pdf of the relevant paper (Liu et al., 2025), its supplementary materials, and a relevant PDB template (Zhu

et al., 2025), and AlphaFold 3-no-MSA predicted pMHC-binder complexes for the published minibinders. Given these inputs, the agent autonomously read the paper, installed relevant tools (RFdiffusion (Watson et al., 2023), ProteinMPNN (Dauparas et al., 2022), and ColabFold (Mirdita et al., 2022)), wrote scripts to run the tools, and managed their execution (Figure 2).

Three main tasks were performed in Round 1:

- 500 RFdiffusion backbone samples with peptide hotspot conditioning
- ProteinMPNN sequence design on each backbone
- AlphaFold2 initial guess for structure prediction and filtering

Three manual interventions were needed:

- A nudge to follow RFdiffusion's installation instructions from its README, as the agent was encountering errors creating an incorrect environment.
- A cleanup of hanging background shells.
- An instruction to run AF2 on the remaining RFdiffusion designs 93-500 – the agent had stopped folding designs after encountering errors. The agent chose to run AF2 only on 10 designs with highest predicted specificity for G12D, increased to 80 on further nudging.

5.1.2. Round 2

After Round 1 was complete, the agent was prompted to run Round 2, consisting of partial diffusion and sequence generation (Appendix D.2). The agent applied partial diffusion (partial_T=20) from four Round 1 seeds, followed by geometry filtering, ProteinMPNN redesign, and AF2 initial-guess evaluation on the top 40 designs per seed.

5.1.3. Computational cost

The two agentic rounds completed in about 1.5 days, with a token cost of $85.57. Together with cost of the GPU server (a single cloud-hosted Nvidia RTX 5090), the total computational cost was roughly $100.

5.1.4. OpenAI Codex

We also attempted the same Round 1 prompt using OpenAI's Codex agent (GPT-5.5). However, the agent stopped at requesting AF3 structures for the already-provided structures from the paper. After a nudge to continue without AF3 structures, it opted to download and run Boltz-2. However, at this point, the agent ran out of tokens. It is likely that this model is capable of running the same tools, but it would need a different prompt or harness to help minimize human intervention.

5.2. Final Design Candidates

Round 1 generated 500 backbones, of which 250 passed a P5 proximity screen (minimum Cβ distance < 9.5 Å). Approximately 2,576 designed sequences were scored for binding specificity; the agent-defined "specificity score" was the change in ProteinMPNN likelihood of the binder given the variant peptide versus the wild-type peptide. Only 4 backbones (binder_7, binder_17, binder_86, binder_93) yielded AF2-predicted pMHC interfaces (pLDDT $\geq$ 67, pAE_interaction below the 31.5 "no contact" floor); the remaining backbones uniformly showed no interface signal despite passing geometry. From these 4 backbones the agent identified 9 experimental candidates (pLDDT 67–75, specificity score +0.037–+0.112).

Round 2 partial diffusion from the 4 seeds produced 800 backbones, 634 passing the geometry screen. Evaluating 160 designs (top 40 per seed by specificity), 41 passed combined filters (pLDDT $\geq$ 67, specificity $\geq$ 0.025). Sequences were 32–68% identical to Round 1. Round 2 improved on Round 1: best pLDDT rose from 75.0 to 77.96 (binder_22_s4, seed b7_t03_s7), best specificity from +0.112 to +0.189 (binder_160_s5, seed b7_t03_s2), and binder_68_s1 (seed b17_s1) reached RMSD = 2.0 Å to its design model. However, these features rarely co-occurred: only *1 design* passed the pLDDT $\geq$ 70 and RMSD $\leq$ 2.0 Å filters.

5.3. Results

Human review of the agent's analysis scripts, decisionmaking log, and results revealed several critical issues. A full review is in Appendix E. Overall, the agent was able to set up the tools and largely correctly use the tools. In the case where the agent mixed up the protein chain order input in partial diffusion, it was able to realize the lack of change in generated backbones and fix the problem without human intervention.

Innovative ways to screen designs before compute-heavy steps After ProteinMPNN sequence generation, instead of proceeding with the relatively long AlphaFold 2 initial guess to fold the designs, the agent used ESMFold to calculate binder pLDDT, which is fast and performs better than AlphaFold2 on α helix dominated miniproteins (Hýsková et al., 2026). To efficiently predict binders' preference for the G12D variant, the agent devised a ProteinMPNN-based "specificity score". Rather than computing the dssp, which requires protein folding, the agent used rough estimates based on amino acid (helix formers and strand formers), and computed aromatic contents (beneficial to hydrophobic

packing), and alanine fractions. However, it still could not control the high alanine fractions.

Incorrect wild-type KRAS sequence The wildtype KRAS residue at codon 12 is Glycine (G), not Valine or Cysteine, which are two other common cancer-driving mutations, and the correct sequence was present in the input files, yet the agent used the wrong residue in both the iPAE off-target comparison and the ProteinMPNN log-likelihood scoring across multiple scripts. Providing the pdb structure of the wildtype KRAS pMHC (e.g., 8I5E) in the prompt may mitigate this problem.

Suboptimal hotspot residue choice The agent included C5 (Aspartic acid (D), polar) as an RFdiffusion hotspot, even though RFdiffusion's documentation explicitly discourages polar hotspots, and Liu et al. (2025), the paper provided to the agent, demonstrated that G12D should be targeted with hotspot residues adjacent to C5. In addition, the agent targeted both terminal anchor residues C1 and C7 as hotspots, constraining the geometric search space against the bent peptide conformation. Adding instructions to read the README page of the RFdiffusion repository may solve the first problem, while the second problem comes from lack of familiarity with pMHC structures and protein design.

Implausible sequences and structures All of the generated minibinders contained long stretches of Alanine (A). While this also occurred in the human Round 1, it was corrected in human Round 2 by partial diffusion and resampling, but the agentic Round 2 did not fix the issue. Also, only 4 of 500 Round 1 backbones folded satisfactorily under AlphaFold2, yet even with this filtering, the structure quality of the selected samples was poor (Table 1) and in some cases there was no contact between the minibinder and the peptide (Figure 3B).

6. Comparison of Final Design Candidates

We observe very different characteristics between designs by the agents (n=41 designs from n=39 backbones) and by the humans (n=15 designs from n=12 backbones) and authors (n=18 designs) (Table 1). Both humans' and authors' designs have high pLDDT (local structural confidence), low binder aligned RMSD (indicating high similarity to the backbones), low iPAE or high ipSAE (good interface), a tight range of radius of gyration (indicating there is an ideal range of protein structural compactness), and low lateral distance between the center of mass of binder and that of the peptide. The converse is true for the agent's designs, meaning that human's designs are more likely to bind than the agent's designs. Notably, while the humans and authors' designs all have at least 3 α helices, 33/41 agents' designs have only 1 or 2 α helices, and all 3-α-helix designs have very different lengths, leading to poor packing and suboptimal contacts with the target. The vertical distance between authors' and agents' designs to target peptide are comparable, while the human's designs are vertically further from the target.

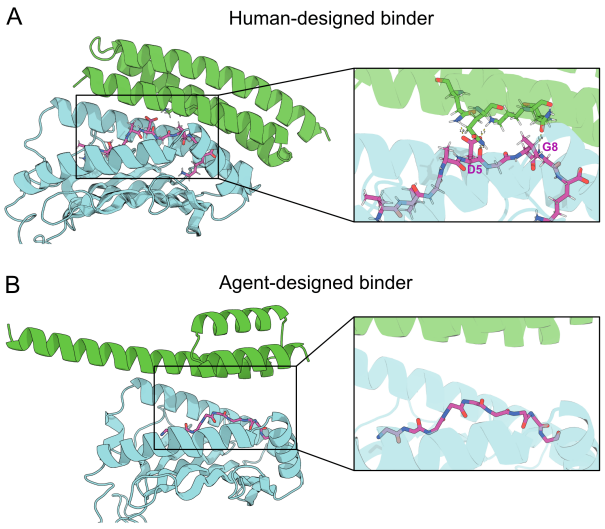

Figure 3. **Predicted structures of binding complexes.** HLA-A*11:01 α1 and α2 domains in complex with KRAS G12D 9-mer peptide (VVGADGVGK) and designed minibinders. A) Human-designed minibinder 34 and B) Agent-designed binder 68. Inset shows peptide and any polar contacts with the minibinder.

7. Discussion

Despite growing interest in LLM agents for scientific research, no published benchmark has systematically compared agent-directed and human-directed campaigns on the generative protein binder design task. Existing evaluations fall into two categories that each capture only part of the problem. Crowdsourced human competitions (Cotet et al., 2025; Kosonocky et al., 2026), experimentally validating designs from over human submissions, provide ground-truth experimental benchmarks for computational methods but involve human designers throughout, not autonomous agents. Agent-focused benchmarks, conversely, have evaluated LLM agents on protein engineering prediction tasks (fitness scoring, sequence classification) rather than on end-to-end generative pipelines (Miller et al., 2025). So far, LLM agents underperformed human baselines on biomedical ML workflows, with no consistent advantage for domain-specialized over generalist agents (Miller et al., 2025; Phylo, 2026). The most proximate precedent is ProteinMCP, which demonstrated that an LLM agent can autonomously orchestrate de novo protein design but without benchmarking candidate quality or experimental hit rates against human-

Table 1. **Structural and quality metrics across three design cohorts.** *Author*: 18 published designs from Liu et al. (2025). *Human R2*: 15 designs from Round 2 of the human-led pipeline (subset with AF2 binder RMSD available). *Agent V1*: 52 designs from the first agentic round. Continuous metrics shown as min–max (median). n_helix_segs breakdown shows counts at 0/1/2/3/… helix segments. iPAE: AF2 initial-guess pAE_interaction. ipSAE: AF3 ipSAE binder–peptide. *AF2 monomer pLDDT. **AF2 initial-guess complex binder pLDDT. [§]Per Liu et al. (2025); all crystal structures confirmed RMSD < 2 Å.

Detailed descriptions of these metrics are in B and C

Metric	**Author** ($n = 18$)	**Human R2** ($n = 15$)	**Agent R2** ($n = 41$)
`binder_plddt` ↑	86.9–97.3 (95.4) *	87.2–96.7 (95.1) *	47.9–78.0 (70.8) **
`binder_rmsd` (Å) ↓	< 2 [§]	0.4–1.1 (0.6)	3.8–60.9 (22.9)
`iPAE` ↓	< 5	3.7–5.7 (4.9)	31.5–31.7 (31.6)
`ipSAE` ↑	0.86–0.97 (0.94)	0.80–0.94 (0.86)	N/A
`rg` (Å)	11.6–14.9 (14.2)	14.7–15.2 (14.9)	17.5–34.1 (18.5)
`n_helix_segs` (#designs at 1/2/… helices)	0/0/6/12	0/0/15/0	5/28/8/0
`lateral_dist` (Å) ↓	1.0–13.2 (4.9)	5.2–12.4 (10.8)	11.4–27.0 (21.9)
`vertical_dist` (Å) ↓	2.0–12.6 (10.6)	12.2–14.4 (13.5)	1.2–12.0 (4.7)

directed campaigns (Xu et al., 2026). To our knowledge, the present work is the first to hold target, evaluation criteria, and pipeline constant while varying the degree of human versus agent control—providing a direct, if preliminary, comparison on a therapeutically relevant generative design task.

We show that Claude Code, a popular agentic tool in software engineering, can rapidly and cheaply conduct an *in silico* protein binder design campaign. The level of human interaction required was low, but not zero. However, the resulting sequences are of much lower quality than a comparable human-directed campaign. While Claude Code is capable of using protein design software, it lacks "biological intuition" to recognize when the outputs do not make sense. We expect that specific training of frontier language models on biological tasks (Claude for Life Sciences, GPT-Rosalind) will improve built-in biological intuition and therefore performance at this task.

8. Future Directions

We have several ideas to increase the success rate, scope, and autonomy of our agentic workflow:

Reducing bias towards specific methods Our prompt was fairly specific and structured, including a highly relevant example paper with approaches to mimic. If we made the prompt more minimal, it might allow the agent to explore novel approaches. The caveat to this freedom is that the model is likely to fall back on its preferred tools—those well-represented in its training set. To increase the diversity of approaches, explicitly asking for new ideas in a separate planning stage may help.

Multi-agent task management We expect harness developments such as multi-agent orchestrator-implementor-evaluator patterns to improve long-running task management https://www.anthropic.com/engineering/harness-design-long-running-apps. These can be incorporated into future prompts.

Aligning the agent with our expectations While Claude is natively able to run tools like RFdiffusion and ProteinMPNN, skill development may help align the model's goals with our expectations. The grill-me skill and other interactive planning skills would be particularly useful in identifying knowledge gaps and assumptions between the prompter and the agent before the agentic process begins. However, to be successful, this approach still depends on the agent knowing what it does not know, which requires built-in experience and intuition using protein design tools.

Software and Data

The human design scripts are deposited online at https://codeberg.org/icml-submission-wb9kc/pmhc_binder_design_human. The agentic scripts and design log are at https://codeberg.org/icml-submission-wb9kc/pmhc_binder_design_claude.

Impact Statement

This paper presents work whose goal is to advance the fields of Machine Learning and Biology. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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., and Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016): 493–500, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://www.nature.com/articles/s41586-024-07487-w. Publisher: Nature Publishing Group.

Bennett, N. R., Coventry, B., Goreshnik, I., Huang, B., Allen, A., Vafeados, D., Peng, Y. P., Dauparas, J., Baek, M., Stewart, L., DiMaio, F., De Munck, S., Savvides, S. N., and Baker, D. Improving de novo protein binder design with deep learning. *Nature Communications*, 14 (1):2625, May 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38328-5. URL https://www.nature.com/articles/s41467-023-38328-5. Publisher: Nature Publishing Group.

Bio, A.-A. Overconfident: What Structure Prediction Confidence Scores Tell Us About Binding, February 2026. URL https://aalphabio.substack.com/p/overconfident-what-structure-prediction.

Cao, L., Coventry, B., Goreshnik, I., Huang, B., Sheffler, W., Park, J. S., Jude, K. M., Marković, I., Kadam, R. U., Verschueren, K. H. G., Verstraete, K., Walsh, S. T. R., Bennett, N., Phal, A., Yang, A., Kozodoy, L., DeWitt, M., Picton, L., Miller, L., Strauch, E.-M., DeBouver, N. D., Pires, A., Bera, A. K., Halabiya, S., Hammerson, B., Yang, W., Bernard, S., Stewart, L., Wilson, I. A., Ruohola-Baker, H., Schlessinger, J., Lee, S., Savvides, S. N., Garcia, K. C., and Baker, D. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910): 551–560, May 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04654-9. URL https://www.nature.com/articles/s41586-022-04654-9. Publisher: Nature Publishing Group.

Cho, Y., Pacesa, M., Zhang, Z., Correia, B. E., and Ovchinnikov, S. Boltzdesign1: Inverting All-Atom Structure Prediction Model for Generalized Biomolecular Binder Design, April 2025. URL https://www.biorxiv.org/content/10.1101/2025.04.06.647261v1. Pages: 2025.04.06.647261 Section: New Results.

Cotet, T.-S., Krawczuk, I., Stocco, F., Ferruz, N., Gitter, A., Kurumida, Y., Machado, L. d. A., 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, April 2025. URL https://www.biorxiv.org/content/10.1101/2025.04.17.648362v2. Pages: 2025.04.17.648362 Section: New Results.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, October 2022. doi: 10.1126/science.add2187. URL https://www.science.org/doi/10.1126/science.add2187. Publisher: American Association for the Advancement of Science.

Glögl, M., Krishnakumar, A., Ragotte, R. J., Goreshnik, I., Coventry, B., Bera, A. K., Kang, A., Joyce, E., Ahn, G., Huang, B., Yang, W., Chen, W., Sanchez, M. G., Koepnick, B., and Baker, D. Target-conditioned diffusion generates potent TNFR superfamily antagonists and agonists. *Science*, 386(6726):1154–1161, December 2024. ISSN 1095-9203. doi: 10.1126/science.adp1779. Place: New York, N.Y.

Hickok, G. H. and Stromnes, I. M. Beyond the native repertoire. *Science*, 389(6758):349–351, July 2025. doi: 10.1126/science.adz6423. URL https://www.science.org/doi/10.1126/science.adz6423. Publisher: American Association for the Advancement of Science.

Householder, K. D., Xiang, X., Jude, K. M., Deng, A., Obenaus, M., Zhao, Y., Wilson, S. C., Chen, X., Wang, N., and Garcia, K. C. De novo design and structure of a peptide–centric TCR mimic binding module. *Science*, 389(6758):375–379, July 2025. doi: 10.1126/science.adv3813. URL https://www.science.org/doi/10.1126/science.adv3813. Publisher: American Association for the Advancement of Science.

Hýsková, A., Maršálková, E., and Šimeček, P. Balancing speed and precision in protein folding: a

comparison of AlphaFold2, ESMFold, and OmegaFold. *Frontiers in Genetics*, 16, January 2026. ISSN 1664-8021. doi: 10.3389/fgene.2025.1715037. URL https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2025.1715037/full. Publisher: Frontiers.

Johansen, K. H., Wolff, D. S., Scapolo, B., Fernández-Quintero, M. L., Risager Christensen, C., Loeffler, J. R., Rivera-de Torre, E., Overath, M. D., Kjærgaard Munk, K., Morell, O., Viuff, M. C., Lacunza, I., Damm Englund, A. T., Due, M., Gharpure, A., Forli, S., Rodriguez Pardo, C., Tamhane, T., Qingjie Andersen, E., Haldrup Björnsson, K., Fernandes, J. S., Voss, L. F., Thumtecho, S., Ward, A. B., Ormhøj, M., Reker Hadrup, S., and Jenkins, T. P. De novo-designed pMHC binders facilitate T cell–mediated cytotoxicity toward cancer cells. *Science*, 389(6758):380–385, July 2025. doi: 10.1126/science.adv0422. URL https://www.science.org/doi/10.1126/science.adv0422. Publisher: American Association for the Advancement of Science.

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., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Publisher: Nature Publishing Group.

Kosonocky, C. W., Abel, A. M., Feller, A. L., Rieffer, A. E. C., Woolley, P. R., Lála, J., Barth, D. R., Gardner, T., Competitors, B. t. B., Ekker, S. C., Ellington, A. D., Wierson, W. A., and Marcotte, E. M. Validation and analysis of 12,000 AI-driven CAR-T designs in the Bits to Binders competition, March 2026. URL https://www.biorxiv.org/content/10.64898/2026.03.03.709355v2. ISSN: 2692-8205 Pages: 2026.03.03.709355 Section: New Results.

Latent Labs Team, Bridgland, A., Crabbé, J., Kenlay, H., Pretorius, D., Schmon, S. M., Hilmkil, A., Bartke-Croughan, R., Rombach, R., Flashman, M., Matteson, T., Mathis, S., Nelson, A. W. R., Yuan, D., Obika, A., and Kohl, S. A. A. Latent-X: An Atom-level Frontier Model for De Novo Protein Binder Design, July 2025. URL http://arxiv.org/abs/2507.19375. arXiv:2507.19375 [q-bio].

Latent Labs Team, Schmon, S. M., Pretorius, D., Mathis, S., Bartke-Croughan, R., Puvanendran, A., Vuckovic, J., Kenlay, H., Vlachynská, M., Bridgland, A., Grishin, I., Over, S., Li, D., Li, B., Crabbé, J., Hilmkil, A., Nelson, A. W. R., Yuan, D., Obika, A., and Kohl, S. A. A. Latent-Y: A Lab-Validated Autonomous Agent for De Novo Drug Design, April 2026. URL http://arxiv.org/abs/2603.29727. arXiv:2603.29727 [q-bio].

Listgarten, J. and Jiang, H. How artificial intelligence is reengineering protein engineering. *Science*, 392(6794):159–166, April 2026. doi: 10.1126/science.aec8444. URL https://www.science.org/doi/10.1126/science.aec8444. Publisher: American Association for the Advancement of Science.

Liu, B., Greenwood, N. F., Bonzanini, J. E., Motmaen, A., Meyerberg, J., Dao, T., Xiang, X., Ault, R., Sharp, J., Wang, C., Visani, G. M., Vafeados, D. K., Roullier, N., Nourmohammad, A., Scheinberg, D. A., Garcia, K. C., and Baker, D. Design of high-specificity binders for peptide–MHC-I complexes. *Science*, 389(6758):386–391, July 2025. doi: 10.1126/science.adv0185. URL https://www.science.org/doi/10.1126/science.adv0185. Publisher: American Association for the Advancement of Science.

Marshall, A. AI turbocharges antibody hunt for binders with drug-like qualities. *Nature Biotechnology*, 44(3):334–337, March 2026. ISSN 1546-1696. doi: 10.1038/s41587-026-03048-w.

Miller, H. E., Greenig, M., Tenmann, B., and Wang, B. BioML-bench: Evaluation of AI Agents for End-to-End Biomedical ML, September 2025. URL https://www.biorxiv.org/content/10.1101/2025.09.01.673319v3. ISSN: 2692-8205 Pages: 2025.09.01.673319 Section: New Results.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, June 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1.

Motmaen, A., Dauparas, J., Baek, M., Abedi, M. H., Baker, D., and Bradley, P. Peptide-binding specificity prediction using fine-tuned protein structure prediction networks. *Proceedings of the National Academy of Sciences of the United States of America*, 120(9):e2216697120, February 2023. ISSN 1091-6490. doi: 10.1073/pnas.2216697120.

Overath, M. D., Rygaard, A., Jacobsen, C. P., Brasas, V., Morell, O., Sormanni, P., and Jenkins, T. P. Predicting Experimental Success in De Novo Binder Design: A Meta-

Analysis of 3,766 Experimentally Characterised Binders, August 2025. URL https://www.biorxiv.org/content/10.1101/2025.08.14.670059v1. ISSN: 2692-8205 Pages: 2025.08.14.670059 Section: New Results.

Pacesa, M., Nickel, L., Schellhaas, C., Schmidt, J., Pyatova, E., Kissling, L., Barendse, P., Choudhury, J., Kapoor, S., Alcaraz-Serna, A., Cho, Y., Ghamary, K. H., Vinué, L., Yachnin, B. J., Wollacott, A. M., Buckley, S., Westphal, A. H., Lindhoud, S., Georgeon, S., Goverde, C. A., Hatzopoulos, G. N., Gönczy, P., Muller, Y. D., Schwank, G., Swarts, D. C., Vecchio, A. J., Schneider, B. L., Ovchinnikov, S., and Correia, B. E. One-shot design of functional protein binders with BindCraft. *Nature*, 646(8084): 483–492, October 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09429-6. URL https://www.nature.com/articles/s41586-025-09429-6. Publisher: Nature Publishing Group.

Pak, M. A., Markhieva, K. A., Novikova, M. S., Petrov, D. S., Vorobyev, I. S., Maksimova, E. S., Kondrashov, F. A., and Ivankov, D. N. Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLOS ONE*, 18(3):e0282689, March 2023. ISSN 1932-6203. doi: 10.1371/journal.pone.0282689. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0282689. Publisher: Public Library of Science.

Phylo. Evaluating AI Agents in Biology, February 2026. URL https://phylo.bio/blog/evaluating-ai-agents-in-biology.

Ragotte, R. J., Tortorici, M. A., Catanzaro, N. J., Addetia, A., Coventry, B., Froggatt, H. M., Lee, J., Stewart, C., Brown, J. T., Goreshnik, I., Sims, J. N., Milles, L. F., Wicky, B. I. M., Glögl, M., Gerben, S., Kang, A., Bera, A. K., Sharkey, W., Schäfer, A., Harkema, J. R., Baric, R. S., Baker, D., and Veesler, D. Designed miniproteins potently inhibit and protect against MERS-CoV. *Cell Reports*, 44(6):115760, June 2025. ISSN 2211-1247. doi: 10.1016/j.celrep.2025.115760.

Sappington, I., Toul, M., Lee, D. S., Robinson, S. A., Goreshnik, I., McCurdy, C., Chan, T. C., Buchholz, N., Huang, B., Vafeados, D., Garcia-Sanchez, M., Roullier, N., Glögl, M., Kim, C. J., Watson, J. L., Torres, S. V., Verschueren, K. H. G., Verstraete, K., Hinck, C. S., Benard-Valle, M., Coventry, B., Sims, J. N., Ahn, G., Wang, X., Hinck, A. P., Jenkins, T. P., Ruohola-Baker, H., Banik, S. M., Savvides, S. N., and Baker, D. Improved protein binder design using beta-pairing targeted RFdiffusion, November 2024. URL https://www.biorxiv.org/content/10.1101/2024.10.11.617496v2. Pages: 2024.10.11.617496 Section: New Results.

Son, A., Park, J., Kim, W., Lee, W., Yoon, Y., Ji, J., and Kim, H. Integrating Computational Design and Experimental Approaches for Next-Generation Biologics. *Biomolecules*, 14(9):1073, September 2024. ISSN 2218-273X. doi: 10.3390/biom14091073. URL https://www.mdpi.com/2218-273X/14/9/1073. Publisher: Multidisciplinary Digital Publishing Institute.

Stark, H., Faltings, F., Choi, M., Xie, Y., Hur, E., O'Donnell, T., Bushuiev, A., Uçar, T., Passaro, S., Mao, W., Reveiz, M., Bushuiev, R., Pluskal, T., Sivic, J., Kreis, K., Vahdat, A., Ray, S., Goldstein, J. T., Savinov, A., Hambalek, J. A., Gupta, A., Taquiri-Diaz, D. A., Zhang, Y., Hatstat, A. K., Arada, A., Kim, N. H., Tackie-Yarboi, E., Boselli, D., Schnaider, L., Liu, C. C., Li, G.-W., Hnisz, D., Sabatini, D. M., DeGrado, W. F., Wohlwend, J., Corso, G., Barzilay, R., and Jaakkola, T. BoltzGen: Toward Universal Binder Design, November 2025. URL https://www.biorxiv.org/content/10.1101/2025.11.20.689494v1. ISSN: 2692-8205 Pages: 2025.11.20.689494 Section: New Results.

van den Berg, J. H., Gomez-Eerland, R., van de Wiel, B., Hulshoff, L., van den Broek, D., Bins, A., Tan, H. L., Harper, J. V., Hassan, N. J., Jakobsen, B. K., Jorritsma, A., Blank, C. U., Schumacher, T. N. M., and Haanen, J. B. A. G. Case Report of a Fatal Serious Adverse Event Upon Administration of T Cells Transduced With a MART-1-specific T-cell Receptor. *Molecular Therapy*, 23(9):1541–1550, September 2015. ISSN 1525-0016. doi: 10.1038/mt.2015.60. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4817886/.

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., and Baker, D. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL https://www.nature.com/articles/s41586-023-06415-8. Publisher: Nature Publishing Group.

Xu, X., Feng, C., Zha, C., He, W., He, M., Xiao, B., and Gao, X. ProteinMCP: An Agentic AI Framework for Autonomous Protein Engineering, March 2026. URL https://www.biorxiv.org/content/10.64898/2026.03.11.711149v1. ISSN: 2692-8205 Pages: 2026.03.11.711149 Section: New Results.

Yang, W., Wang, S., Lee, G. R., Zhang, J. Z., Courbet, A., Juergens, D., Wang, X., Schlichthaerle, T., Abedi, M., Ragotte, R., An, L., Kalvet, I., Pellock, S., Mihaljevic, L., Glasscock, C., Pillai, A., Broerman, A., Ennist, N., Haefner, E., McNamara-Bordewick, N., Haydon, I., Stewart, L., Bhardwaj, G., and Baker, D. The past, present and future of de novo protein design. *Nature*, 652:1139–1152, April 2026. ISSN 0028-0836. doi: 10.1038/s41586-026-10328-7. URL https://ui.adsabs.harvard.edu/abs/2026Natur.652.1139Y. ADS Bibcode: 2026Natur.652.1139Y.

Yu, Q., Guo, L., Qin, X., Huang, X., Tian, B., Wang, H., Liu, Y., Lang, Y., Wang, D., Shen, Z., Lin, J., and Chen, M. High-Affinity Protein Binder Design via Flow Matching and In Silico Maturation, January 2026. URL https://www.biorxiv.org/content/10.64898/2026.01.19.700484v2. ISSN: 2692-8205 Pages: 2026.01.19.700484 Section: New Results.

Zhang, G., Liu, C., Li, W., Lu, J., Li, A., and Zhu, L. Beyond evolution: *De novo* designed protein toolkit rewriting the rules of synthetic biology. *Biosafety and Health*, 7(5):306–311, October 2025. ISSN 2590-0536. doi: 10.1016/j.bsheal.2025.09.004. URL https://www.sciencedirect.com/science/article/pii/S2590053625001314.

Zhu, J., Chen, Z., Xu, X., Wang, Y., Liu, P., Wen, M., Wang, Q., He, Y., Jin, H., Xue, H., Wang, S., Xu, K., and Zhao, L. Structure guided analysis of KRAS G12 mutants in HLA-A*11:01 reveals a length encoded immunogenic advantage in G12D. *Communications Biology*, 9 (1):26, December 2025. ISSN 2399-3642. doi: 10.1038/s42003-025-09285-0. URL https://www.nature.com/articles/s42003-025-09285-0. Publisher: Nature Publishing Group.

A. Comparison of published and human-designed minibinders

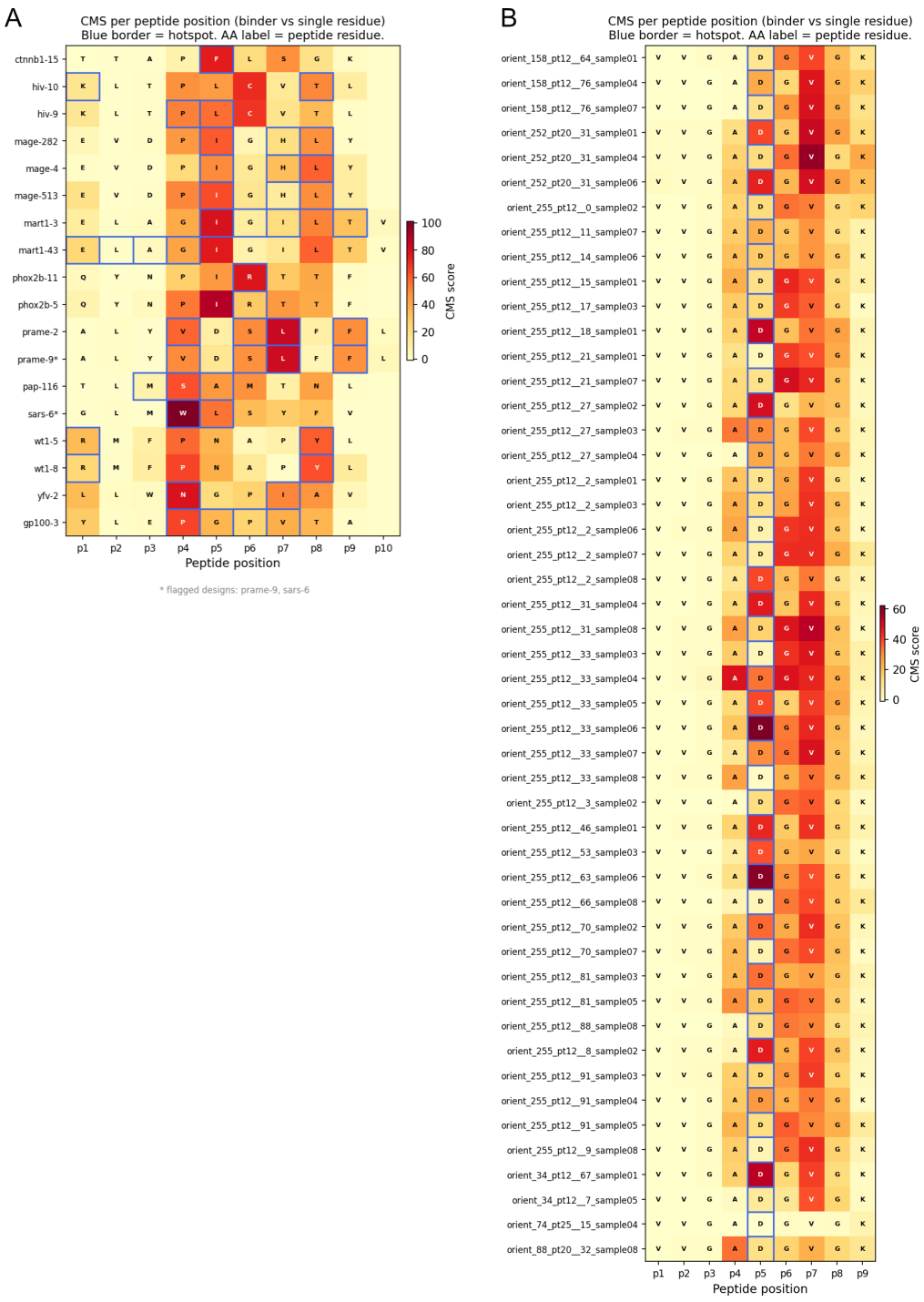

Figure A.1. **CMS scores comparison.** Comparison of contact molecular surface (CMS) scores (Cao et al., 2022) between A) published minibinders (Liu et al., 2025) and B) human-designed minibinders, at each position in the HLA-displayed peptide (x), for each minibinder (y).

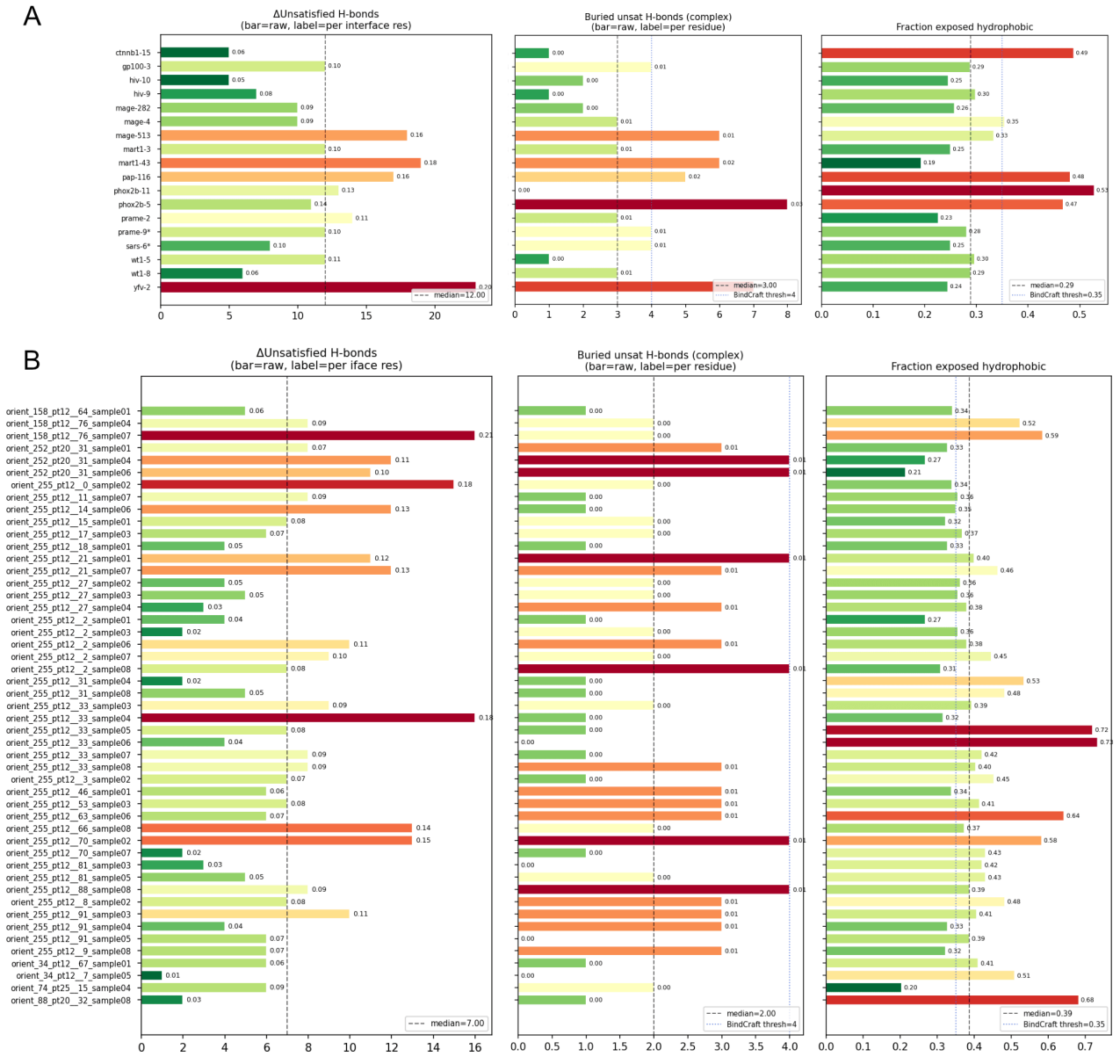

Figure A.2. **H-bonds and hydrophobicity comparison.** Comparison of hydrogen bonding and exposed hydrophobic residues between A) published minibinders (Liu et al., 2025) and B) human-designed minibinders. ΔUnsatisfied H-bonds: difference in number of potential hydrogen bond donors or acceptor atoms without a partner between the bound complex and unbound subunits (lower is better). Buried unsatisfied H-bonds: number of H-bonding atoms buried in the core with an insufficient number of H-bonds (lower is better). Fraction exposed hydrophobic: fraction of hydrophobic residues exposed on the solvent-accessible surface (lower is better).

B. Backbone geometry screen criteria

Residue count (60–100), helix segment count (≥2 by DSSP), radius of gyration within 13-23, no Cα clashes, high peptide contacts, and SVD-based orientation aligned to the MHC-I groove.

- Secondary structure (≥2 alpha helices)
- Radius of gyration (13-23)
- No clashes (≥4.0 Å)

- Peptide contact C alpha <8.0 Å for P5 (G12D)
- backbone principle axis and peptide principle axis $\leq$ 35 degree
- displacement of the binder COM relative to the peptide COM (max lateral distance $\leq$ 15 Å)

C. AlphaFold3-based multi-metric filtering

Hard filters (any failure = 1 hard flag):

- dG_separated $<$ -50 REU
- dG_separated/dSASAx100 $>$ -2.3
- dSASA_int $<$ 2000 Å^2
- hbonds_int / nres_int $<$ 0.07
- delta_unsatHbonds / nres_int $>$ 0.20 (calibrated to authors: 4/18 hit $>$ 0.15, max author = 0.20 for yfv-2)
- surface_hydrophobicity $>$ 0.60 (hard; 0.35 = BindCraft soft thresh)
- interface_n_K $>$ 3
- ipsae_binder_peptide $<$ 0.80

Soft flags (reported but do not count toward hard flag total):

- sc_value $<$ 0.55
- packstat $<$ 0.55
- buns_delta_unsat $\geq$ 4 (soft: 7/18 validated authors hit this)
- binder_aligned_rmsd $>$ 1.0 (flag for structural inspection)

Specificity signal (reported separately, used for tiering):

- n_saltbridge_hotspot_p5 $\geq$ 1 (R/K contacting G12D neoepitope)
- cms_hotspot_p5 $\geq$ 25 (CMS at p5)
- n_contacts_hotspot_p5_polar $\geq$ 1

Tiering (after hard flags counted):

- Tier 1: 0 hard flags + salt bridge or hbond at p5
- Tier 2: 0 hard flags, no salt bridge or hbond at p5 (sorted by cms_hotspot_p5)
- Redesign candidate: 1 hard flag (hydrophobicity only) + salt bridge/hbond at p5 + CMS score $\geq$ 25
- Soft flag only: 1 hard flag (non-hydrophobicity), or soft flags only
- Deprioritize: 2+ hard flags
- Exclude: 4+ hard flags

Contact map analysis (addition beyond Liu et al.): We compute three contact types between binder and peptide: all-atom contacts $<$ 4.0 Å; polar hydrogen-bond geometry (donor–acceptor $<$ 4.0 Å, D–H···A angle $\geq$ 120°, D···A $<$ 3.5 Å); and hydrophobic contacts $<$ 4.0 Å.

D. Agent Prompts

D.1. Minibinder design round 1

```
You are a computational protein design scientist who wants to design miniprotein binders
    to pMHC presenting the KRAS G12D peptide. All relevant info is in ./pmhc_design folder,
     including:
- PDB file of the target: pmhc_design/input/9UV8.pdb

See this paper for details on a relevant approach:
- pdf: pmhc_design/science.adv0185.pdf
- supplements: pmhc_design/science.adv0185_sm.pdf
- Additional: pmhc_design/science.adv0185.pdf
- Experimentally validated designs:
   - sequences: pmhc_design/paper_output/science.adv0185_data_s1.xlsx
   - AF3 predicted structures of the binders with the targeted peptide MHC complex (no MSA
       for the designed binders): pmhc_design/paper_output/af3_nomsa
   - Experimentally determined structures: pmhc_design/paper_output/mmdb_9O5S.pdb

Designs will be tested for these features:
- Folding confidence
- RMSD between the backbone and the actual binder from the predicted structure after
    alignment
- Net charge
- Secondary structure (percentage of residues in alpha helices, beta sheets, loops)
- CMS scores for the 5th peptide of KRAS G12D
- Number of hydrophobic and polar interactions between the binder and the peptide (
    interface)
- ipSAE between the peptide and the binder
- Energetics calculated after FastRelax
- Shape complementarity and packstat between the binder and the target
- Number of unsaturated hydrogen bonds at the interface
- Number of buried unsaturated hydrogen bonds within the complex
- Number of lysines and methionines at the binding interface
- Fraction of exposed hydrophobic residues
- Presence of poly alanines
However, you do not need to run all of these metrics.

You have access to a RTX 5090 on this system. You may install and run standard protein
    design models, such as RFdiffusion, ProteinMPNN, ESM, AlphaFold 2 monomer, AlphaFold 2
     initial guess, and generate input to run AlphaFold 3 on the web server (limited to 30
     jobs per day). A conda virtual environment is available at /venv/main, with pytorch
    cuda 13.0 installed. You may install standard python packages to that virtual
    environment as needed.

I have created an git repo at ./pmhc_design. Add any scripts, code, and analyses you need
    to run. You may commit to it as needed.
Keep a log of your design process, including any decisions made given intermediate results.
```

D.2. Minibinder design round 2

```
Perform partial diffusion (check the pmhc_design/science.adv0185_sm.pdf for details) on
    the top-ranking few designs and redo the sequence generation and downstream
    evaluations in the paper. At the end, compare generated sequences with round 1.
    Designs will be tested for these features:
- Folding confidence
- RMSD between the backbone and the actual binder from the predicted structure after
    alignment
- Net charge
- Secondary structure (percentage of residues in alpha helices, beta sheets, loops)
- CMS scores for the 5th peptide of KRAS G12D
- Number of hydrophobic and polar interactions between the binder and the peptide (
    interface)
- ipSAE between the peptide and the binder
```

```
- Energetics calculated after FastRelax
- Shape complementarity and packstat between the binder and the target
- Number of unsaturated hydrogen bonds at the interface
- Number of buried unsaturated hydrogen bonds within the complex
- Number of lysines and methionines at the binding interface
- Fraction of exposed hydrophobic residues
- Presence of poly alanines
However, you do not need to run all of these metrics.
```

E. Human review of agent results

RFdiffusion The hotspot selection is suboptimal, as RFdiffusion struggles to target polar residues effectively, yet C5 (the G12D mutation site) was included as a hotspot. Additionally, including both C1 and C7 as hotspots makes the design significantly harder to succeed, given the bent structure described in the Nature Communications paper associated with the uploaded 9UV8.pdb. The input structure was also not truncated, which unnecessarily consumes computing resources since chain B for the beta globulin is not required.

ProteinMPNN No constraints were applied to the charges or types of residues, and there were no considerations made to preserve the interface residues. The run temperature was set to 0.1, but also 0.3, which may be too high for early design rounds.

AF2 Screening (no initial guess) ColabFold was used for screening. Designs were screened using ESMFold pLDDT and ProteinMPNN scores, selecting the top 5. However, early rounds require a much larger number of designs, as the scores are not particularly discriminatory when all designs perform poorly in the first round. The approaches are smart but need to be applied to more designs.

AF2 Prediction (pMHC fold and initial guess) The iPAE calculation uses the wrong wild-type amino acid: it should be G, not C. The initial guess also used a different model, alphafold2_multimer_v3, though the MSA mode is correct.

Score Specificity (first pass) A new per-residue score based on Kyte-Doolittle hydrophobicity was introduced, which could be useful for building a loss function. Rather than instructing ProteinMPNN to omit certain amino acids as in BindCraft, it would be more useful to focus on surface hydrophobicity instead. The wild-type residue for KRAS was incorrect: it should be G, but V was used. For sequence properties, rather than computing DSSP, rough estimates based on amino acid propensities (helix formers, strand formers) should be used, along with aromatic content and alanine fractions. The ProteinMPNN log likelihood metric is, in practice, not particularly useful when we have backbones at different lengths. It is, however, very useful to rank the sequences after we narrow down to promising backbones. Contact scoring between the peptide and binder was computed based on C-alphas, which is acceptable for early rounds, but later analyses will need to account for amino acid type and contact type, distinguishing polar from hydrophobic contacts.

Score Backbones Backbone scoring should be performed immediately after backbone generation. The number of peptide contacts based on C-beta is a good metric, as is the focus on residues C4, C5, and C7. Hotspot coverage is a reasonable approach, though the hotspot selection itself is poor. The pep_frac vs. mhc_frac metric is good, and requiring a polar residue at C5 is the correct consideration.

Score Specificity (second pass) It is unclear how this script differs from `05_score_specificity.py`. The wrong wild-type residue for the peptide was used again.

Partial Diffusion Following instructions from the human, the run started with 4 backbones, which should be evaluated. Hotspots were included for partial diffusion and should also be used in the human version. Diffusion was run for 20 steps, which may be too many — the output needs to be checked.

Analyze Designs Rosetta was not accessible, which should be remedied by providing access. Lysine (K) and methionine (M) counts were calculated together rather than separately. Exposed hydrophobicity could not be calculated without Rosetta. For ipSAE, a distance cutoff should be applied. The analysis of interaction modes between the binder and the target is good. The wild-type pMHC complex PDB should have been provided from the outset.

Summary The agents should be prompted to consult the README of the relevant repositories. The RFdiffusion README explicitly advises against using polar residues as hotspots, and this reflects a gap in practical knowledge that can nonetheless be bridged easily. Insufficient context and tools were provided: Rosetta was not accessible, the wild-type PDB was not supplied, and the agents were not instructed to use evaluation metrics such as net charge and surface hydrophobicity to iteratively improve designs. On the positive side, the overall process was very fast and cost-effective, making it well-suited for exploring a wide range of ideas, and the use of ESMFold and ColabFold was an innovative choice. A major recurring error was the use of the wrong wild-type residue for KRAS on multiple occasions, pointing to a gap in biological knowledge. Overall, this is a solid start, and with greater human input and guidance, the workflow has strong potential.