A180 · Can AI Agents Design Proteins? Agentic vs. Human-Directed De Novo Minibinder Design for a KRAS Neoantigen

Presenter: Yilan Wang — Harvard Medical School

Authors: Yilan Wang, Aaron Kollasch

Session: Day 2 · Fri Oct 2, 2026 · 4:15–5:15 PM

Agentic AI—large language model (LLM) agents that autonomously plan, write code, and operate scientific software—is rapidly entering biology, yet it remains unclear whether such agents can match human experts on real generative protein-design tasks. We present, to our knowledge, the first head-to-head benchmark of agentic versus human-directed de novo protein binder design on a therapeutically relevant target: a minibinder to the KRAS G12D neoantigen peptide VVGADGVGK presented on HLA-A*11:01, one of the hardest peptide–MHC design challenges. We compare three modes of LLM integration along a spectrum of human control: (1) human-directed design with an LLM as a coding assistant; (2) a fully autonomous Claude Code agent given only the target and broad objectives; and (3) an agent operating under continuous expert supervision and harness engineering. Using a custom in silico evaluation suite spanning structure and sequence quality, Rosetta interface energetics, and AlphaFold-based specificity metrics, the human campaign produced 17 high-confidence hits, while the minimally supervised and human-supervised agent campaigns yielded 41 and 6 top candidates, respectively. Agents cheaply explored large parallel design spaces and, when supervised, recovered human-like binding geometry and G12D-specific salt bridges. However, unsupervised agents made domain-specific errors—incorrect target sequences, poor hotspot choices, and miscalibrated specificity thresholds—that expert oversight corrected. Our results show that agentic AI can accelerate protein design, but domain expertise remains essential for correct binding orientation, calibrated evaluation, and viable candidate selection.

Can AI Agents Design Proteins? Agentic vs. Human-Directed De Novo Minibinder Design for a KRAS Neoantigen

Yilan Wang[1,*] Aaron Kollasch

[1]*Department of Systems Biology, Harvard Medical School, Boston, MA, USA*

*Presenting author. Correspondence: wangyilan0304@gmail.com

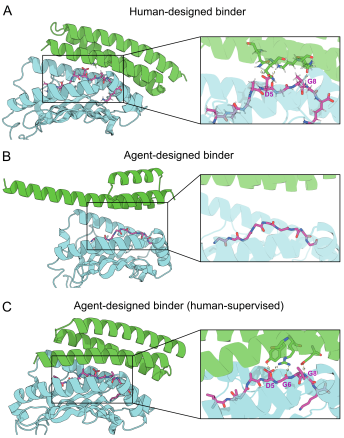

Figure. Predicted structures of binding complexes: HLA-A*11:01 α1 and α2 domains (cyan) in complex with the KRAS G12D 9-mer peptide (magenta) and designed minibinders (green). (A) Human-designed binder, (B) minimally supervised agent-designed binder, and (C) human-supervised agent-designed binder. Insets show the peptide and its polar contacts with the minibinder; supervised designs recover the D5-anchored salt-bridge geometry that confers G12D specificity.

Agentic AI—large language model (LLM) agents that autonomously plan, write code, and operate scientific software—is rapidly entering biology, yet it remains unclear whether such agents can match human experts on real generative protein-design tasks. We present, to our knowledge, the first head-to-head benchmark of agentic versus human-directed de novo protein binder design on a therapeutically relevant target: a minibinder to the KRAS G12D neoantigen peptide VVGADGVGK presented on HLA-A*11:01, one of the hardest peptide–MHC design challenges. We compare three modes of LLM integration along a spectrum of human control: (1) human-directed design with an LLM as a coding assistant; (2) a fully autonomous Claude Code agent given only the target and broad objectives; and (3) an agent operating under continuous expert supervision and harness engineering. Using a custom *in silico* evaluation suite spanning structure and sequence quality, Rosetta interface energetics, and AlphaFold-based specificity metrics, the human campaign produced 17 high-confidence hits, while the minimally supervised and human-supervised agent campaigns yielded 41 and 6 top candidates, respectively.

We constructed our own *in silico* benchmarking criteria, calibrated against the experimentally validated, published minibinder designs for pMHC complexes in Liu et al. (*Science*). Across two rounds, the campaigns collectively evaluated roughly 2,300 RFdiffusion backbones and tens of thousands of backbone–sequence pairs. The human and human-supervised agentic designs align with the Liu et al. designs, sharing a consistent, favorable profile: high predicted local confidence (binder pLDDT 94–95), low RMSD to their backbones, strong interface metrics (low iPAE, high ipSAE), a tight radius of gyration, and a small lateral offset between binder and peptide. The minimally supervised agent designs were the opposite on nearly every axis: all human-supervised designs adopted three or more α-helices, whereas 33 of 41 minimally supervised agent designs used only one or two helices, and the three-helix remainder varied widely in length, giving poor packing and weak peptide contacts.

Left unsupervised, the agent made errors a protein-design expert would not: it scored off-target specificity against an incorrect wild-type KRAS residue, selected polar and terminal-anchor hotspots contrary to tool guidance, and produced implausible poly-alanine sequences. Supplying full scientific context, curated literature, calibrated thresholds, and careful context and harness engineering closed much of this gap—the supervised agent recovered human-design-like topologies and the Arg–Asp(p5) salt bridge underlying G12D specificity (only 0/17 human designs, versus 6/6 minimally supervised designs, suffered ineffective p5 polar contacts). This required heavy human involvement: 64 manual interventions versus 3, a more than 20-fold increase.

Both agentic campaigns were fast and inexpensive—about 1.5 to 6 days and roughly $100–$150 including a single cloud GPU—showing that agents can cheaply run large parallel campaigns and surface novel design ideas. Yet across metrics the human-directed pipeline still prevailed, and only expert oversight secured correct binding orientation, calibrated specificity, and viable candidate selection. This is, to our knowledge, the first benchmark to hold target, evaluation criteria, and pipeline constant while varying the degree of human versus agent control; we expect targeted training of frontier models on biological tasks to narrow the remaining gap.