HUMANS VS. AUTONOMOUS AGENTS IN DE NOVO PROTEIN BINDER DESIGN: A POST-COMPETITION ANALYSIS OF THE MUNI X ADAPTYV TREM2 HACKATHON

Anonymous authors
Paper under double-blind review

ABSTRACT

We present, to our knowledge, the first experimentally validated comparison of human-led and autonomous-agent de novo protein binder design. In a single-day hackathon, human teams and six agents submitted 81 and 60 designs under common constraints; 100 designs from nine human teams and six agents were tested. Of these, 89 expressed and 37 displayed binding to TREM2. Human and agent hit rates were 25/65 (38.5%) and 12/35 (34.3%), respectively (two-sided Fisher's exact P=0.83), with no clear cohort advantage. The strongest binders had per-design mean K_D values of 1.113 nM (human) and 3.636 nM (agent) by SPR. Experimental stability measurements were not available for the tested designs. Sequence-validated Protenix and Chai-1 predictions placed the strongest binder and all 36 mini-protein hits at TREM2's hydrophobic tip, outside the principal crystallographic contact sites of scFv-2 and scFv-4; 46/49 expressed mini-protein non-binders were also predicted to contact this region. Retrospective Protenix ipSAE and ipTM distinguished binders from non-binders (AUROC 0.782 and 0.791) but correlated weakly with affinity among 36 fitted binders (Spearman rho=0.293 and 0.196), supporting binder triage but not reliable affinity ranking. A separate 15-round, human-directed agentic search produced ten candidates; nine yielded fitted BLI mean K_D values below 10 nM and three below 1 nM. In this setting, the follow-up produced numerically stronger reported top-end affinities than the original hackathon, although the campaigns differed in duration, selection and assay. Anthropic's subsequent Claude Science study selected TREM2 as a campaign target and benchmarked its autonomous binder-design results against the publicly released hackathon and agentic follow-up datasets. Together, these results establish the first experimentally validated benchmark of human and agentic protein design, showing similar observed hit rates under common submission constraints, while also highlighting the value of iterative multi-round agentic search.

1 INTRODUCTION

Computational protein design has reached an inflection point driven by AI-based structure prediction and generative models. BindCraft achieves 10–100% binding success rates across diverse targets (Pacesa et al., Nature 2025), a dramatic improvement from the 0.01% hit rates reported in early de novo binder campaigns (Cao et al., Nature 2022). RFdiffusion (Watson et al., Nature 2023), ProteinMPNN (Dauparas et al., Science 2022), and BoltzGen (Stark et al., bioRxiv 2025) have further expanded the design toolkit.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

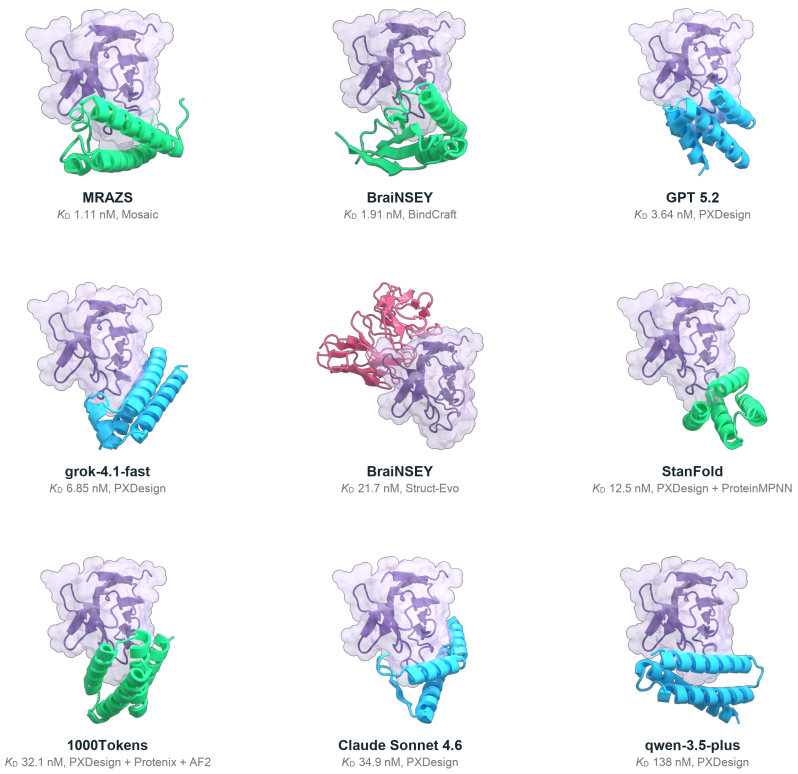


Figure 1: Selected TREM2 binders from the hackathon. Protenix-predicted complexes show TREM2 in lavender, human-team binders in green, agent-designed binders in cyan, and a reference-like scFv in pink; labels report team, mean SPR (K_D), and design method.

Community-driven competitions benchmark protein design under standardized experimental conditions. The Adaptyv EGFR and Nipah challenges and Bits to Binders have generated experimental datasets for evaluating design methods and computational scores (Cotet et al., 2025; Kosonocky et al., 2026; Overath et al., 2025). Prior competitions compared design *methods* — participants used different tools but the same human-driven workflow of target analysis, design generation, and filtering. A parallel trend in AI is the emergence of autonomous agents capable of end-to-end scientific workflows, from literature review to tool selection to design generation and quality control. This raises a fundamental question: can autonomous AI agents match or exceed human expert performance when given the same task and constraints? At the time of the event, no prior competition or experimentally-validated benchmark had tested this.

TREM2 (Triggering Receptor Expressed on Myeloid Cells 2) is a transmembrane receptor on microglia that regulates neuroinflammation and amyloid-beta plaque clearance. Loss-of-function variants (e.g., R47H) increase late-onset Alzheimer's risk 2–4x (Guerreiro et al., NEJM 2013). At the time of the event, no de novo mini-protein binder existed for TREM2, making it both a clinically relevant and scientifically interesting target for this comparison.

We report results from the muni x Adaptyv TREM2 Binder Design Hackathon (February 28, 2026, San Francisco): to our knowledge, the first controlled comparison of human teams and fully autonomous AI agents from different model providers under identical submission rules, with experimental validation. Participants designed de novo binders for TREM2's IgSF domain (residues 19–132, PDB: 5ELI). We compare computational scores, experimental outcomes, design methods and sequence diversity; examine predicted mini-protein contact footprints relative to the principal crystal sites of two characterized scFvs; and test confidence scores for binder discrimination and affinity ranking. Finally, we report ten additional designs from a post-hackathon campaign that iterated computational generation and evaluation before Bio-layer Interferometry (BLI) testing. Anthropic's

Claude Science study compared its TREM2 results against both the publicly released hackathon and agentic follow-up results (Claude Science & Shanehsazzadeh, 2026).

2 METHODS

2.1 COMPETITION DESIGN

Event format. Nine human teams (40 participants) and six autonomous AI agents competed in a single-day in-person hackathon co-organized by muni and Adaptyv Bio. Human teams used muni's platform alongside custom agentic or human-in-the-loop workflows. Autonomous agents operated without human intervention during the design phase and received the same prompt.

Submission constraints. Submissions were limited to ten single-chain protein sequences of at most 250 amino acids, using only standard amino acids. Designs were required to target the IgSF domain (residues 19–132), not the stalk region (133–172), and differ by at least ten amino acids from known binders (scFv-2, scFv-4, VHB937 CDRs).

2.2 COMPUTATIONAL SCORING AND DESIGN SELECTION

Submitted target–binder complexes were evaluated using Boltz-2 (Passaro et al., 2025) through Adaptyv's publicly released scoring pipeline. The primary ranking metric was ipSAE (interaction prediction Score from Aligned Errors; Dunbrack, 2025). The larger target-to-binder or binder-to-target score was used to select the 100 highest-scoring designs for experimental validation. Throughout, ipSAE refers to this Boltz-2-derived score unless another predictor or variant is named. Pre-experimental hackathon scores were kept separate from later retrospective predictions.

2.3 CELL-FREE EXPRESSION AND BINDING CHARACTERIZATION

Selected designs were submitted to Adaptyv Bio for cell-free protein expression and binding characterization by surface plasmon resonance (SPR), with additional bio-layer interferometry (BLI) measurements for a subset. Laboratory annotations distinguished non-expression from expressed non-binding outcomes. Per-design K_D values were arithmetic means of included fitted values, calculated separately by assay; missing fits were not imputed. Record-inclusion criteria and assay provenance are provided in Supplementary S3.

2.4 SEQUENCE DIVERSITY AND STATISTICAL ANALYSIS

Human–agent differences in fitted affinity were assessed using a two-sided Mann–Whitney U test on pK_D, defined as -log10(K_D in molar units). Hit fractions and expression outcomes were compared using two-sided Fisher's exact tests. Binder discrimination was evaluated using ROC-AUC among expressed, tested designs; score distributions were compared by binding outcome and cohort using two-sided Mann–Whitney U tests. Metric–affinity associations were assessed using Spearman correlation with pK_D among fitted binders. Design-level bootstrap intervals, Wilson proportion intervals and endpoint-specific Benjamini–Hochberg adjustments were used. Analyses were conditional on experimental selection, and design-level resampling did not account for dependence arising from shared teams or design lineages.

2.5 POST-HACKATHON AUTORESEARCH

A separate post-hackathon computational autoresearch campaign used NVIDIA's Proteina-Complexa framework to generate and evaluate TREM2 binder candidates (Didi et al., 2026). Over 15 in silico rounds, the agent varied targeted epitope, binder-length range, search algorithm and search parameters. Candidates were refolded with Boltz-2 and ranked by the larger directional default ipSAE score. Ten selected designs were tested separately by BLI, with mean $K_D \pm$ sample s.d. calculated from two fitted replicates per successfully loaded design. Detailed methods are provided in Supplementary Appendix A and S3–S7.

3 RESULTS

3.1 BOTH COHORTS PRODUCED SINGLE-DIGIT-NANOMOLAR BINDERS WITH SIMILAR OBSERVED HIT RATES

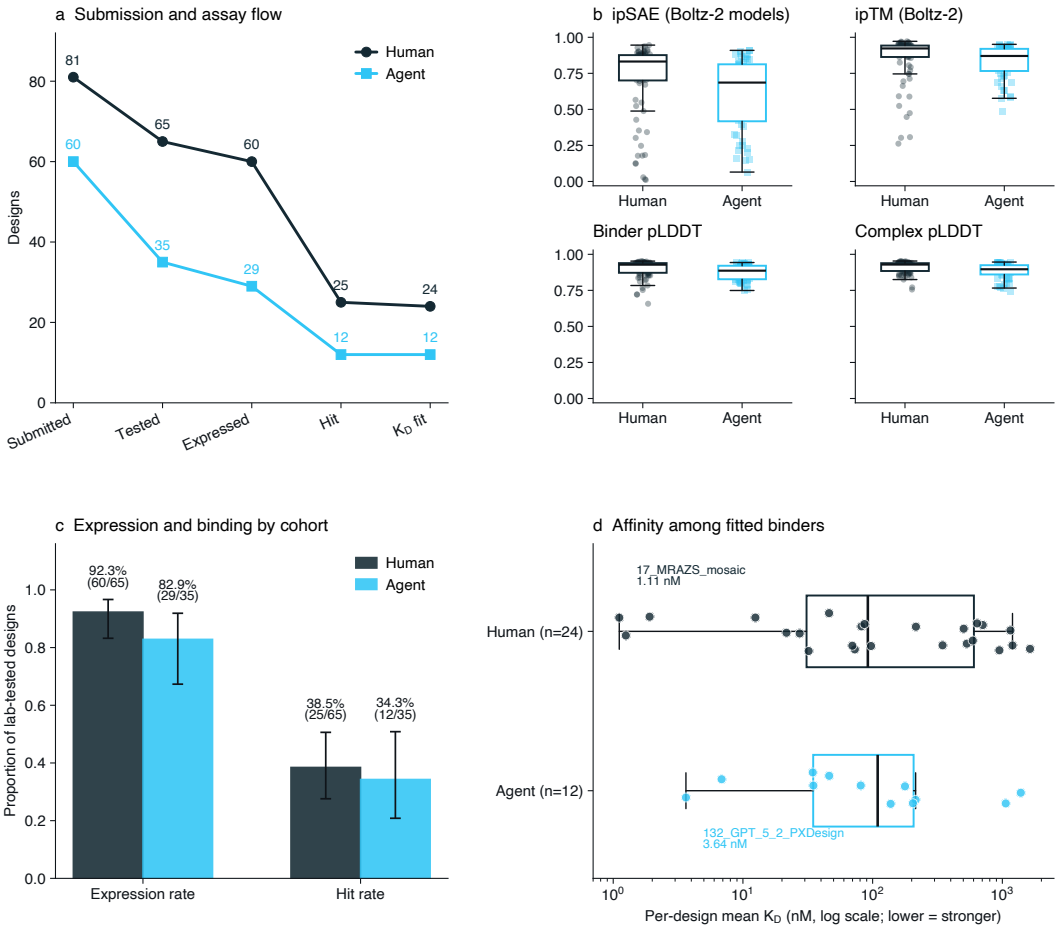


Figure 2: Submission flow, computational confidence scores and experimental outcomes. a, Flow of human-team and autonomous-agent submissions through experimental selection, expression, binding classification and fitted-affinity analysis. Untested designs have no assigned binding outcome. **b,** Distributions of the four computational confidence scores recorded with the submissions: ipSAE (Boltz-2 models), ipTM (Boltz-2), binder pLDDT and complex pLDDT. Points show individual designs and boxes show medians and interquartile ranges. **c,** Expression and binding-hit rates among experimentally tested designs (human, n = 65; agent, n = 35). Bars show observed proportions, error bars show Wilson 95% confidence intervals, and labels give percentages and counts. **d,** Canonical per-design mean K_D values for 24 human-team and 12 agent binders with fitted affinities, shown on a logarithmic nanomolar scale. Points show individual designs; boxes show medians and interquartile ranges, with whiskers extending to 1.5 times the interquartile range. Lower K_D indicates stronger binding; labels identify the strongest measured binder in each cohort. Assay provenance is reported in Supplementary S3.

Human submissions totaled 81 designs, and agent submissions totaled 60 designs; 100 were selected for experimental validation (Figure 2a).

Human-team submissions had higher values for ipSAE (Boltz-2 models; mean $\pm$ s.d., 0.717 $\pm$ 0.254 versus 0.605 $\pm$ 0.247; two-sided Mann–Whitney U test, p = 0.00085), ipTM (Boltz-2), binder pLDDT and complex pLDDT (Figure 2b; Supplementary S5).

Of the 65 human-team and 35 agent designs tested, 60 (92.3%) and 29 (82.9%) expressed, while 25 (38.5%) and 12 (34.3%) were classified as binding hits, respectively (Figure 2c). Fitted K_D values were not available for NVF design 5, which was classified as a weak binding hit; it was therefore included in hit-rate analysis but excluded from the affinity analysis (Supplementary S1).

The cohort difference in hit rates was not significant (two-sided Fisher p = 0.828; odds ratio 1.20). The estimated difference was 4.2 percentage points, but its 95% bootstrap interval was wide (-15.2 to 23.5); therefore, these data cannot show that one cohort performed better or prove that the true rates were equal. Among the 36 binders with fitted K_D values, estimated binding strengths also spanned similar ranges (two-sided Mann–Whitney U = 134, p = 0.750; Figure 2d). Details for cohort and team outcomes are reported in Supplementary S1.

The highest-affinity human design was MRAZS design 17 (K_D = 1.11 nM), and the highest-affinity agent design was GPT 5.2 design 132 (K_D = 3.64 nM) (Figure 1). The leaderboard and representative SPR sensorgrams are provided in the separate supplementary PDF (Appendix A, Table A1 and Figure A1).

Human-led teams used de novo, hallucination, optimization and diversification workflows, whereas all autonomous agents used PXDesign. Design-category definitions and counts are provided in Supplementary Appendix A (Table A2) and S1.

ESM-2 projections placed human and agent designs in overlapping regions, with several local team- and agent-associated groupings. Sequence-length distributions overlapped, with medians of 85 amino acids for human submissions and 80.5 for agent submissions; human-team submissions included the longest sequences (two-sided Mann–Whitney U = 2834, p = 0.0913; Supplementary S4).

3.2 Model confidence scores distinguish binders but weakly rank affinity

Among the expressed, tested designs, ipSAE (Boltz-2 models) showed modest binder discrimination (AUROC 0.652), whereas ipTM (Boltz-2) was weaker (0.604). In a retrospective analysis, Protenix predictions distinguished the outcomes more strongly: AUROC 0.782 for ipSAE and 0.791 for ipTM (Figure 3a,c). These were the highest AUROCs among 15 model scores, and their binder/non-binder score differences remained significant after adjustment across the 15 tests (q = 0.0000447 each).

Among the 36 fitted binders, Protenix ipSAE and ipTM had Spearman rho values of 0.293 and 0.196 with pK_D, respectively; neither passed the adjustment of 15 score-versus-affinity comparisons (q = 0.248 and 0.368; Figure 3d). ipSAE (Boltz-2 models) retained a modest positive association (rho = 0.397; adjusted q = 0.0655 in the four-original-score affinity family; Figure 3b,d). Thus, some model scores discriminated binders in this selected population, but model-confidence scores did not provide a reliable individual-affinity ranking. Full estimates, uncertainty intervals and sensitivity analyses are reported in Supplementary S5.

3.3 High-affinity mini-protein binding is predicted at a TREM2 surface distinct from two antibody epitopes

The structural analysis included 85 expressed, tested mini-protein designs (36 hits and 49 non-binders), excluding four antibody-family designs (Supplementary Appendix A).

The strongest hackathon binder, design 17, was predicted by both Protenix and Chai-1 to engage the TREM2 hydrophobic tip, outside the principal scFv-2 and scFv-4 crystal-contact sites. All 36 mini-protein hits contacted at least one hydrophobic-tip residue in the consensus predictions (Figure 4a,c,d). Tip contact was also predicted for 46/49 expressed mini-protein non-binders (Figure 4e,f), so predicted tip engagement alone did not distinguish binding outcomes.

None of the 36 hit consensus footprints overlaps the selected principal scFv-2 or scFv-4 crystal-contact sets (Szykowska et al., 2021; Figure 4a,b). This distinction does not extend to every antibody-associated crystal-packing contact. Footprint-size statistics, secondary crystal contacts and phosphatidylserine comparisons are reported in Supplementary Appendix A and S6.

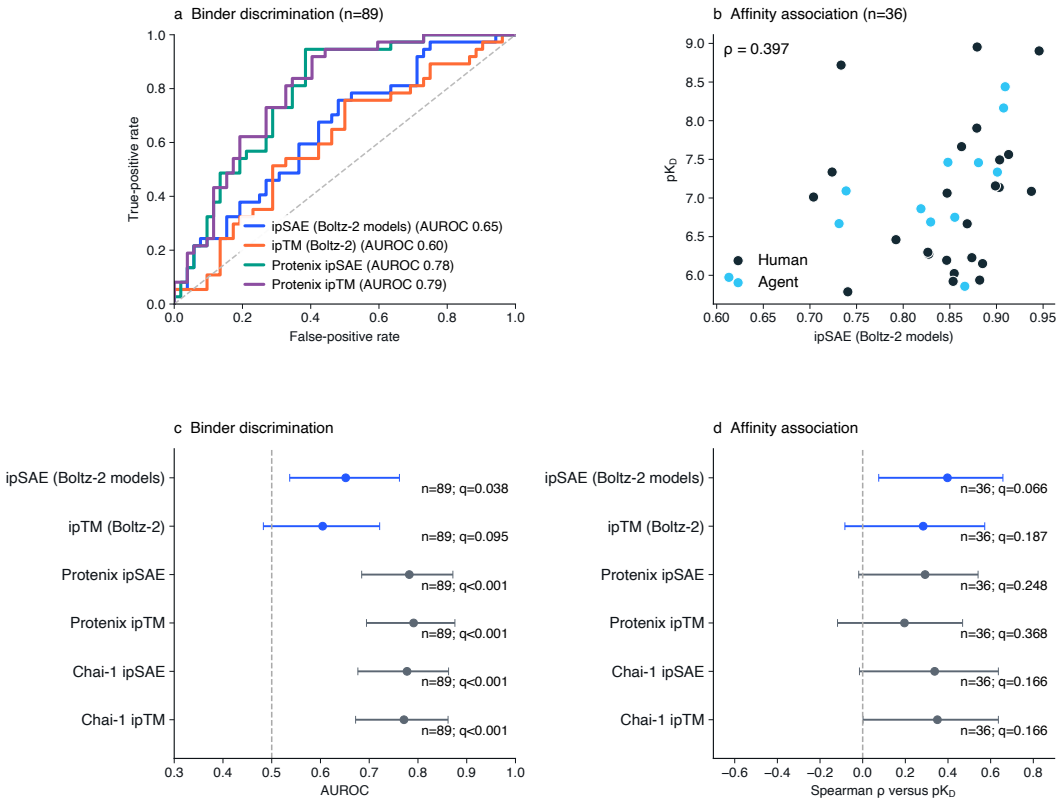


Figure 3: Binder discrimination and binder-only affinity association are distinct endpoints. **A**, ROC curves compare ipSAE and Boltz-2 ipTM with later Protenix ipSAE and ipTM among 89 experimentally tested designs that expressed (37 binders and 52 non-binders). **B**, ipSAE against canonical pK_D for the 36 binders with fitted affinities. **C**, ROC-AUC estimates and 95% design-level percentile-bootstrap intervals for selected submitted and later metrics. **D**, Binder-only Spearman correlations and paired design-bootstrap 95% intervals.

The hydrophobic tip has been implicated in TREM2 ligand binding and independent mini-protein predictions; these models support a predicted binding surface distinct from the two principal scFv sites, not discovery of a new epitope or evidence of ligand competition (Kober et al., 2016; Perera et al., 2026). The models omitted explicit glycans, so potential steric hindrance from TREM2 glycans in the predicted binding poses was not assessed.

3.4 Agentic computational search yielded three sub-nanomolar binders

Of the ten candidates tested by BLI, nine loaded successfully and had mean fitted K_D values below 10 nM; muni-AR-07 had no assay loading. Three designs—muni-AR-03, muni-AR-09 and muni-AR-04—had subnanomolar means of 0.232, 0.583 and 0.619 nM, respectively (Figure 5b). These were numerically below the three lowest eligible hackathon leaderboard means, although the campaigns used different assays (BLI versus SPR). Representative sensorgrams are shown in Figure 5c; full fitted values, replicate variability and computational rankings are reported in Supplementary S7.

4 Discussion

4.1 Mini-protein binders are predicted to engage a distinct TREM2 surface

Experimental structures, mutagenesis or competition assays are required to establish the binding mode and determine whether these designs alter ligand binding or TREM2 function.

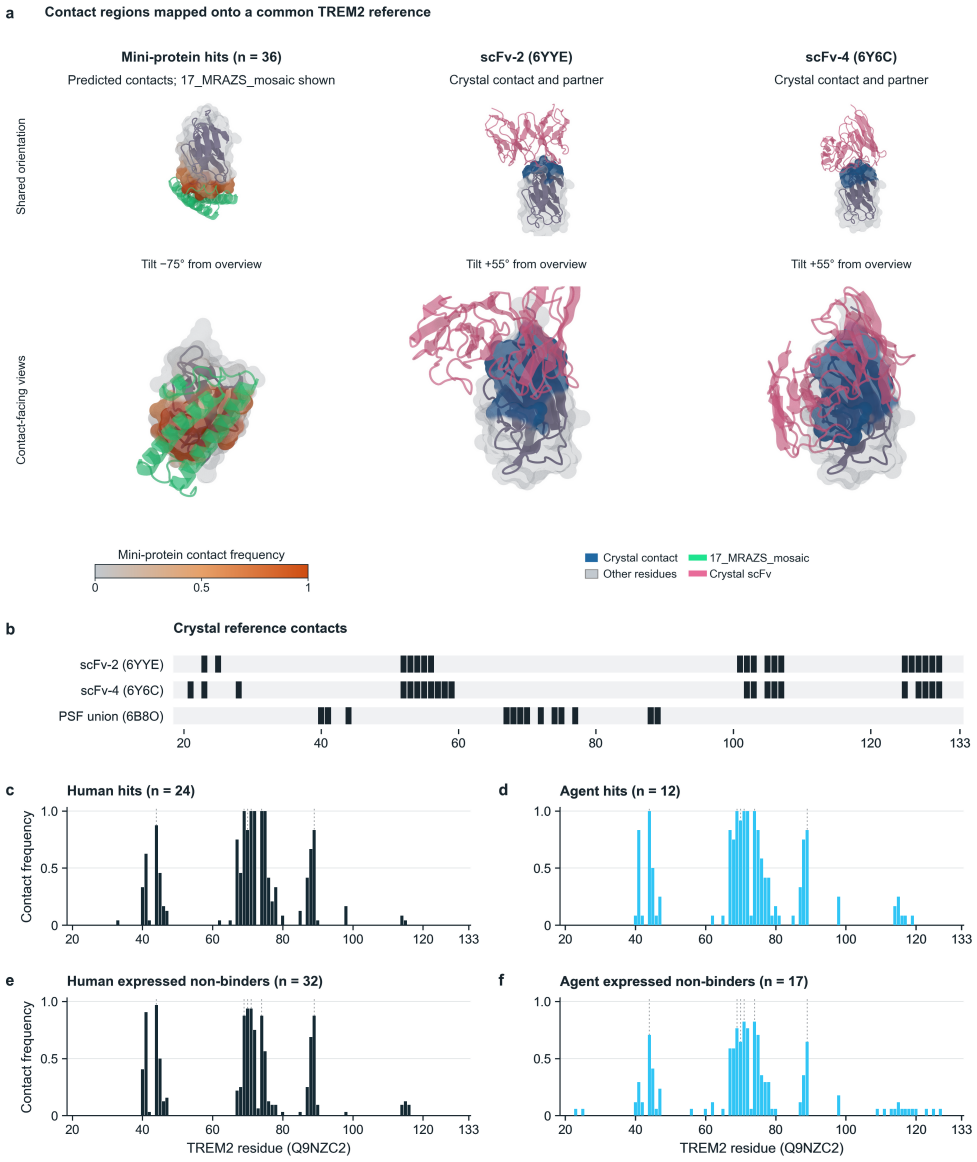


Figure 4: Predicted mini-protein contacts converge on a TREM2 surface distinct from two antibody epitopes. a, Residue maps on a common Protenix TREM2 reference. Small overview maps share one orientation; larger contact-facing maps tilt toward the mini-protein or antibody-contact regions. Gray-to-orange encodes predicted consensus contact frequency across 36 mini-protein hits, blue marks the principal scFv-2 (6YYE; 19 residues) or scFv-4 (6Y6C; 21 residues) crystal-contact masks, and gray marks other residues. The translucent surface is shown over a TREM2 cartoon to reveal the receptor fold; opacity does not encode data. These are residue maps, not observed simultaneous complexes. **b,** Crystal-contact tracks; black cells mark sequence-mapped TREM2 residues within 4.5 Å of the selected partner. Phosphatidylserine (6B8O) is represented by the 13-residue PSF-only contact union across crystallographic receptor/ligand copies. **c–f,** Predicted consensus contact frequencies for human hits (**c**, n = 24), agent hits (**d**, n = 12), human expressed non-binders (**e**, n = 32) and agent expressed non-binders (**f**, n = 17). Human cohorts are shown in navy and agent cohorts in cyan. Consensus requires heavy-atom separation $\leq$ 4.5 Å and target-residue burial >1 Å^2 independently in Protenix and Chai-1. Guides mark W44, L69, W70, L71, F74 and L89 in full-length UniProt numbering.

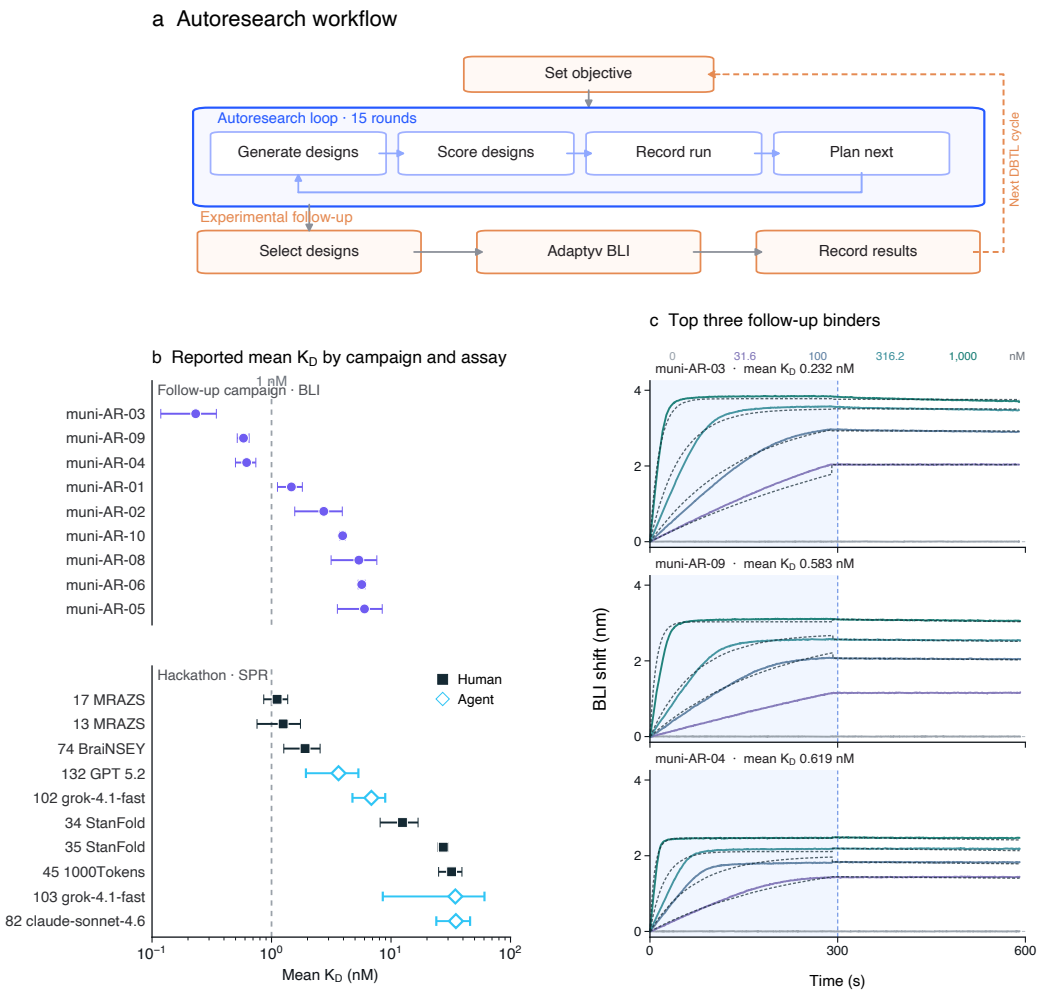


Figure 5: Autoresearch workflow and experimental follow-up. a, Autoresearch workflow. Fifteen computational rounds varied generation and search settings before Boltz-2 refolding, ipSAE scoring, and BLI selection. The return arrow denotes the outer design-build-test-learn cycle, where assay results influence the next objective. **b,** Mean $K_D \pm$ sample s.d. across the available included fitted replicates for nine measured autoresearch designs (BLI) and ten eligible hackathon designs (SPR), separated by campaign and assay on a common logarithmic scale. Replicate counts varied by design and assay; the measured follow-up designs each had two fitted replicates. Design 79 is excluded from this ranked display; muni-AR-07 had no assay loading. The dashed line marks 1 nM. **c,** Representative raw BLI concentration series and supplied fits for muni-AR-03 replicate 1, muni-AR-09 replicate 2 and muni-AR-04 replicate 1 at 0, 31.6, 100, 316.2 and 1,000 nM. The vertical dashed line marks the association-to-dissociation transition.

4.2 Model confidence scores support binder triage, not affinity ranking

Scores calculated from the Boltz-2 models also performed differently, showing that confidence metrics are not interchangeable across structure-prediction systems. These scores are therefore somewhat helpful for enriching a panel for likely binders, but are still imperfect proxies for ranking individual candidates. Reproducibility varies with hardware and random seeds, and directly optimizing for ranking metrics ("hill-climbing") risks producing designs that fail experimentally. Despite imperfect scoring functions, the complementary signals observed across confidence metrics and structural descriptors suggest that future selection strategies could use composite objective scoring functions to mitigate some biases and variance introduced with ipSAE-only scoring.

4.3 AGENTIC WORKFLOWS CAN EXTEND BINDER DESIGN

While the results demonstrate that agents can also complete a full design workflow within a short deadline, they do not establish agent superiority. We would need a larger sample size and a longer campaign that would involve multiple design-build-test-learn loops. Regarding agent failure modes, we found little evidence of agent-specific rule failures (Supplementary Appendix A and S6). The final designs did not indicate obvious misunderstanding of the target structure or challenge constraints, but they also do not reveal whether the agents made and corrected errors during the design process.

All agents also used PXDesign for binder design; in contrast, human-led teams used a broader selection of generative, hallucination, and optimization workflows. Because tool choice was confounded with team and only selected candidates were tested, the per-method hit counts do not establish that any individual design method was better than another. Detailed prompts, method-specific skill files, or explicit instructions could help to broaden the workflows agents use; whether this increases candidate diversity or experimental hit rates is yet to be determined.

The subsequent campaign demonstrates the potential of agentic systems to sustain protein design projects beyond a time-limited challenge, extending researchers' capacities to explore a wide search space while leaving responsibility for experimental objectives and biological interpretation for human direction. Reliable deployment requires agents to integrate deeper domain knowledge, recover from failures, and interpret constraints reproducibly.

4.4 LIMITATIONS AND OUTLOOK

Our analysis includes 100 selected submissions, with small subgroups and related designs from shared teams and toolchains. Incomplete records of compute use and design decisions limit our ability to separate the contributions of human expertise, agent sampling and error recovery. The proposed binding epitope lacks structural validation, and affinities were measured using different assays across campaigns. TREM2 has N-linked glycosylation sites at N20 and N79 (Shirotani et al., 2022). Our Protenix and Chai-1 models omitted explicit glycans, leaving potential steric hindrance that could prevent binding despite a favorable predicted interface unassessed. Binding to the recombinant assay target does not establish recognition of native cell-surface TREM2. For the assays conducted, binding alone does not establish specificity, stability, function or therapeutic value. There are promising directions for future investigation to test whether agents can learn from both successful and unsuccessful designs. Such studies would establish a basis for adaptive discovery campaigns where agents use compounding data — beyond binding data — to iteratively revise their hypotheses.

REPRODUCIBILITY STATEMENT

Experimental and computational procedures are described in Methods and Supplementary Appendix A; supporting analyses and records are provided in Supplementary S1–S7 and Supplementary Data S1–S7. Analysis code and saved inputs are included in the supplementary materials. Binding affinities, expression data and raw BLI sensorgrams are publicly available on Proteinbase for the hackathon and autoresearch campaign. Submitted designs and competition details are on the competition page. Supporting result tables are provided in the supplementary materials; the ipSAE scoring pipeline is available in the Adaptyv repository.

AI USAGE STATEMENT

GPT-Astra was used to streamline and revise writing, check consistency across the manuscript and supplementary materials, retrieve and verify references, assist with data formatting and supporting code, review interpretation of results, and revise selected figures. The corresponding author reviewed and approved any proposed revisions. Experimental measurements came from our laboratory records. The content, the scientific claims, the figure content and interpretation, and the conclusions are ours.

REFERENCES

Adaptyv Bio. nipah_ipsae_pipeline. GitHub repository, 2026a. URL https://github.com/adaptyvbio/nipah_ipsae_pipeline. Accessed 17 September 2026.

Adaptyv Bio. Nipah competition results. Proteinbase, 2026b. URL https://proteinbase.com/collections/nipah-binder-competition-results. Accessed 17 September 2026.

Nick Boyd, Sam Guns, and Escalante Bio. Mosaic: Composite-objective protein design. GitHub repository, 2025. URL https://github.com/escalante-bio/mosaic. Accessed 17 September 2026.

Verena Brand et al. TREM2 stabilizing antibodies. International patent publication WO 2020/079580 A1, 2020. URL https://patents.google.com/patent/WO2020079580A1/en. Published 23 April 2020.

ByteDance AML AI4Science Team et al. Protenix—advancing structure prediction through a comprehensive AlphaFold3 reproduction. *bioRxiv*, 2025. doi: 10.1101/2025.01.08.631967. URL https://doi.org/10.1101/2025.01.08.631967. Preprint.

Salvatore Candido et al. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735. URL https://doi.org/10.64898/2026.06.03.729735. Preprint.

L. Cao et al. Design of protein-binding proteins from the target structure alone. *Nature*, 605:551–560, 2022. doi: 10.1038/s41586-022-04654-9. URL https://doi.org/10.1038/s41586-022-04654-9.

Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955. URL https://doi.org/10.1101/2024.10.10.615955. Preprint.

Claude Science and Amir Shanehsazzadeh. Autonomous de novo protein binder design with Claude. Technical report, Anthropic, August 2026. URL https://www-cdn.anthropic.com/30bf50e22a01388bb29bf077ee3f244531594b7a.pdf. 18 August 2026.

Peter J. A. Cock et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25:1422–1423, 2009. doi: 10.1093/bioinformatics/btp163. URL https://doi.org/10.1093/bioinformatics/btp163.

Tudor-Stefan Cotet et al. Crowdsourced protein design: Lessons from the adaptyv EGFR binder competition. *bioRxiv*, 2025. doi: 10.1101/2025.04.17.648362. URL https://doi.org/10.1101/2025.04.17.648362. Preprint.

Justas Dauparas et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378:49–56, 2022. doi: 10.1126/science.add2187. URL https://doi.org/10.1126/science.add2187.

Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael M. Bronstein, Martin Steinegger, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Scaling atomistic protein binder design with generative pretraining and test-time compute. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qmCpJtFZra. Oral presentation.

Roland L. Dunbrack, Jr. Rēs ipSAE loquuntur: What's wrong with AlphaFold's ipTM score and how to fix it. *bioRxiv*, 2025. doi: 10.1101/2025.02.10.637595. URL https://doi.org/10.1101/2025.02.10.637595. Preprint.

Richard Evans et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021. doi: 10.1101/2021.10.04.463034. URL https://doi.org/10.1101/2021.10.04.463034. Preprint.

Dominik Feuerbach et al. Methods of treatment using anti-TREM2 antibodies. International patent publication WO 2026/062582 A1, 2026. URL https://patents.google.com/patent/WO2026062582A1/en. Published 26 March 2026.

A. Graf, D. Feuerbach, D. Shimshek, J. Sovago, G. Peraus, A. Zhu, C. Agoropoulou, A. Bobirca, and C. Serban. VHB937, a TREM2 stabilizing and activating fully human monoclonal antibody: A phase 2 study design in Alzheimer's disease and rationale. *The Journal of Prevention of Alzheimer's Disease*, 12:88–89, 2025. URL https://secure.key4events.com/key4register/images/client/1892/files/CTAD25%20Abstract%20Book.pdf. 18th Clinical Trials on Alzheimer's Disease (CTAD), abstract P023. San Diego, 1–4 December 2025.

Rita Guerreiro et al. TREM2 variants in Alzheimer's disease. *New England Journal of Medicine*, 368:117–127, 2013. doi: 10.1056/NEJMoa1211851. URL https://doi.org/10.1056/NEJMoa1211851.

Daniel L. Kober et al. Neurodegenerative disease mutations in TREM2 reveal a functional surface and distinct loss-of-function mechanisms. *eLife*, 5:e20391, 2016. doi: 10.7554/eLife.20391. URL https://doi.org/10.7554/eLife.20391. Primary publication for PDB 5ELI.

Clayton W. Kosonocky et al. Validation and analysis of 12,000 AI-driven CAR-T designs in the bits to binders competition. *bioRxiv*, 2026. doi: 10.64898/2026.03.03.709355. URL https://doi.org/10.64898/2026.03.03.709355. Preprint.

Zeming Lin et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://doi.org/10.1126/science.ade2574.

H. Long et al. Preclinical and first-in-human evaluation of AL002, a novel TREM2 agonistic antibody for Alzheimer's disease. *Alzheimer's Research and Therapy*, 16:235, 2024. doi: 10.1186/s13195-024-01599-1. URL https://doi.org/10.1186/s13195-024-01599-1.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3:861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

C. J. Mummery et al. The TREM2 agonistic antibody AL002 in early Alzheimer's disease: a phase 2 randomized trial. *Nature Medicine*, 32:1708–1716, 2026. doi: 10.1038/s41591-026-04273-1. URL https://doi.org/10.1038/s41591-026-04273-1.

Novartis. A clinical trial to learn about the effects of VHB937 in people with early Alzheimer's disease. Clinical trial listing, NCT07094516, 2026. URL https://www.novartis.com/clinicaltrials/study/nct07094516. Updated 26 June 2026. Accessed 25 September 2026.

Novartis AG. Q2 2026 results: Investor presentation, July 2026. URL https://www.novartis.com/sites/novartis_com/files/q2-2026-investor-presentation.pdf. 21 July 2026, slide 68. Accessed 17 September 2026.

Max D. Overath, Andreas S. H. Rygaard, Christian P. Jacobsen, Valentas Brasas, Oliver Morell, Pietro Sormanni, and Timothy P. Jenkins. Predicting experimental success in de novo binder design: A meta-analysis of 3,766 experimentally characterised binders. *bioRxiv*, 2025. doi: 10.1101/2025.08.14.670059. URL https://doi.org/10.1101/2025.08.14.670059. Preprint.

Martin Pacesa et al. One-shot design of functional protein binders with BindCraft. *Nature*, 646:483–492, 2025. doi: 10.1038/s41586-025-09429-6. URL https://doi.org/10.1038/s41586-025-09429-6.

Saro Passaro et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707. URL https://doi.org/10.1101/2025.06.14.659707. Preprint.

Doniesha Perera, Emmanuel Ajiboye, Kalana Pitakatuwana, Sophia Wier, Vincent Duong, and Haifan Wu. XL-MS and de novo protein design identified a common motif for TREM2 binding. *bioRxiv*, 2026. doi: 10.64898/2026.04.23.720433. URL https://doi.org/10.64898/2026.04.23.720433. Preprint.

Protenix Team et al. PXDesign: fast, modular, and accurate de novo design of protein binders. *bioRxiv*, 2025. doi: 10.1101/2025.08.15.670450. URL https://doi.org/10.1101/2025.08.15.670450. Preprint.

Keiro Shirotani, Daisuke Hatta, Naoki Wakita, Kaori Watanabe, and Nobuhisa Iwata. The role of TREM2 N-glycans in trafficking to the cell surface and signal transduction of TREM2. *Journal of Biochemistry*, 172(6):347–353, 2022. doi: 10.1093/jb/mvac073. URL https://doi.org/10.1093/jb/mvac073.

A. Shrake and J. A. Rupley. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of Molecular Biology*, 79:351–371, 1973. doi: 10.1016/0022-2836(73)90011-9. URL https://doi.org/10.1016/0022-2836%2873%2990011-9.

Hannes Stark et al. BoltzGen: Toward universal binder design. *bioRxiv*, 2025. doi: 10.1101/2025.11.20.689494. URL https://doi.org/10.1101/2025.11.20.689494. Preprint.

Athena Sudom et al. Molecular basis for the loss-of-function effects of the Alzheimer's disease-associated R47H variant of the immune receptor TREM2. *Journal of Biological Chemistry*, 293:12634–12646, 2018. doi: 10.1074/jbc.RA118.002352. URL https://doi.org/10.1074/jbc.RA118.002352.

Aleksandra Szykowska et al. Selection and structural characterization of anti-TREM2 scFvs that reduce levels of shed ectodomain. *Structure*, 29:1241–1252.e5, 2021. doi: 10.1016/j.str.2021.06.010. URL https://doi.org/10.1016/j.str.2021.06.010.

Michel van Kempen et al. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42:243–246, 2024. doi: 10.1038/s41587-023-01773-0. URL https://doi.org/10.1038/s41587-023-01773-0.

Joseph L. Watson et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620:1089–1100, 2023. doi: 10.1038/s41586-023-06415-8. URL https://doi.org/10.1038/s41586-023-06415-8.

Katherine Yenko. Autoresearch and muni CLI. muni Research, 2026a. URL https://www.muni.bio/research/muni-makeover-cli-autoresearch. Campaign report; code at https://github.com/bioArena/autoresearch_trem2.

Katherine Yenko. Finding three sub-nanomolar binders, agentically. muni Research, 2026b. URL https://www.muni.bio/research/closing-the-autoresearch-loop. Experimental follow-up report.