Autonomous *de novo* protein binder design with Claude

Claude Science[1] Amir Shanehsazzadeh[1,*]

August 18, 2026

[1]Anthropic, San Francisco, CA, USA

*Correspondence: ashanehsazzadeh@anthropic.com

Abstract

Deep-learning methods for protein structure generation, sequence design, and structure prediction now enable *de novo* binder design against many targets using only tens of designs. A design campaign nonetheless demands expertise that spans target biology, structural modeling, and a fast-moving set of computational tools, as well as days of orchestrating software and compute. We asked how much of this expertise and labor an AI agent could supply. We wrote the working knowledge of a binder design campaign into a single protocol prompt that specifies no epitope, scaffold, or sequence for any target. Working from it, and without human input into any design decision, Claude Opus 4.8 and Mythos Preview ran 24- to 48-hour campaigns against 16 targets. They researched each target, chose epitopes, installed and ran open-source protein design and structure-prediction models, optimized their candidates *in silico*, and delivered 30 ranked designs per target. Two independent contract research organizations synthesized every design exactly as delivered and measured its binding, with 15 of the 16 targets giving interpretable measurements. Claude designed binders against 14 of them, and 354 of 1,320 designs bound, a hit rate of 27%; among the designs ranked first for each target in each campaign, 49% bound. Designing against all targets at once in a single 48-hour session, Mythos Preview and Opus 4.8 achieved hit rates of 26.7% and 22.6%; designing against one target at a time in 24-hour sessions, Mythos Preview's hit rate was 35.1%. On the E3 ligase subunit RBX1, recently the subject of an open design competition in which 9 of 245 *de novo* designs bound, 28 of Claude's 90 designs bound. The tightest bound with a K_D of 3.9 nM, against 45 nM for the competition's winning entry re-synthesized and measured on the same plate. Although cross-species reactivity was only a secondary objective of the prompt, 130 of the 233 binders tested against the mouse ortholog of their target also bound it. Every model Claude used is open-source, which places campaigns of this kind within reach of any laboratory. We release the prompts, computational models of all 1,440 designs, and the binding data for the 1,320 designs with reliable measurements as a reproducible protocol for autonomous binder design and a benchmark dataset for the field.[1]

Introduction

There has been considerable progress in *de novo* protein binder design. Deep-learning methods now generate binder structures in the context of the target [1–15], design sequences to fold into them [16,17], and, through structure prediction of the designed complex [18–20], identify the designs most likely to bind before any are synthesized [21,22]. Early campaigns screened 10^4 to 10^5 designs to find binders [23,24], while recent methods find them among tens to hundreds [1,3,14]. Despite this, a protein design campaign still requires a significant number of expert decisions: which region of the target to model (the target construct) and which surface on it to bind, which methods and settings to use, how stringently to filter, whether redesign is worth its compute, and which designs to order. Furthermore, orchestrating the specialized models for protein design and structure prediction, together with the compute resources they require, remains a significant bottleneck and a barrier to entry.

[1]Available at https://huggingface.co/datasets/Anthropic/claude-protein-binder-design.

AI agents built on large language models have begun to plan and carry out experiments in chemistry and biology [25–28], and they have more recently been applied to protein binder design, both in pipelines that propose or rank designs and in agent-assisted design efforts whose products were tested experimentally [29–36]. We hypothesized that such agents could supply a significant portion of the expertise and labor that a binder design campaign requires, and we set out to determine how successfully Claude could autonomously carry out protein design campaigns against multiple targets. To test this, we wrote the working knowledge of a campaign into a single protocol prompt. The prompt covers the stages of a campaign, the open-source tools for each stage, and the criteria for selecting designs to test. It does not specify an epitope, scaffold, target construct, or sequence for any target. We acted only at the beginning and the end of the process: we chose the targets, giving each to Claude only by name, UniProt accession, organism, and oligomeric state, provided a cloud GPU account with a fixed budget and time limit, placed the synthesis orders, and interpreted the binding data. Everything in between, from the choice of target region and epitope to the final ranking, was left to Claude, which scored its candidates with an ensemble of co-folding predictors that we had validated beforehand on a public benchmark [22] (Methods) and closed each campaign by returning 30 ranked designs per target with a record of its decisions.

Claude Opus 4.8 and Mythos Preview ran campaigns in two formats. In multi-target campaigns lasting 48 hours, each model designed against the same 14 targets at once. In single-target campaigns lasting 24 hours, Mythos Preview designed against every target, and Opus 4.8 against latent GDF-8, mature GDF-8, and TNFα. Two independent contract research organizations (CROs), Adaptyv Bio and Twist Bioscience, synthesized every design exactly as delivered and characterized it for binding. One of the 16 targets, the mature GDF-8 dimer, aggregated under assay conditions and gave no interpretable measurement at either CRO, so we exclude its 120 designs from all analyses, which leaves 1,320 designs on 15 targets.

We report the outcome for every delivered design, including those from failed campaigns, together with the prompts, the design models, and both CROs' measurements.[2]

Methods

Ensembling ESMFold2 and Protenix v2 scores produces state-of-the-art filtering for *de novo* protein binder designs

Claude Science was used to benchmark ESMFold2 and ESMFold2-Fast [37,38] and Protenix v2 [39,40] as filters for *de novo* binder designs on the Overath et al. [22] dataset (3,532 designs against 13 targets, 391 experimentally confirmed binders). Each model was run with five seeds and one sample per seed, and designs were scored by $\text{ipSAE}_{\text{min}}$ [41], taking the maximum over seeds, which ranked binders best or near-best among the confidence metrics we tested for every model. By macro-AP (average precision within each target, averaged over targets), ESMFold2-Fast (0.62) and ESMFold2 (0.61) each exceeded the strongest predictor reported by Overath et al. (AlphaFold3 [20], 0.55), and a per-target z-score ensemble of the three models was significantly more predictive still (0.66; +0.11 over AlphaFold3, 95% CI 0.02 to 0.26, and +0.04 over ESMFold2-Fast alone, 0.02 to 0.07; Figure M1). We added self-consistency DockQ (sc-DockQ), the DockQ [42] between each model's predicted complex and the designed complex, as three further z-scored terms, one per model, weighted 4:1 in favor of $\text{ipSAE}_{\text{min}}$. On its own sc-DockQ is a weaker filter (0.45), and at this weight it leaves discrimination unchanged (0.65), so the ranking score measures co-folding confidence first and agreement with the designed pose second. The campaign protocol instructs Claude to rank its designs by this score, computed from one seed per model during screening and from five seeds for the final ranking.

A single protocol prompt guides Claude on how to run a protein binder design campaign

Claude ran every campaign autonomously from one protocol prompt (about 16,000 words), loaded as the system prompt of every agent, together with a short kickoff message. The prompt sets out what an expert designer knows about running such a campaign: its stages, the tools for each stage, and the criteria a design must meet to be selected for testing. It leaves the decisions to Claude. In each campaign, Claude:

- researched each target's biology and available structures;

[2] Available at https://huggingface.co/datasets/Anthropic/claude-protein-binder-design.

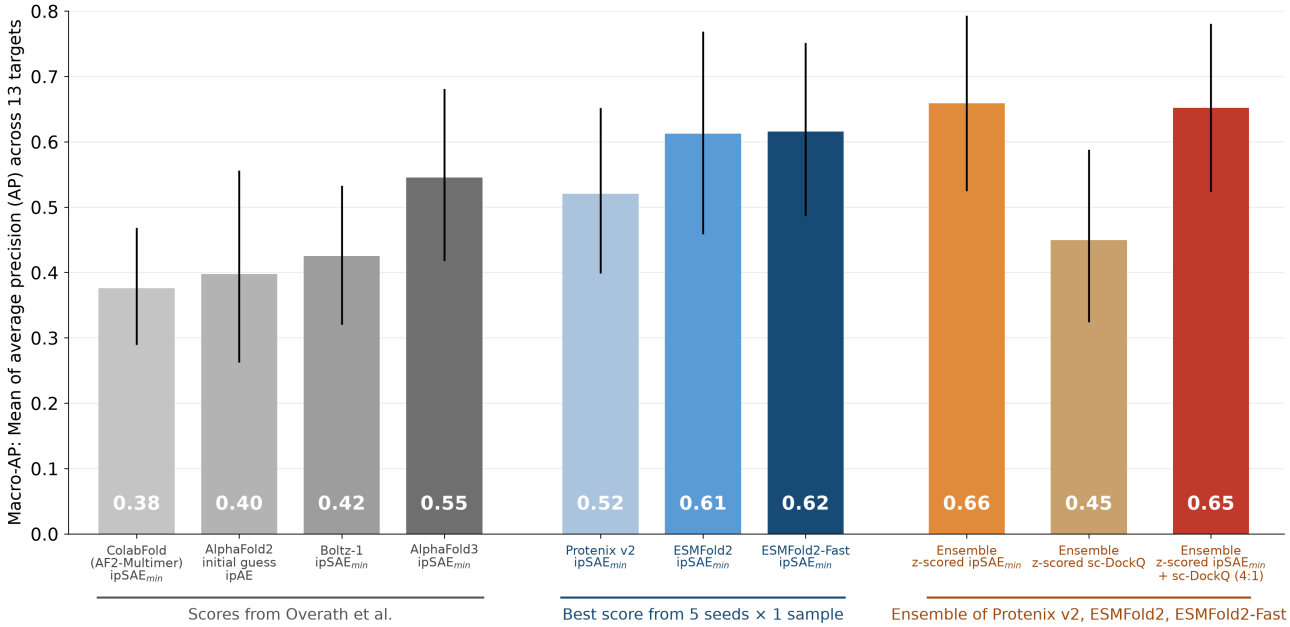


Figure M1. Ensembling ESMFold2 and Protenix v2 scores on the Overath et al. dataset. Macro-averaged precision (macro-AP) of each score for separating the 391 binders from the non-binders among the 3,532 designs of the dataset: average precision is computed within each of the 13 single-chain targets and averaged over targets (whiskers, 95% bootstrap intervals over targets). Left, the four predictors highlighted by Overath et al. [22], scored with the values released with the dataset: ColabFold [43] AlphaFold2-Multimer [19], Boltz-1 [44], and AlphaFold3 [20] (one seed, best of three samples) by ipSAE$_{min}$ [41], and AlphaFold2 initial guess [21] by ipAE. Middle, the three predictors used in this work, Protenix v2 [40], ESMFold2 [37,38], and ESMFold2-Fast, each run five times and scored by the ipSAE$_{min}$ of its best run. Right, ensembles of those three, formed by z-scoring each score within a target and averaging: ipSAE$_{min}$ alone (orange); self-consistency DockQ against the designed complex (sc-DockQ [42]) alone (tan); and ipSAE$_{min}$ with the sc-DockQ terms added at one-quarter weight (red), the ranking score prescribed in the campaign protocol. The sc-DockQ terms do not change discrimination and are kept as a check that the predicted pose is the designed one.

- chose the region of the target to model and the epitopes to pursue;
- decided which public backbone-generation and sequence-design tools to use, and in what proportions;
- removed candidates that resembled known proteins, duplicated one another, or had problematic sequence composition before spending compute on scoring;
- checked that the ranking score (Figure M1) recognized known binders of each target, then scored candidates with one seed to screen and five seeds to rank;
- ran further rounds of *in silico* optimization on its top-scoring candidates when it expected them to improve the set;
- selected and ranked the 30 designs per target, which were synthesized exactly as delivered, and kept a complete record of its decisions.

About a third of the prompt is scientific guidance (Figure M2). It states the objective of a campaign: for each target, 30 novel single-chain miniproteins of 50 to 120 residues, to be synthesized and measured for binding at Adaptyv Bio and Twist Bioscience. It specifies each target by name, UniProt accession [45], organism, oligomeric state, and any obligate cofactor or bound partner (the sgRNA of SpCas9, the prodomain of latent GDF-8); several single-target prompts also name the extracellular domain. It directs Claude toward biologically relevant surfaces, preferably ones where a bound miniprotein could plausibly affect function, and allows a novel epitope where Claude can justify one. It sets two secondary objectives, each subordinate to affinity and hit rate on the primary antigen (generally the human protein): cross-reactivity with the cynomolgus, mouse, and rat orthologs, and a share of designs that are not all-α. Finally, it provides a reading list and an accompanying document corpus: Adaptyv Bio's benchmark and competition reports [46–53] and case studies [30,31,54], the ProteinBase pages and result collections for its competitions [55–62] and for published design methods [63–66], and the AlphaProteo, Cao et al., RFdiffusion, BindCraft, and BoltzGen papers [1,3,5,14,24,67]. The corpus also holds the method paper of every tool the prompt names and a set of published binder sequences from ProteinBase that was used, with UniRef90 [68], in the novelty screen.

The remaining two thirds of the prompt are what makes autonomy work at this scale: how to pace a fixed GPU budget against the wall clock, how to delegate work to a two-layer team of sub-agents and supervise it, how to verify results before reporting them, and when to report. Those sections were iterated over test campaigns until Claude reliably sustained 24- to 48-hour campaigns and used the full budget as intended, and were then frozen before the campaigns reported here. The single-target prompt differs from the multi-target prompt mainly in resources and scope, with its pacing parameters (screening volume, minimum optimization rounds, concurrency ceilings) scaled to match: 24 hours and a compute budget of USD $10,000 for a single-target campaign instead of 48 hours and USD $50,000 for a multi-target campaign, spent on cloud GPUs through Modal,[3] with 30 designs per target in both. All prompts and kickoff messages are made publicly available.

Setup and launch. A single operator prepared Claude Science before each campaign by providing the following: a funded cloud GPU account (Modal) with a per-campaign container limit; full connectors to Slack, Google Drive, Gmail, and bioRxiv; an allowlist of network domains; the credentials the tools would need; and the offline reference corpus. No design or prediction tool was pre-installed; the prompt has Claude build each one from its public repository into a container image and validate it in the first hour. The operator then loaded the protocol as the system prompt, sent the kickoff message (the targets are named in the prompt itself), and let the campaign run. Two multi-target campaigns (14 targets each, one per model, Claude Opus 4.8 and Claude Mythos Preview) and the single-target campaigns (Claude Mythos Preview on each of the 16 targets; Claude Opus 4.8 on TNFα, latent GDF-8, and mature GDF-8) ran this way. Claude Opus 4.8 was run at the maximum reasoning-effort ("thinking") setting and Claude Mythos Preview at the high setting. During a campaign the operator watched dashboards and account limits and approved access prompts that carried no scientific content. The only human messages that entered any campaign were short, non-technical instructions to resume after a session had died for infrastructure reasons; Claude otherwise detected and worked around infrastructure incidents on its own. When a campaign closed, the 30 ranked designs it returned for each target were ordered for expression and screening without modification. For TNFα, in addition to the 30 designs delivered by the Opus 4.8 single-target campaign, we ordered a supplementary set of 30 designs from the same campaign's scored pool. We selected these after the campaign had closed, using a rule fixed in advance on the campaign's own co-folding scores. Claude designed and scored all 30 but did not select or rank them, so they are included in the TNFα totals and hit rates but excluded from the analyses of Claude's ranking. Orders were placed with sequences shuffled and with no indication of model, campaign, or rank.

Re-scoring. Confidence values recorded during a campaign came from whichever predictors, settings, and target constructs Claude chose at the time, so they may not be comparable across campaigns or targets. After the campaigns closed, and independently of the wet lab work, we re-scored every ordered design under one protocol with ten publicly available co-folding predictors: Protenix v2 [39,40], ESMFold2 and ESMFold2-Fast [37,38], AlphaFold-Multimer v3 [19,43], Boltz-2 [69], Chai-1 [70], OpenFold3 [71], RoseTTAFold3 [72], OpenDDE [73], and the AlphaFold3 inference code with OpenFold3 weights [74]. Each was run five times per design, without templates, on one shared target construct per target, with an unpaired multiple-sequence alignment for the target (none for the single-sequence ESMFold2-Fast) and the design as a single sequence; AlphaFold-Multimer does not model nucleic acids and was not run on the SpCas9 ribonucleoprotein. For each design and predictor we keep the run with the highest $\text{ipSAE}_{\text{min}}$ and record that value and, where a design-time model of the complex exists, the self-consistency DockQ against it. Every *in silico* value in the Results that is not labeled design-time, including all of Figure 4, comes from this re-score. No experimental structure of any design or complex was determined, so every pose in this report is a prediction.

Open-source design and structure-prediction models used in the campaigns

Every design and prediction model available to Claude was open-source. The protocol prompt names candidate tools by category, all pre-cleared for license and installability, and requires a set of seven designated structure-generation methods to contribute at least 50 backbones each to every target's scored pool; beyond that, Claude chose among them for each target. The counts below are taken from the released per-design provenance, which records the tools that produced each of the 1,315 tested designs.

Structure generation and co-design. Ten generators contributed designs that were ordered: PXDesign [7] (358 designs), RFdiffusion3 [2] (267), Genie 3 [8] (185), FreeBindCraft [4] (135), a PyRosetta-free build of BindCraft [3], BoltzGen [5]

[3] Modal Labs, https://modal.com.

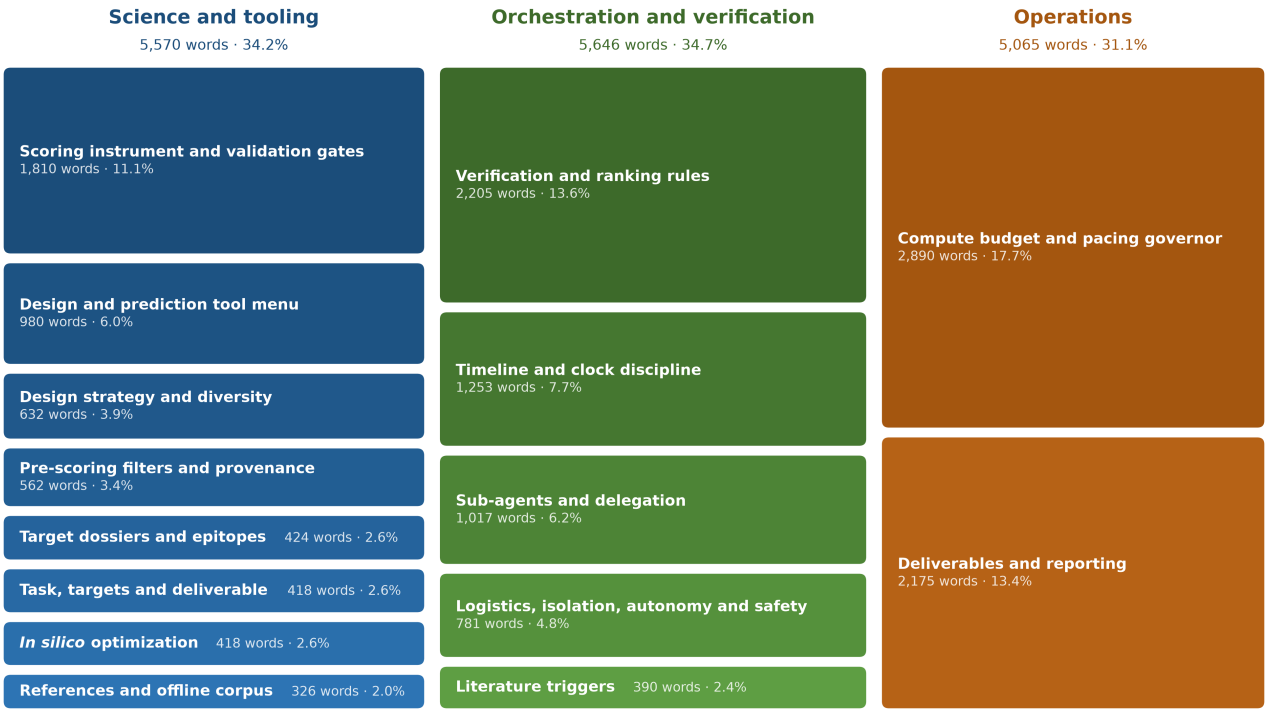


Figure M2. Anatomy of the campaign protocol prompt. The multi-target protocol prompt (about 16,000 words), which every Claude agent in a campaign receives as its system prompt, divided into 15 thematic blocks; tile areas are proportional to word counts, and percentages are of the whole prompt. Science and tooling (blue, 34.2%) is the working knowledge Claude designs from: target dossiers and epitope selection, design tools and strategy, pre-scoring filters, the ranking score of Figure M1, *in silico* optimization, and the 30-design deliverable. Orchestration and verification (green, 34.7%) and operations (orange, 31.1%) are what let Claude sustain a 24- to 48-hour campaign on its own: sub-agent delegation, clock discipline, verification and ranking rules, the compute budget and its pacing governor, and reporting. The single-target prompts differ mainly in resources and scope.

(134), RFdiffusion [1] (118), Proteina-Complexa [9,11] (100), FoldCraft [12] (14), BoltzDesign1 [6] (2), and Protein Hunter [13] (2). The prompt also named Mosaic [75] and HalluDesign [76], which did not contribute designs that were ordered. *In silico* optimization used partial diffusion with sequence redesign [77], inverse-folding resampling, predict-and-redesign cycling against a co-folding model, and point mutagenesis.

Sequence design. SolubleMPNN, the soluble variant of ProteinMPNN [16,78], designed the sequences of 1,133 tested designs, ProteinMPNN with default weights designed 21, and SolubleCaliby, the soluble variant of Caliby [17], designed 111. The remaining 50 sequences were produced jointly with the backbone by the generator (FreeBindCraft, Proteina-Complexa, or BoltzGen). Sequence plausibility was scored with ESM Cambrian (ESMC) log-likelihoods [79], and secondary structure was assigned with DSSP [80].

Structure prediction and scoring. The three predictors in the ranking score were ESMFold2, ESMFold2-Fast [37,38], and Protenix v2 [39,40]. As substitutes the prompt allowed AlphaFold-Multimer v3 [18,19] through ColabFold [43], the AlphaFold 3 inference architecture with OpenFold3 weights [71,74], Chai-1 [70], Boltz-1 [44], and Boltz-2 [69], with AF_unmasked [81] as a diagnostic. The *post hoc* re-score (above) added OpenFold3 [71], RoseTTAFold3 [72], and OpenDDE [73]. Designs were scored with confidence metrics, such as ipSAE [41,82] and ipTM, and for self-consistency with DockQ [42,83]. Novelty and redundancy screens used MMseqs2 [84] against UniRef90 and a corpus of published binder sequences assembled from ProteinBase [66] for sequence comparisons, and TM-align [85] and Foldseek [86] for structural comparisons. AlphaFold 3 weights [20,87], Rosetta and PyRosetta [88,89], and ESM3 [90] were excluded on license grounds.

Experimental validation by two independent contract research organizations (CROs)

The 1,320 designs across the 15 targets were sent, as delivered by the campaigns, to two contract research organizations (CROs) that made and measured them independently and in different formats (Figure M3A). Neither CRO saw the other's data; the design models; or the model, campaign, or rank behind a sequence; and each CRO's data were labeled without sight of the other's (below). The 120 designs against mature GDF-8, from four campaigns, were expressed and screened at both CROs, but the target aggregated under assay conditions and bound the sensor surfaces non-specifically in both formats, causing uninterpretable responses. Those measurements were inconclusive, and mature GDF-8 is excluded from every count and analysis in this report.

Adaptyv Bio (Lausanne, Switzerland) received all 1,320 designs. They expressed each design individually by cell-free synthesis and immobilized it on the sensor, one design copy per site, with the target in solution as analyte (SpCas9 as the sgRNA-loaded ribonucleoprotein; RBX1 alone (with Zn); latent GDF-8 as the pro-complex). Binding was read by surface plasmon resonance in single-cycle format: five target concentrations (1 µM, 316, 100, 31.6, and 10 nM) with 5 min association each and no regeneration [91]. A minority of replicates were read by bio-layer interferometry [92] over a shorter series instead. Two or more replicate ligands were run per design where expression allowed. In all, Adaptyv Bio returned a result for 1,296 of the 1,320 designs (336 binders, 899 non-binders, and 61 that did not express); a design that entered the assay but did not express counts as a tested non-binder. Adaptyv Bio fitted each replicate they judged analyzable with 1:1, bivalent-analyte, and two-state models. For each design they delivered a binder or non-binder classification with a strength grade, the kinetic constants, and an expression class. Known binders run as plate controls confirmed target activity wherever a working control was available, including MBP, where no design bound; the TrkA positive controls did not express, and the latent GDF-8 plate ran without controls.

Twist Bioscience (South San Francisco, CA, USA) received 1,260 of the 1,320 designs, those against every target except latent GDF-8, and expressed them in parallel with Adaptyv Bio as human IgG1 Fc fusions in HEK293 cells, so that each Fc dimer carries two copies of the design. Each design–Fc was captured on an anti-Fc surface of a high-throughput SPR array, with one plate per target. The target was then titrated over six three-fold concentrations from 1 µM to 4.1 nM (from 150 nM for SpCas9), one cycle per concentration with 5 min association and 10 min dissociation, and no regeneration between cycles. Every plate carried the human target as one or two independent preparations. Eight plates also carried the mouse ortholog, six of them the cynomolgus ortholog as well, and five carried the unrelated protein CLEC12A as an off-target control. The SpCas9 plate carried apo SpCas9 beside two ribonucleoprotein preparations, and the RBX1 plate presented RBX1 as the CUL1–RBX1 complex. For each design and target preparation, Twist Bioscience delivered a 1:1 kinetic fit, a steady-state fit, and a quality flag on the fit, together with the expression titer and analytical size-exclusion chromatography (aSEC) profile of the Fc fusion. Twist Bioscience reports rate constants only where a titration supports a kinetic fit and regards a design with reported constants as a binder; because this criterion differs from Adaptyv Bio's, every Twist Bioscience label used here was re-derived from the raw titrations as described below.

Automated labeling. The two CROs return different readouts: Adaptyv Bio classifies each design as a binder or non-binder, whereas Twist Bioscience reports fitted rate constants, where the titration supports them, together with a fit-quality flag. We therefore labeled each dataset on its own, blind to the other, and then combined the two labels by a fixed rule.

For Twist Bioscience, each design was labeled binder, non-binder, or uninformative on its target-form preparations from the raw titration cycles and the delivered fit fields, with no Adaptyv Bio data in view. A binder label required two or more dose-ordered concentrations responding above the plate's reference and isotype floor; curved association rather than a linear ramp; measurable dissociation, or a steady-state fit within 30-fold of the kinetic fit; an R_{max} plausible for the amount of design–Fc captured; no comparable response on the CLEC12A off-target control where it was run; and the correct target form (for SpCas9, only the ribonucleoprotein preparations count). A design whose Fc fusion did not capture is uninformative rather than negative. The fit-quality flag alone neither makes nor breaks a label.

For Adaptyv Bio, their classification is kept as their label. Independently, every replicate trace was graded by the same criteria (dose-ordered response, curvature, dissociation), blind to that classification and to the Twist Bioscience data, so that the two readouts are judged by one standard and the rare response that is a square solvent step or a slow linear drift is recognized.

Both blind passes combined objective trace metrics with two independent visual reads of every design that showed any

response (agreement between readers, 98% for Twist Bioscience and 96% for Adaptyv Bio on positive versus not); no label was changed on metrics alone.

Combination. A design is a binder when (1) Adaptyv Bio reported it as a binder and its own traces support that, at high confidence when the blind Twist Bioscience label is also positive; (2) Adaptyv Bio did not, but both blind trace grades are positive; or (3) the blind Twist Bioscience label is positive and Adaptyv Bio did not measure the design or its measurement was uninformative (for example, the design did not express). A design is a non-binder when neither CRO shows positive evidence; when a response that Adaptyv Bio did not classify as binding has no Twist Bioscience support; or when an Adaptyv Bio binder classification rests on a trace graded as a step or drift and Twist Bioscience gives no support. A blind Twist Bioscience positive that the rule could not reconcile with the Adaptyv Bio result, most often an explicit Adaptyv Bio negative, was never promoted automatically: fourteen such designs were reviewed one by one against both CROs' raw traces, and six were labeled binders. Of the 1,235 designs measured at both CROs, Adaptyv Bio's classification and the presence of a Twist Bioscience fit agreed for 1,099 (89%; Cohen's $\kappa = 0.71$); where they disagreed, the integrated call followed the Adaptyv Bio result for 118 designs and the Twist Bioscience result for 18. In all, the integrated call set aside one of the 336 designs that Adaptyv Bio classified as binders and added 19 that it had not (12 classified as non-binders, 5 that did not express, and 2 that were not tested). Two plates carry a standing caveat: the SpCas9 ribonucleoprotein plate at Twist Bioscience, titrated from 150 nM, gave low responses plate-wide and had no positive control in the capture format; and on the RBX1 plate the target was the CUL1–RBX1 complex, in which CUL1 occludes part of the RBX1 surface, so a Twist Bioscience negative there carries less weight. The rule grades each binder as high, medium, or low confidence. It was written after both CROs had reported and is applied identically to every design; every input it uses is a released column, so each label can be recomputed.

Affinities. The equilibrium dissociation constant (K_D) of record of a design (`kd_nM_final` in the release) is the geometric mean of the specific 1:1 kinetic fits available for it: Adaptyv Bio's replicate fits, our standardized re-fits of the same replicates where they resolve the off-rate, and Twist Bioscience's per-lot fits on the human target where the blind pass judged the response specific. The release also lists each source separately and the range they span. At both CROs the target is the analyte, so avidity can arise only from the target. Values for monomeric targets are therefore 1:1 affinities (at Twist Bioscience the two design copies on each captured Fc are independent sites for a monomeric target), whereas values for the five oligomeric targets (VEGF-A, TNFα, 15-PGDH, Nipah G, and latent GDF-8), whose multivalent analyte can bridge neighboring design copies, are apparent affinities at both CROs. When we cite one CRO's value rather than the K_D of record, we name the CRO. When the off-rate sat at the fitting floor, or the kinetic and steady-state fits disagreed by more than 30-fold (as on the BHRF1 and TREM2 plates, where dissociation was too slow to measure), we report the steady-state value or a bound. We report values below the assay floor of about 100 pM as bounds, and all values to two significant figures.

Statistics and software. Proportions were compared with two-sided Fisher's exact tests. The association of design rank with binding was tested by permuting ranks within each campaign's list for a target (2,000 permutations), and the excess of the co-folding score's average precision over chance by a sign test across targets; correlations are Spearman rank correlations. Intervals on the Overath et al. dataset comparisons (Figure M1) are 95% percentile intervals from a bootstrap over targets. The dependence of mouse cross-reactivity on the number of substituted footprint residues was tested with a Cochran–Armitage trend test stratified by target [93,94]. Analyses used SciPy [95], scikit-learn [96], Biopython [97], and Matplotlib [98].

Figure M3. Experimental validation by two independent CROs. **A**, The two assay formats (Methods). Adaptyv Bio immobilized each cell-free-expressed design and flowed the target over it at five increasing concentrations (10 nM to 1 µM) in a single cycle, in duplicate, and classified each design as a binder or non-binder. Twist Bioscience captured each design as an Fc fusion on an SPR array and titrated the target over six three-fold concentrations (4.1 nM to 1 µM), one cycle each, and reported kinetic and steady-state fits, with a quality flag, wherever the titration supported them. **B**, Raw traces of six designs, re-plotted from the released data: Adaptyv Bio on the left of each pair (blue, dark to light with decreasing concentration; replicate 1) and Twist Bioscience on the right (green; human target). Panel headers give each CRO's readout as delivered (+, an Adaptyv Bio binder classification or a quality-passing Twist Bioscience fit; −, neither), and the arrow gives the integrated call. Dashed lines mark the start of dissociation, and traces without a response are drawn on a fixed ±15 RU scale. **i**, RBX1, positive at both CROs. **ii**, EGFR, captured at both with no response. **iii**, IL-7Rα, a binder on Adaptyv Bio evidence alone: a curved, dose-dependent response in both replicates and none at Twist Bioscience. **iv**, VEGF-A, reported by Adaptyv Bio as a weak binder but labeled a non-binder: the response is a square step without curvature whose fit sits at the upper bound of the assay, and Twist Bioscience saw no binding. **v**, TREM2, a binder on Twist Bioscience evidence alone: the design did not express at Adaptyv

Bio; at Twist Bioscience the response is curved, saturating, and reversible, the kinetic and steady-state fits agree, and the cynomolgus and mouse orthologs also bind (not shown). **vi**, Nipah G, quality-passing micromolar fits at Twist Bioscience but labeled a non-binder: no response at Adaptyv Bio, and at Twist Bioscience the responses are rectangular steps that scale linearly with concentration and recur on the CLEC12A off-target control on the same surface (inset, red, same scale), the signature of a bulk refractive-index step rather than binding.

Results

Working from the protocol prompt, Claude Opus 4.8 and Mythos Preview ran every campaign from target research to producing a final ranking of 30 designs per target, and we ordered said designs for synthesis and testing without modification (Methods). Data for all targets except mature GDF-8 were interpretable; this leaves 1,320 designs on 15 targets. We labeled each CRO's data blind to the other's and combined the two labels into a single call per design, the integrated call (Methods). By the integrated call, 354 of the 1,320 designs bound their targets, a hit rate of 26.8%. Claude obtained binders against 14 of the 15 targets, although successes against two of these targets were marginal.

The 15 targets span several classes (Fig. 1): cytokines and growth factors (TNFα [99], VEGF-A [100], and latent GDF-8, the pro-domain complex of myostatin [101,102]); cell-surface receptors (PD-L1 [103], TREM2 [104], IL-7Rα [105], TrkA [106], and EGFR [107]); a viral attachment glycoprotein (Nipah virus glycoprotein G, hereafter Nipah G [108,109]); a viral BCL-2 homolog (BHRF1 [110]); the RING subunit of an E3 ubiquitin ligase (RBX1 [111]); and two enzymes, SpCas9, which was assayed as the sgRNA-loaded ribonucleoprotein [112], and 15-PGDH [113]. We chose two further targets as tests rather than for their biology: BBF-14, a β-barrel that was itself designed *de novo* [78,114], and the maltose-binding protein of *Escherichia coli* (MBP) [115].

Claude achieves high hit rates and designs high-affinity binders

In the multi-target format each model designed against 14 targets at once, 13 of which are analyzed here. Opus 4.8 produced 88 binders from 390 designs (22.6%) and Mythos Preview produced 104 from 390 (26.7%). In the single-target format, Mythos Preview produced 158 binders from 450 designs (35.1%), and 143 from 390 (36.7%) on the 13 targets shared with the multi-target campaigns. Hit rates varied much more between targets than between campaigns. Pooled over campaigns, they ranged from 72 binders out of 90 designs on TREM2, 54 of 90 on VEGF-A, and 49 of 90 on IL-7Rα to 3 of 90 on BBF-14, 1 of 30 on 15-PGDH, and 0 of 90 on MBP. For the same model on the same 13 targets, the single-target format produced more binders than the multi-target format (143 against 104 of 390; Fisher's exact $p = 0.003$) and had the highest hit rate of the three campaigns on 7 of the 13, tying on 3 more, one of them MBP, where no campaign produced a binder. A multi-target campaign did better on BHRF1 and TrkA, and TNFα binders came only from Opus 4.8 campaigns, as described below. The single-target format also gave each target a dedicated 24-hour campaign and 2.8 times the per-target compute budget (Methods), so we cannot separate the effect of focus from that of budget.

All 354 binders have an equilibrium dissociation constant (K_D) of record, defined in Methods. Of these, 194 bound below 100 nM, 90 below 10 nM, and 42 below 1 nM. The values span five orders of magnitude, from the assay floor of about 100 pM to a few micromolar. Six TREM2 binders sit at that floor and are reported as below 100 pM (Methods). Five of the targets are oligomeric (TNFα, VEGF-A, Nipah G, latent GDF-8, and 15-PGDH) and account for 100 of the 354 binders; affinities measured against them are avid and are therefore apparent values. Among the 254 binders to monomeric targets, 142 bound below 100 nM, 78 below 10 nM, and 38 below 1 nM. On nine targets the best design reached single-digit nanomolar affinity or better, with apparent values for three of them (TNFα, VEGF-A, and latent GDF-8).

Claude's ranking of its designs is calibrated with binding

Each campaign ended with Claude ranking its 30 designs per target from most to least promising, on the basis of its co-folding scores and the protocol's written criteria for interface quality, diversity, and risk (Methods). Because we never changed this order, the value of the ranking can be read directly from the results (Fig. 2). The average precision of Claude's rank for identifying binders, computed within each target and averaged over the 13 targets with enough binders and non-binders to evaluate (all except MBP, with no binders, and 15-PGDH, with one), was 0.48, compared to an expected 0.35 for

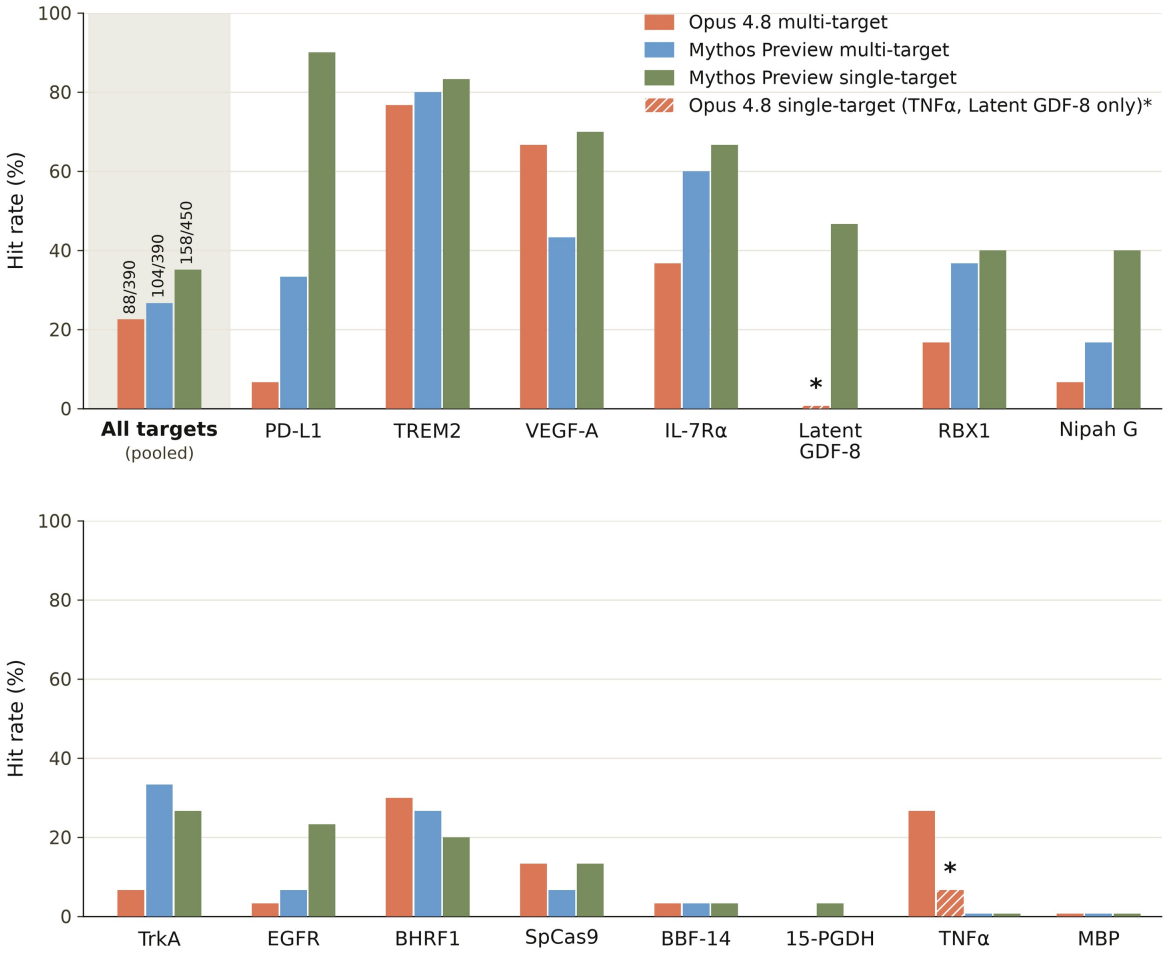


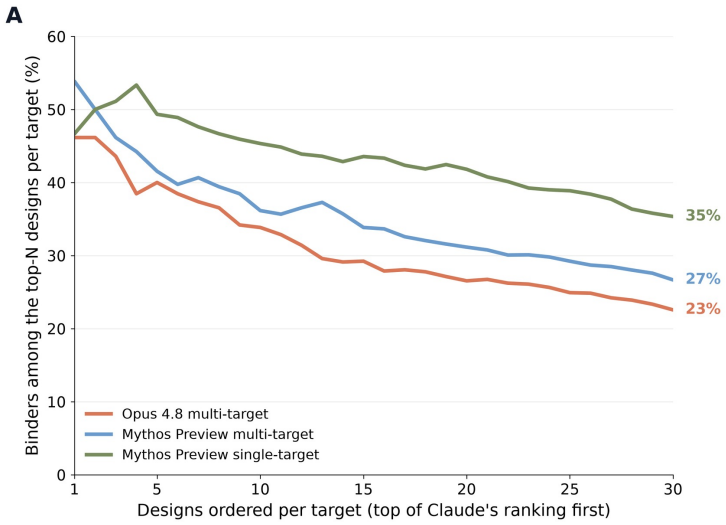
Figure 1. Claude designed binders against 14 of 15 targets. Hit rate of each campaign on each target: binders by the integrated call over the 30 designs delivered per target. Opus 4.8 multi-target (orange) and Mythos Preview multi-target (blue) each designed against these targets in one 48-hour session (13 of their 14 targets are analyzed here); Mythos Preview single-target (green) gave each of the 15 targets its own 24-hour session. The shaded column pools each campaign's targets (binders/delivered designs printed on the bars). Hatched bars (*) are the two Opus 4.8 single-target campaigns, on latent GDF-8 (0/30) and on TNFα (4/60: the 30 delivered designs together with the supplementary 30 drawn from the same campaign's pool; Methods). Targets are ordered by single-target hit rate. Latent GDF-8 and 15-PGDH were run in the single-target format only, and latent GDF-8 was tested at Adaptyv Bio only.

an uninformative ordering (2,000 permutations of the ranks within each campaign's ranking of each target; one-sided $p < 0.001$). Under the same evaluation the re-computed co-folding score reached 0.52 (described below), and the order Claude delivered follows that score closely (median Spearman $\rho = 0.86$ within a target's ranking). Claude's ranking therefore performed about as well as the co-folding score on which it was largely based, but not better.

The practical consequence is that binders were concentrated at the top of Claude's rankings (Fig. 2A). Pooling the 41 rankings of 30 designs from the three campaigns that covered 13 or more targets (one ranking per target and campaign), the hit rate was 49% for the top-ranked design alone, 44% over the top five, and 39% over the top ten, against 28% over all 30. Testing only the top-ranked design would still have found a binder in 20 of the 41 rankings, and on 12 of the 14 targets that yielded one at all; testing the top five, in 27 rankings and on 13 targets, missing only BBF-14 (Fig. 2B).

Claude successfully orchestrates multiple open-source protein design models

The prompt required Claude to generate at least 50 backbones per target with each of seven designated structure-generation methods, and to draw each target's 30 designs from at least three of them, with no single method contributing more than half of the designs. Within those limits Claude decided what to filter, optimize, combine, and deliver, and it assembled a different pipeline for each target and campaign (Fig. 3A). Ten structure-generation methods contributed designs that were ordered (Methods), and the seven designated methods account for 1,297 of the 1,315 tested (Fig. 3B). Claude redesigned



B

Targets with at least one binder among the top N designs per target

N	Opus 4.8 multi-target	Mythos Preview multi-target	Mythos Preview single-target
1	6/13	7/13	7/15
2	7/13	7/13	9/15
3	7/13	7/13	11/15
4	7/13	8/13	11/15
5	8/13	8/13	11/15
10	8/13	9/13	11/15
15	10/13	10/13	11/15
20	11/13	10/13	13/15
25	12/13	10/13	13/15
30	12/13	11/13	13/15

Figure 2. Design rank is calibrated with binding. **A**, Hit rate over the N highest-ranked designs of each target, pooled over a campaign's targets, for the three campaigns run on 13 or more targets (Opus 4.8 and Mythos Preview multi-target, 13 targets each; Mythos Preview single-target, 15). Ranks are the order in which Claude delivered its 30 designs per target; nothing was filtered or re-ranked afterward. Values at the right margin are the hit rates over all 30 designs, to which each curve converges. **B**, Number of each campaign's targets with at least one binder among the top N designs, for selected N. The targets never covered are MBP (all three campaigns) and TNFα (both Mythos Preview campaigns). The Opus 4.8 single-target campaigns and supplementary TNFα set are not shown.

sequences mainly with SolubleMPNN [16,78] and put most designs through one or more rounds of *in silico* optimization before selecting them. In all, the tested designs came through 24 distinct combinations of structure-generation and sequence-design methods before optimization (Fig. 3A). Because Claude chose the tools, the hit rates per method are descriptive and not a controlled comparison of the methods. With that caveat, each of the seven main methods, which contributed 100 or more designs each, produced binders at hit rates between 22% and 43% (Fig. 3B), and the binders from every method spanned nanomolar to micromolar affinities with similar medians (Fig. 3C).

The prompt also specified no epitope. For ten of the targets, a structure of the complex with a natural protein partner or ligand defines that partner's epitope [103,105,107,108,110,111,116–120]. Counting a design as targeting this epitope when at least 30% of its modeled interface residues fall within it, 91% to 100% of the tested designs targeted the natural epitope on eight of these ten targets, 60% did so on Nipah G, and 14% did so on RBX1, a target discussed below. Claude thus concentrated its designs on the surfaces that natural partners use.

The designs are also novel in sequence. After the campaigns, and independently of the novelty screen that Claude ran during them, we searched every tested design against every protein chain in the Protein Data Bank [121]. For 1,285 of the 1,315 designs (98%) no chain aligned at 30% sequence identity or more ($E \leq 0.1$), and no design exceeded 56% identity to any entry. Most of the 30 designs with a detectable match resembled other designed helical-repeat scaffolds rather than any binder of their target, and among the 354 binders the only match was an immunoglobulin Fc fragment, at 37% identity to one SpCas9 binder.

Co-folding scores are calibrated with binding

To ask how much of the outcome the co-folding scores could have anticipated, we re-scored every tested design after the campaigns under uniform settings, using ESMFold2, ESMFold2-Fast, and Protenix v2. We ran each predictor with five seeds on one shared target construct per target, took each predictor's maximum $ipSAE_{min}$ over its seeds, and scored each design by the mean of these three maxima (Methods; Fig. 4A). We measured how well this score separated binders from non-binders within each target by average precision, for which the chance level is the target's hit rate. Average precision exceeded chance on 12 of the 13 evaluable targets (all except MBP and 15-PGDH; sign test, $n = 13$, $p = 0.003$), with a mean of 0.52 against a mean chance level of 0.31 (Fig. 4B). These values understate the enrichment that was available before selection, because the delivered designs had already been filtered on the same three predictors, which compresses the range of their scores. An ensemble of the seven re-scoring predictors that were not part of the campaigns' ranking score reached

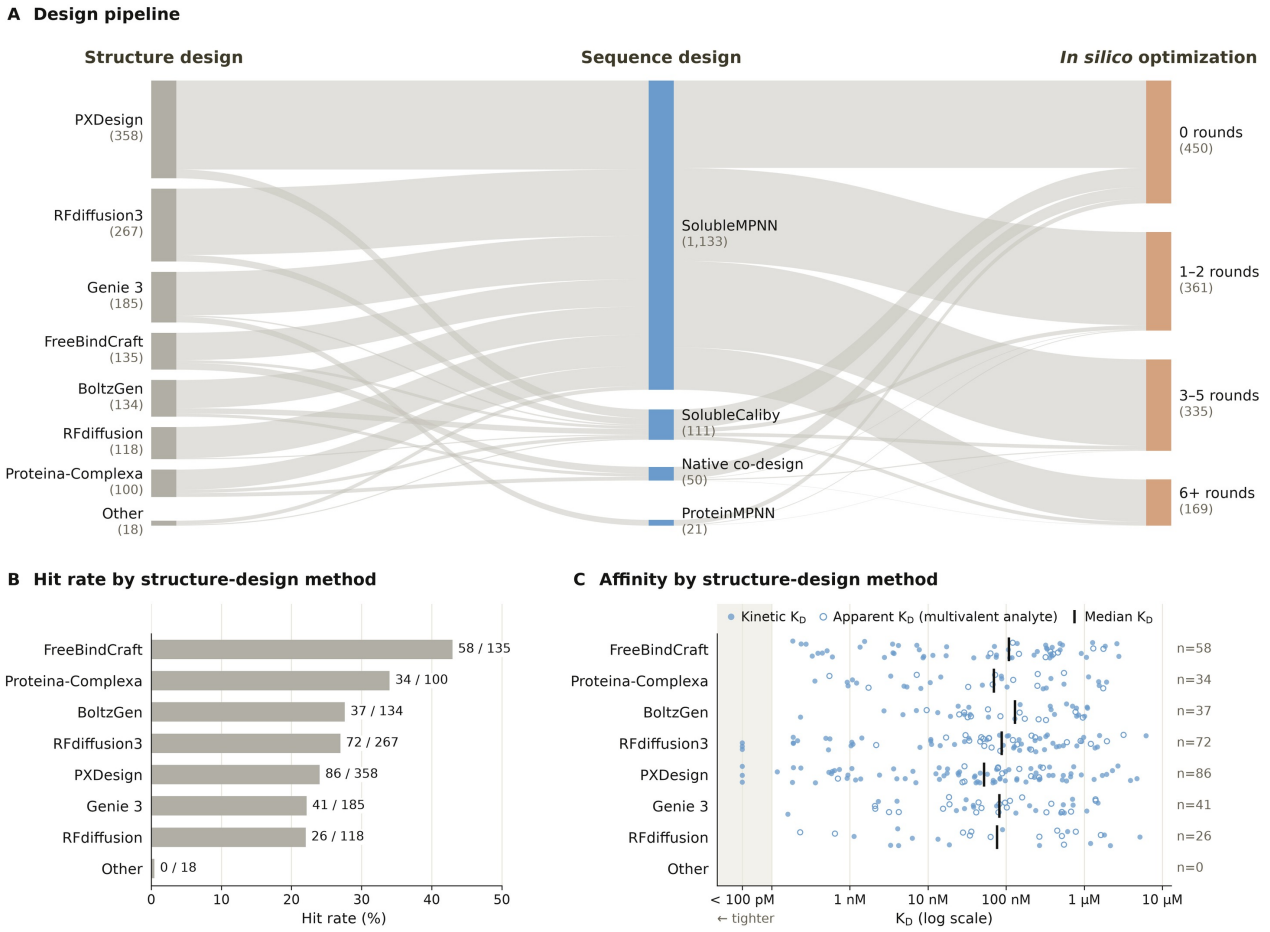


Figure 3. The design pipelines Claude assembled, and outcomes by structure-generation method. A, Sankey diagram of the 1,315 tested designs from the structure-generation method that produced each backbone (left), through the sequence-design method applied to it (middle), to the number of *in silico* optimization rounds before delivery (right). Node heights are proportional to the number of designs; methods contributing fewer than ten designs are grouped as "Other". **B**, Hit rate by structure-generation method (binders/tested designs). Because Claude chose the methods for each target and campaign, these rates describe its campaigns rather than compare the methods; most Genie 3 designs, for example, come from the campaign and targets with the lowest hit rates. **C**, K_D of every binder by structure-generation method (log scale; black ticks, medians; n, binders). Open points are apparent values for the multivalent analytes (TNFα, VEGF-A, Nipah G, latent GDF-8, and 15-PGDH), and values tighter than 100 pM are drawn in the shaded band at the assay floor.

0.57 on the same designs.

Across targets the scores were informative as well, although less sharply. A target's hit rate rose with the median score of its designs (Spearman ρ = 0.61 over the 15 targets, p = 0.02; ρ = 0.67 over the 13 targets whose target constructs the predictors resolve), so that targets with high-scoring pools generally yielded more binders (Fig. 4C). The relationship was not steep enough, however, to have flagged the failures in advance: the three least successful targets, MBP, BBF-14, and 15-PGDH (0/90, 3/90, and 1/30), had median scores of 0.68 to 0.70, only slightly below those of successful targets such as VEGF-A and Nipah G (both 0.72). Among the 354 binders, higher scores were only weakly associated with tighter binding (Spearman ρ = 0.16 pooled over targets, and 0.25 after standardizing score and affinity within each target).

Claude's designs are competitive with entries to Adaptyv Bio's protein design competitions

Six of the targets (TREM2, RBX1, latent GDF-8, Nipah G, EGFR, and 15-PGDH) have been the subject of open protein design competitions or hackathons run by Adaptyv Bio [51,55,56,58,62,122,123], with results deposited in ProteinBase [66]. In each, entries from multiple expert and non-expert groups were expressed and measured for binding in the same laboratory, using the same assay as our designs, which allows for retrospective comparison. It is not a controlled benchmark,

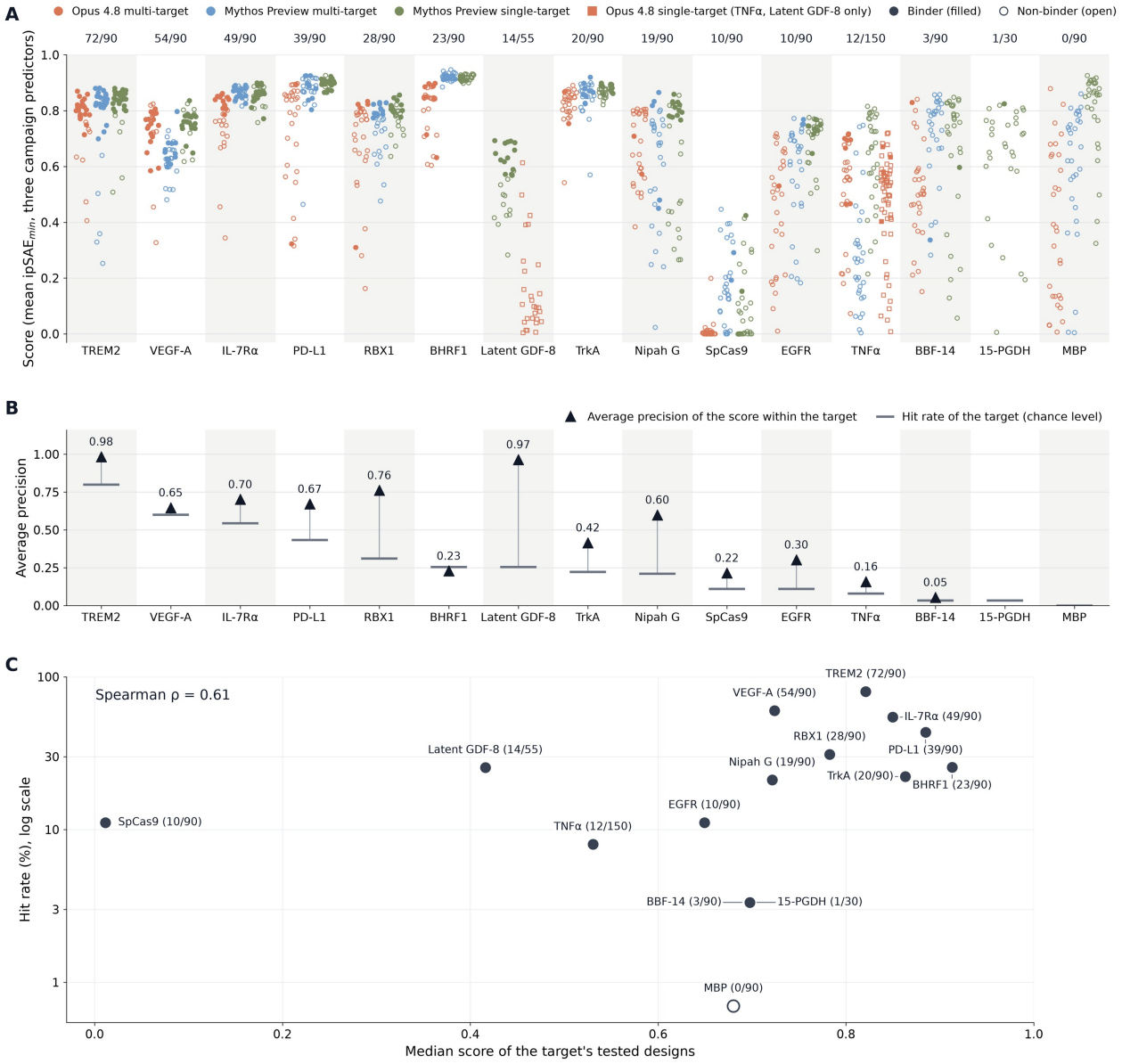


Figure 4. Co-folding scores are calibrated with binding. A, Re-scored confidence of every tested design (n = 1,315) by target, with targets ordered by hit rate (binders/tested above each column). For each of the three predictors used during the campaigns (ESMFold2, ESMFold2-Fast, and Protenix v2), a design's value is its maximum ipSAE$_{min}$ over five seeds on the shared target construct; the score plotted is the mean of the three (Methods). Colors distinguish campaigns as in Figure 1 (orange squares, the Opus 4.8 single-target campaigns); filled symbols are binders and open symbols non-binders. Scores for SpCas9, and for many latent GDF-8 designs, are near zero because most co-folds do not resolve their target constructs (the sgRNA-loaded enzyme and the pro-complex). **B**, Within-target average precision of the score (triangles) against each target's hit rate (horizontal marks), the value expected of a random ordering, for the 13 targets with at least three binders and three non-binders (all except MBP and 15-PGDH); here each predictor's maximum is z-scored within the target before the three are averaged. BHRF1 is the one target ranked below chance. **C**, Median score of each target's tested designs against its hit rate (log scale; labels, binders/tested). MBP, with no binders, is drawn as an open circle below the axis floor, and the BBF-14 and 15-PGDH points coincide.

because the competitions differ from our campaigns in the number of entries, the design budget per team, and the permitted formats (Fig. 5), and the reports or result collections of four of them (EGFR, TREM2, RBX1, and Nipah G) were on the reading list that accompanied the protocol prompt (Methods). For each competition we use the organizers' final tally of the *de novo* field, that is, entries classified as *de novo* by sequence identity, in minibinder and other non-antibody and non-peptide formats, and counted as binders when Adaptyv Bio's measurement gave a fitted K_D. We count Claude's binders by the integrated call, as in the rest of this report, which on these six targets gives 12 more binders than Adaptyv Bio's classification alone (7 on TREM2 and 5 on RBX1), and we pool its designs over campaigns. Counting them instead exactly as the organizers counted entries, by Adaptyv Bio's measurement alone, lowers our count only on TREM2 (65 rather than 72 of 90), RBX1 (23 rather than 28 of 90), and latent GDF-8 (13 rather than 14), and changes none of the comparisons below.

Claude's designs bound at higher rates than the field on four of the six targets: TREM2 (72/90 against 36/94 in the TREM2 hackathon [31,55]), RBX1 (28/90 against 9/245 [58]), Nipah G (19/90 against 69/666 [62]), and latent GDF-8 (14/60 against 6/100 [56,122], all 14 from the Mythos Preview campaign), with Fisher's exact $p < 0.01$ in each case; the latent GDF-8 comparison carries the caveat that our plate ran without controls (Methods). On the other two targets the rates were indistinguishable from the field: EGFR (10/90 against 20/284 [51,53]; $p = 0.26$) and 15-PGDH (1/30 against 1/106 [123]; $p = 0.39$). Separately from the hackathon, Adaptyv Bio also measured the Muni Proteina-Complexa auto-research collection [32], ten TREM2 designs generated with Proteina-Complexa in an automated workflow, of which 9 bound. At the same number of designs, 10, 9, and 10 of the ten top-ranked designs from our three TREM2 campaigns were binders. The tightest Muni design, which we included on our TREM2 plate as a control, bound below the 100 pM floor of the assay, as did the best design from each of our three campaigns. Figure 5B shows the hit rates for each of our campaigns separately.

For RBX1, we re-synthesized the winning entry of the GEM × Adaptyv RBX1 competition [57,58] and measured it on the same plate as our designs (Fig. 5A). On that plate the top-ranked design of the Mythos Preview single-target campaign bound free RBX1 at 3.9 nM (replicates of 4.1 and 3.7 nM), whereas the competition winner bound at 45 nM (compared with the 26 nM reported in the competition itself). Pooled with our re-fit of the same curves and the Twist Bioscience measurement, its K_D of record is 7.0 nM (Methods). The top-ranked design of the Opus 4.8 multi-target campaign bound at 30 nM, within the 1.8-fold plate-to-plate variation seen for the winner. Twist Bioscience presented the target as the CUL1–RBX1 complex, and there the Mythos Preview design bound with a K_D of 8.4 nM and gave no response against the CLEC12A off-target binding control. Against the same CUL1–RBX1 complex, the Opus 4.8 design gave a slowly associating response that did not saturate within the injections (Twist Bioscience fitted it at 106 to 400 nM, with the fit quality flagged on one lot). Although Claude designed against free RBX1, its designs largely avoided the face of RBX1 that CUL1 covers in the assembled ligase (PDB 1LDJ [111]): only 17% of the modeled footprint residues over all tested RBX1 designs lie on that face, and only 3 of the 28 RBX1 binders place half or more of their footprint there.

In affinity, the tightest binder from at least one of our campaigns bound as tightly as or more tightly than the tightest competition entry on five of the six targets (Fig. 5C). On Nipah G our binders were weaker; there the tightest competition entries target the stalk of the glycoprotein rather than the receptor-binding head domain that our designs address. Apart from RBX1, these comparisons are across separate experiments in the same laboratory and are indicative only.

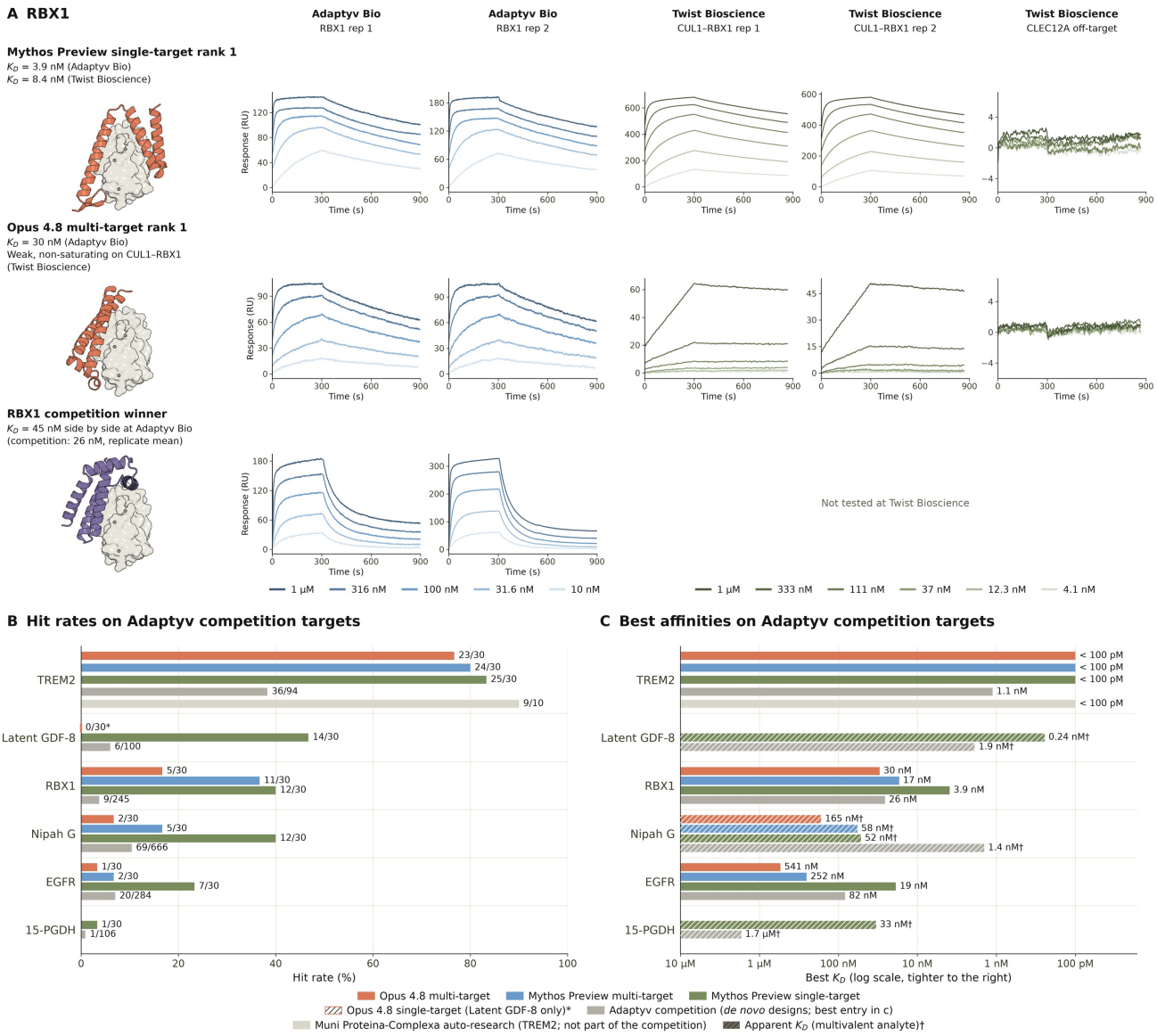


Figure 5 (legend on the following page).

Figure 5. Comparison with open design competitions measured in the same laboratory. A, RBX1. The top-ranked designs of the Mythos Preview single-target and Opus 4.8 multi-target campaigns, and the winning entry of the GEM × Adaptyv RBX1 competition [57,58], which we re-synthesized and measured on the same Adaptyv Bio plate as our designs. Left, each binder on the RBX1 RING domain (zinc ions as gray spheres), drawn with one RBX1 model and camera: design models for our two designs (orange) and a co-fold of the winner's published sequence (violet), for which no design model is available. Right, SPR sensorgrams colored by analyte concentration (keys below). Adaptyv Bio (blue; two replicate surfaces) flowed free RBX1 over the immobilized binder in a single cycle of five concentrations; Twist Bioscience (green) titrated two preparations of the CUL1–RBX1 heterodimer, and the unrelated protein CLEC12A as an off-target control, over the Fc-captured design. The winner was not in the Twist Bioscience panel. K_D values are each CRO's 1:1 kinetic fits (Adaptyv Bio, replicate means). **B**, Hit rates on the six targets for which Adaptyv Bio has run an open competition or hackathon [46,50]. Colored bars, Claude's campaigns (binders by the integrated call over 30 delivered designs; hatched , *the Opus 4.8 single-target campaign on latent GDF-8, a target tested at Adaptyv Bio only). Gray bars, the* de novo* entries of each competition by the organizers' final tally on ProteinBase [66] (non-antibody and non-peptide formats; binders confirmed with a fitted K_D): the TREM2 hackathon, in which AI-agent teams competed with human teams [30,31,55,124]; the GEM × Adaptyv RBX1 competition [58,59]; the Nipah G competition [60–62]; the EGFR competition, rounds 1 and 2 [48,49,51–53]; the Berlin Bio × Adaptyv 15-PGDH competition [123]; and the EPFL Protein Design Week × Adaptyv myostatin competition, tested against the same latent pro-complex [56,122]. The light bar under TREM2 is the Muni Proteina-Complexa auto-research collection [32], ten designs measured at Adaptyv Bio outside the hackathon. **C**, Best K_D per campaign on the same targets, on a reversed logarithmic axis so that longer bars are tighter: Adaptyv Bio's value for our tightest binder, and the ProteinBase replicate mean for the tightest *de novo* entry of each competition (for latent GDF-8 [122] and 15-PGDH [123], the values the organizers reported; the 15-PGDH competition had a single binder). Values below the 100 pM floor of the single-cycle assay are shown as < 100 pM. Hatched bars (†) are apparent values for multivalent analytes (the latent GDF-8 pro-complex, the Nipah G tetramer, and the 15-PGDH dimer), where avidity contributes on both sides of the comparison. The tightest Nipah G competition entries are reported by their authors to bind the G stalk rather than the receptor-binding head domain that our designs target.

Claude designed species cross-reactive binders against TNFα, a challenging, therapeutically relevant target

TNFα is the target of five approved biologics [125]. It is a compact homotrimer [99] whose receptor-binding grooves lie across the interfaces between subunits [116], and multiple *de novo* design efforts have reported no binders against it [5,7,14]. Twelve of Claude's 150 TNFα designs bound (8.0% hit rate), all of them designed by Opus 4.8: 8 from the multi-target campaign, 2 from the single-target campaign, and 2 from the supplementary set drawn from that campaign's pool (Methods). None of the 60 designs from the two Mythos Preview campaigns bound. Because each campaign ran once and the campaigns used different structure-generation methods, we cannot attribute this difference to the models. The twelve binders represent four of the 92 distinct backbones tested on this target: all eight multi-target binders are sequence variants of one Genie 3 backbone, and the four single-target binders derive from three PXDesign backbones. Figure 6 shows five of the twelve. In their models, multi-target ranks 2, 3, and 5 sit across an inter-subunit groove of the trimer, and all three also bind cynomolgus and mouse TNFα, although rank 3 binds the cynomolgus protein only weakly. Single-target ranks 9 and 15, from two of the PXDesign backbones, bind human TNFα and lose binding to mouse TNFα, and they differ on cynomolgus TNFα: rank 9 gave only a weak response, whereas rank 15 bound it at least as tightly as the human protein; soluble human TNFα differs from the cynomolgus protein at 4 of 157 positions and from the mouse protein at 33, several of them in the receptor-binding groove. Rank 15 was the tightest of the twelve, with an apparent K_D of 0.70 nM (0.24 nM in Adaptyv Bio's fit). Because TNFα is a trimer, all of these affinities are apparent values (Methods).

Claude designed fold-diverse binders with β-sheets

Computationally designed binders are predominantly α-helical bundles. Designs rich in β-structure are harder to make because their strands must pair in register, and they are more prone to misfolding and aggregation [126–128]. The prompt therefore set fold diversity as a secondary objective. At least 10% of the designs delivered for each target were to be not all-α by DSSP, that is, at least 3 of 30, unless meeting this quota would have displaced materially higher-scoring designs. By the DSSP fold class recorded for each design model in the release, 126 of the 1,320 delivered designs (9.5%) are not all-α, and the quota of three was met in 18 of the 43 rankings of 30 designs (one per target and campaign). These designs bound less often than the all-α designs: 16 of 125 tested (12.8%) versus 338 of 1,190 (28.4%). This difference is not controlled for target or method. Fifteen of the 354 binders, from ten distinct backbones against six targets, have design models that contain at least 20% β-strand. Figure 7 shows one design per backbone. They came from five structure-generation methods and from both Claude models. Three of them are all-β folds: BoltzGen β-sandwiches against VEGF-A and Nipah G, and a Proteina-Complexa design against SpCas9. The others are mixed α/β folds in which a small sheet packs against one or two helices, including RFdiffusion3 designs against PD-L1 and TREM2, Genie 3 designs against TREM2 and BBF-14, a Proteina-Complexa design against Nipah G, and a FreeBindCraft design against VEGF-A. Their affinities span almost three orders of magnitude. SpCas9 rank 29 bound at a few nanomolar, as did VEGF-A rank 22 as an apparent value against the dimer. TREM2 rank 16 bound at 37 nM at Adaptyv Bio and 24 nM at Twist Bioscience, and Nipah G rank 2 at about one

micromolar. Whether these designs adopt their modeled folds has not been tested experimentally.

Claude designed species cross-reactive binders against eight targets

Cross-species reactivity, which preclinical development generally requires, appeared in the prompt only as a secondary objective. Twist Bioscience titrated the mouse ortholog beside the human antigen on eight targets, and the cynomolgus ortholog on six of them, so that cross-reactivity could be read out for most binders on those targets (Fig. 8). Counting any reproducible response, including weak micromolar ones, 154 of the 179 binders with an evaluable cynomolgus titration bound the cynomolgus ortholog; 135 of them with a sub-micromolar fit, most within a few fold of the human K_D, and 19 only weakly. By the same count, 130 of the 233 binders tested against the mouse ortholog bound it, but here the outcome depended strongly on the target. On TREM2, 68 of 69 binders bound the mouse protein and on VEGF-A 32 of 53 did, whereas only 18 of 40 did so on IL-7Rα, 4 of 8 on EGFR, 3 of 12 on TNFα, 2 of 14 on TrkA, and 2 of 36 on PD-L1 (both of them micromolar); the single 15-PGDH binder bound the mouse ortholog weakly.

For every binder on these targets we counted the footprint residues, defined as target residues within 5 Å of the design in the model, that are substituted in the ortholog. This count did not predict which binders would cross-react. Pooled over targets, the fraction of binders that bound the mouse ortholog did not decline with it: 36%, 71%, 50%, and 52% of the binders with 0, 1 to 2, 3 to 4, and 5 or more substituted footprint residues bound, and within targets there was no significant trend (Methods). Among the 129 binders with a 1:1 fit in both species, weak fits included, the loss of affinity in mouse was likewise unrelated to this count (Spearman $\rho = 0.07$), and binders whose footprints contained a cynomolgus-substituted residue bound the cynomolgus protein as often as those whose footprints did not (35 of 41 against 119 of 138). Mouse cross-reactivity was instead a property of the target, and one that overall sequence divergence does not explain: mouse TREM2 and mouse PD-L1 are equally diverged from their human counterparts, at 73% identity each.

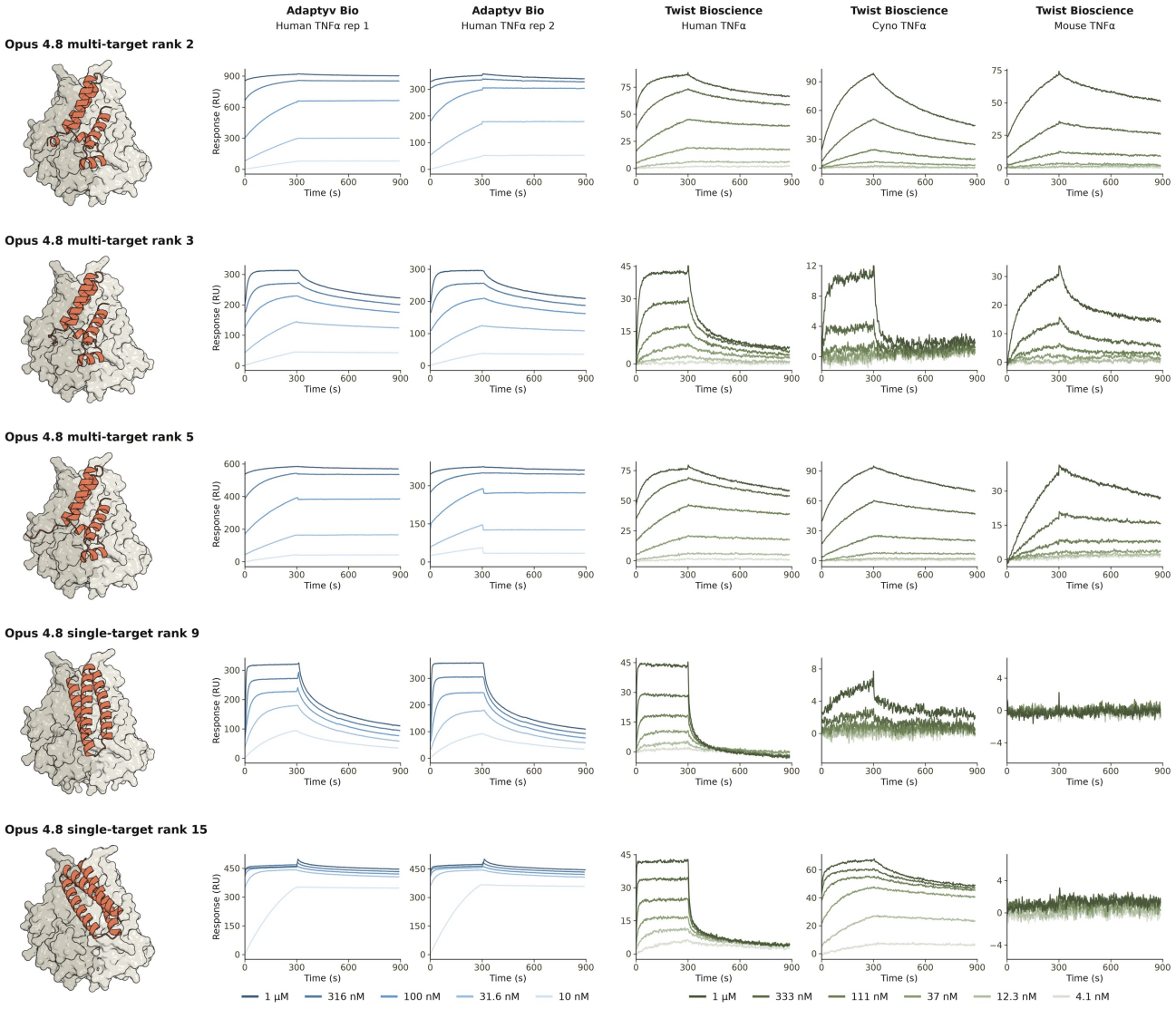


Figure 6. Claude designed binders against the TNFα trimer. Five of the twelve TNFα binders, all from Opus 4.8 campaigns: multi-target ranks 2, 3, and 5, which share one Genie 3 backbone, and single-target ranks 9 and 15, from two PXDesign backbones. Left, each binder (orange) on the TNFα trimer (light surface): design models for ranks 9 and 15, and Protenix v2 co-folds for ranks 2, 3, and 5, whose design files lack the target. Right, SPR sensorgrams from Adaptyv Bio (blue; two replicate surfaces; human TNFα flowed over the immobilized design) and Twist Bioscience (green; human, cynomolgus, and mouse TNFα over the Fc-captured design), colored by analyte concentration (keys below); dissociation begins at 300 s. Because the trimer was the analyte at both CROs, all affinities for these designs are apparent values.

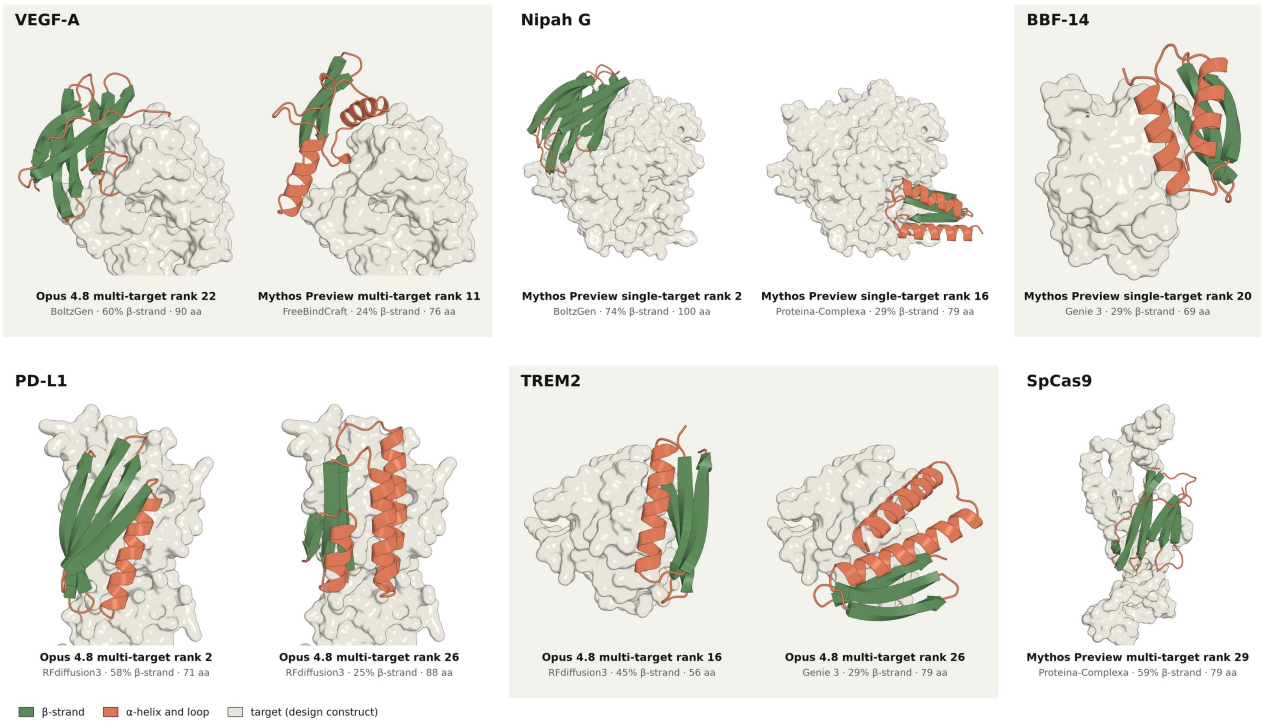


Figure 7. Claude designed binders that contain β-sheets. Models of ten binders with at least 20% β-strand content, one per generated backbone, grouped by target: the VEGF-A receptor-binding domain dimer, the Nipah virus glycoprotein G head domain, the *de novo* β-barrel BBF-14, the PD-L1 and TREM2 ectodomains, and the 275-residue SpCas9 domain with bound sgRNA that Claude used as its target construct. β-strands are green, helices and loops orange, and the target construct is a light surface in one fixed orientation per target. Labels give the campaign and rank, the structure-generation method, the β-strand content, and the binder length. Design models are shown except for VEGF-A rank 22 and TREM2 rank 26 (Protenix v2 co-folds). PD-L1 rank 26 is a lower-confidence binder (concentration-dependent binding at Adaptyv Bio, no response at Twist Bioscience).

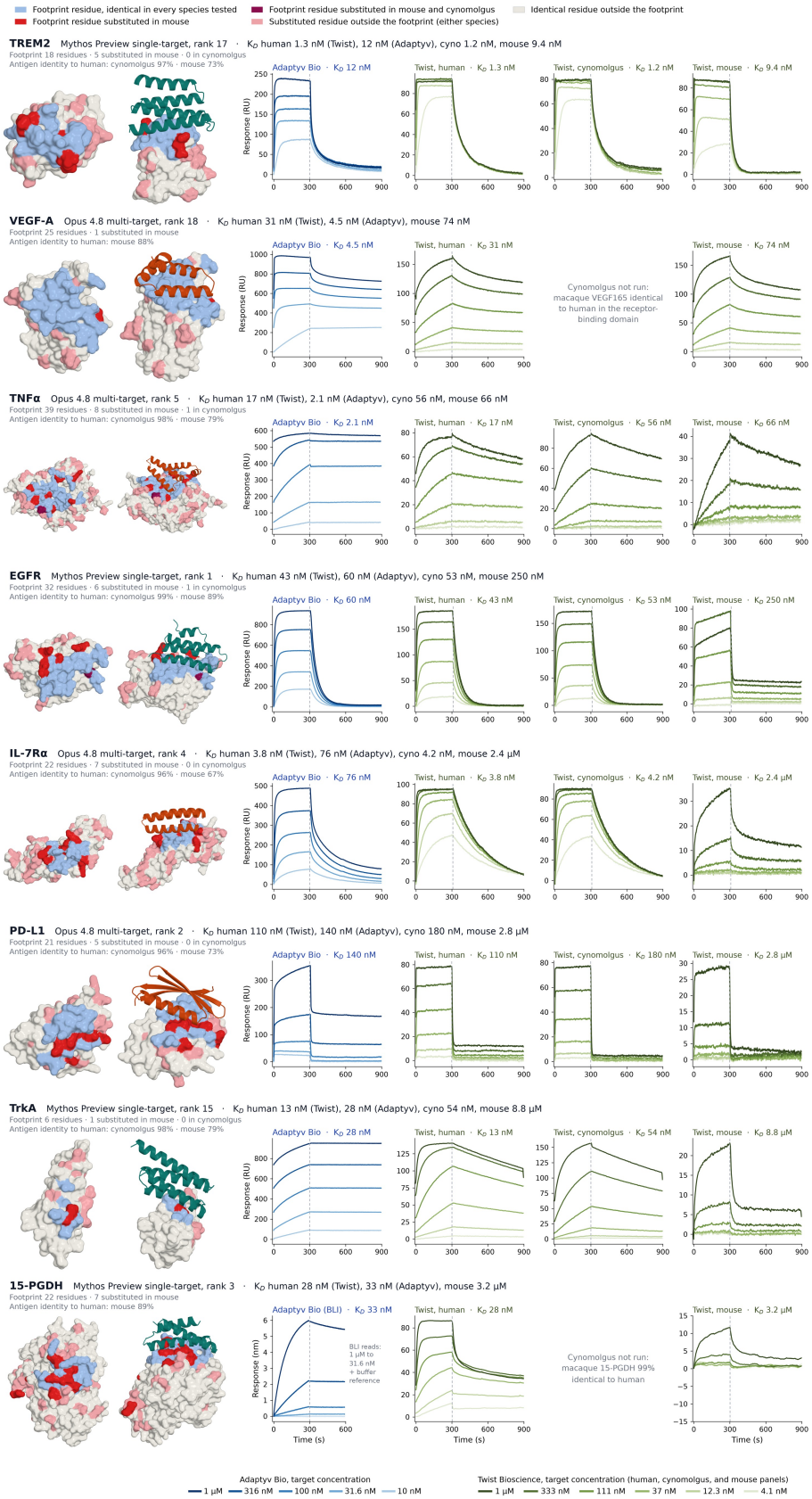


Figure 8 (legend on the following page).

Figure 8. Most of Claude's binders cross-react with their target's cynomolgus ortholog, and mouse cross-reactivity depends on the target. One representative binder for each of the eight targets for which designs were screened against an ortholog of the human antigen (cynomolgus and mouse for six targets; mouse only for VEGF-A and 15-PGDH, whose cynomolgus orthologs are identical or nearly identical to the human antigen over the region used). Each is a design with a clean 1:1 human titration, a reported fit on every ortholog for which any binder to that target had one, and a modeled footprint of at least ten residues; we show a tight, legible example rather than the tightest, and on IL-7Rα, PD-L1, TrkA, and 15-PGDH, where no binder combined tight human binding with a mouse fit inside the titration range, the weak mouse response of the selected design. Row headers give the campaign, rank, and each CRO's 1:1 K_D (apparent for the oligomeric VEGF-A, TNFα, and 15-PGDH); gray lines give the footprint counts and the overall antigen identity. Left image, the target surface under the design (removed), with footprint residues (within 5 Å of the design in the model) colored blue if identical in every species tested, red if substituted in mouse, and magenta if substituted in mouse and cynomolgus; rose marks substitutions outside the footprint. No footprint residue in any row is substituted in cynomolgus alone. Right image, the modeled complex (design models for TREM2, PD-L1, and TrkA; Protenix v2 co-folds for the rest). Traces, left to right: Adaptyv Bio, human target (blue; replicate 1; SPR, or bio-layer interferometry for 15-PGDH); Twist Bioscience, human, cynomolgus, and mouse target (green), dark to light with decreasing concentration; dashed lines mark the start of dissociation. All structures are predictions, and over all binders on these plates footprint substitutions did not predict ortholog binding (Results).

Claude struggled against some targets

BBF-14 is a 110-residue β-barrel that does not exist in nature: it was itself designed *de novo* [78,114], and BindCraft has produced binders against it [3] (Fig. 9). It therefore tests whether design methods work on a surface with no evolutionary history. Three of Claude's 90 designs bound it, one from each of the three campaigns and each from a different backbone: Mythos Preview multi-target rank 27 at 0.37 μM, Mythos Preview single-target rank 20 at 0.72 μM, and Opus 4.8 multi-target rank 21 at 0.98 μM. None of the three responded to the CLEC12A off-target control. MBP, the maltose-binding protein of *Escherichia coli* [115], is a two-lobed periplasmic protein with a large, convex, and predominantly polar surface, against which designed binders have been reported [54,129]. None of Claude's 90 MBP designs bound by the integrated call, although the known MBP binder run as a plate control bound on the Adaptyv Bio plates. The nearest candidate gave a small, reproducible response at Twist Bioscience that failed the quality criteria, and at Adaptyv Bio, which classified it as a non-binder, a concentration-dependent response in only one of two replicates. As noted above, the co-folding scores gave little warning of either failure: the designs against both targets scored nearly as well as those against productive targets (Fig. 4C).

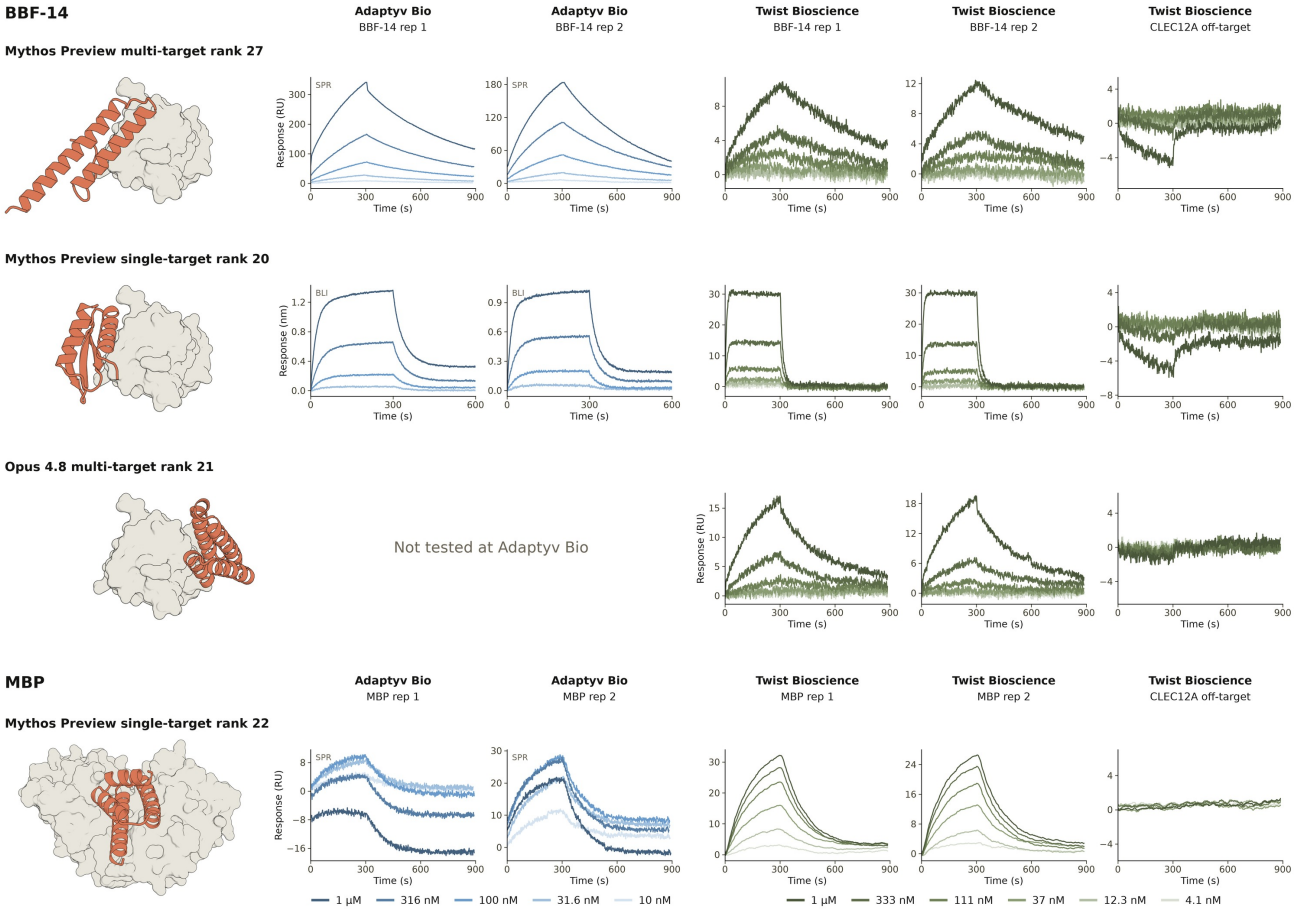


Figure 9. Claude designed three weak binders to the *de novo* β-barrel BBF-14 and none to maltose-binding protein (MBP). Left, model of each design (orange) on its target; right, sensorgrams from Adaptyv Bio (blue; two replicate surfaces; target flowed over the immobilized design) and Twist Bioscience (green; two target preparations and the unrelated protein CLEC12A as an off-target control, over the Fc-captured design), colored by analyte concentration (keys below). BBF-14 rows: the three binders, each from a different backbone (Mythos Preview multi-target rank 27, Mythos Preview single-target rank 20, and Opus 4.8 multi-target rank 21); none responded on CLEC12A. Rank 21 was not tested at Adaptyv Bio, and the Adaptyv Bio panels for rank 20 show bio-layer interferometry data from the vendor's online record that were not part of the delivered results and enter neither the integrated call nor the tested counts. MBP row: no design bound; the candidate shown (Mythos Preview single-target rank 22) gave a small, reproducible response at Twist Bioscience that failed quality criteria and a concentration-dependent response in only one of two Adaptyv Bio replicates, and Adaptyv Bio classified it as a non-binder.

Discussion

The question behind these campaigns was whether an AI agent, given a written protocol and open-source tools but no help with any individual decision, could carry a binder design campaign from target research to a ranked set of designs that bind. The results say that it can on most targets. Claude produced binders against 14 of the 15 targets with interpretable data, at hit rates of 27% over all designs and 49% for its top-ranked designs. On the six targets with open competitions measured in the same laboratory, its hit rates matched or exceeded those of the *de novo* entries, although the results of four of those competitions were available to it during design [51,55,58,62]. It also produced twelve binders, on four distinct backbones, against TNFα, a target on which multiple design efforts had reported none [5,7,14]. The failures are as informative as the successes: nothing bound MBP, three designs bound BBF-14, and one bound 15-PGDH, and in each case the co-folding scores that guided the campaign gave little warning (below).

Claude worked as an expert designer would with the same tools: it directed most of its designs at the surfaces that natural partners bind, and it relied on co-folding confidence scores to filter and rank them. We did not run a matched campaign by human experts, and we do not claim that Claude's designs are better than an expert would obtain with the same tools and budget. What the campaigns establish is that an AI agent can exercise this expertise autonomously, across 16 targets at once and within one to two days. We re-scored every tested design after the fact under uniform settings. Within a target, the ensembled confidence of the three campaign predictors separated binders from non-binders on 12 of the 13 evaluable targets (mean average precision 0.52, against 0.31 expected by chance). Across targets, the median score also rose with hit rate, but not steeply enough to have flagged the failures in advance. Designs against MBP, BBF-14, and 15-PGDH scored nearly as highly as designs against targets that yielded many binders, such as VEGF-A and Nipah G. Among the binders, the scores were also only weakly related to affinity. A confident co-fold was therefore a useful requirement for selection but not a guarantee of binding, and experimental screening remains the only way to learn which targets a campaign has succeeded on and the affinity of the designed binders.

The chief limitation of this study is that its evidence is binding, not structure or function. A binder here is a design that gave a concentration-dependent binding signal at one or both CROs under the integrated call. No design was tested for activity or solved structurally, so every pose shown is a prediction, and the affinities on the five oligomeric targets are apparent values. We note that *de novo* binders that pass co-folding filters have generally been found to bind as designed when solved [1,3,24]. Second, we count designs per sequence, although sequence variants of one backbone are not independent. Counting only the best-ranked sequence from each of the 809 generated backbones, 200 bound (24.7%), and the main comparisons are unchanged. Third, we ran each combination of model, format, and target once, apart from the supplementary TNFα order described in Methods, and campaigns against the same target could differ widely. Model, format, and run-to-run variation are therefore confounded, and we have described campaigns rather than models. Finally, most of the targets chosen are extensively characterized in the literature.

Language models have been coupled to protein design tools before, in pipelines that propose or rank binders [29,33,34], in agent systems whose designs were validated experimentally [27,31,32,35], and, more recently, in a hackathon in which autonomous AI agents designed TREM2 binders alongside human teams [30,31]. The present study differs in scale and in completeness. Claude made every design decision on 16 targets, from the choice of target construct to the final ranking, and we synthesized every design it delivered and report every measurement, including those from the campaigns that failed. Our work also differs in how little human intervention the campaigns required. We chose the targets, wrote the protocol, and placed the orders, but we did not install a design tool, choose an epitope, or rank a design, and every model Claude used is open-source [1–3,5,7,8,11,16]. The same frozen protocol served all 16 targets, suggesting that campaigns of this kind should be within reach of laboratories that have targets of interest but no expertise in computational protein design.

We release the prompts, models of all 1,440 designs with Claude's ranks, per-design provenance, and co-folding model predictions, and both CROs' binding data for the 1,320 designs with reliable measurements. The prompts allow any laboratory to rerun the protocol as it stands or to improve on it. The data provide a benchmark for filtering and ranking methods that complements curated meta-analyses and community collections [22,46,66], and the per-design provenance offers a record of Claude's design decisions for those who wish to study how an AI agent conducts such work.

Data, code, and prompt availability

The materials are released at https://huggingface.co/datasets/Anthropic/claude-protein-binder-design as three archives, under CC BY 4.0 for data and the MIT license for scripts. The prompt release contains the multi-target and single-target protocol prompts, the kickoff messages, and the accompanying document corpus, with credentials and internal identifiers redacted. The data release covers all 1,440 designs on the 16 targets: for every design the sequence, the design model (retained by the pipeline for 1,309 of the 1,440 designs), the seed-best co-folding predictions and scores, and the per-design provenance, and for the 1,320 designs on the 15 targets analyzed here both CROs' measurements (raw sensorgrams, fits, and Adaptyv Bio's classifications) and the integrated call. The structure and PAE release contains the co-folding models, scores, and predicted-aligned-error matrices for every seed of the ten predictors (AlphaFold-Multimer v3, which does not model nucleic acids, folded the SpCas9 designs against the apoprotein only). All design and structure-prediction tools are the open-source programs cited in Methods.

Acknowledgments

We thank Adaptyv Bio and Twist Bioscience for expression, purification, and binding measurements performed as contract research, and for answering questions about assay conditions, controls, and previously generated results. We thank Kiko Wemmer, John Ferrie, Lindsay Blohm, Cecilia Latapie, and Adrian Sadoogh for establishing the research operations and infrastructure that supported this work, Alexander Tarashansky and Aleksei Lorenz for improvements to Claude Science, and Richard Shuai, Vincent Fan, Nathan Frey, Colby Souders (Twist Bioscience), and Tudor-Stefan Cotet, Daniel Nakhaee-Zadeh Gutierrez, and Liza Kozlova (Adaptyv Bio) for reviewing this work.

Author contributions

Claude Science executed the design campaigns (run by Claude Opus 4.8 and Claude Mythos Preview) autonomously within the written protocol, from the choice of target construct and epitope through structure generation, sequence design, filtering, and optimization to selection and ranking of the 30 designs per target, and installed and operated the design and prediction tools. It also performed the *post hoc* re-scoring, the integrated call, and the statistical analyses, built the figures and the data release, and drafted the manuscript. A.S. conceived the study; defined the campaign formats, budgets, and protocol; selected the targets and assay antigens; wrote the campaign prompts and assembled the External Resource Corpus; provisioned tools, credentials, and cloud compute, and monitored infrastructure without intervening in design decisions; placed the synthesis and characterization orders with both CROs; directed the analysis and the writing; and reviewed and takes responsibility for the content of the report. The only designs not autonomously selected by Claude were the 30 in the supplementary TNFα set, which A.S. chose from the campaign's scored pool.

Competing interests

A.S. is an Anthropic employee. Adaptyv Bio and Twist Bioscience were paid to conduct the experimental validation in this study.

Use of AI tools

This report evaluates an AI system and was also prepared with one. The design campaigns were run by Claude Opus 4.8 and Claude Mythos Preview, as described in Methods. The subsequent work, including the *post hoc* re-scoring, the labeling of the binding data, the statistical analyses, the figures, the data release, and the drafting and revision of the text, was carried out by Claude models operating in Claude Science under the direction of the corresponding author, who checked the analyses against the released data and approved the final text. Every binding statistic reported for our designs can be recomputed from the released tables, and the structural statistics from the released structure files.

References

1. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
2. Butcher, J. *et al.* De novo design of all-atom biomolecular interactions with RFdiffusion3. *bioRxiv* https://doi.org/10.1101/2025.09.18.676967 (2025).
3. Pacesa, M. *et al.* One-shot design of functional protein binders with BindCraft. *Nature* **646**, 483–492 (2025).
4. Ring, A. M. FreeBindCraft: BindCraft with optional PyRosetta bypass. https://github.com/cytokineking/FreeBindCraft (2025).
5. Stark, H. *et al.* BoltzGen: Toward universal binder design. *bioRxiv* https://doi.org/10.1101/2025.11.20.689494 (2025).
6. Cho, Y., Pacesa, M., Zhang, Z., Correia, B. E. & Ovchinnikov, S. Boltzdesign1: Inverting All-Atom structure prediction model for generalized biomolecular binder design. *bioRxiv* https://doi.org/10.1101/2025.04.06.647261 (2025).
7. Protenix Team *et al.* PXDesign: Fast, modular, and accurate de novo design of protein binders. *bioRxiv* https://doi.org/10.1101/2025.08.15.670450 (2025).
8. Lin, Y. *et al.* Fast and Ultra-Capable protein design: Advancing the frontier through atomistic SE(3)-equivariance with Genie 3. *bioRxiv* https://doi.org/10.64898/2026.05.01.722168 (2026).
9. Geffner, T. *et al.* Proteina: Scaling flow-based protein structure generative models. *arXiv* https://doi.org/10.48550/arXiv.2503.00710 (2025).
10. Geffner, T. *et al.* La-Proteina: Atomistic protein generation via partially latent flow matching. *arXiv* https://doi.org/10.48550/arXiv.2507.09466 (2025).
11. Didi, K. *et al.* Scaling atomistic protein binder design with generative pretraining and test-time compute. *arXiv* https://doi.org/10.48550/arXiv.2603.27950 (2026).
12. Rustamov, K. R. & Baev, A. Y. Fold-Conditioned de novo binder design via AlphaFold2-Multimer hallucination. *bioRxiv* https://doi.org/10.1101/2025.07.02.662497 (2025).
13. Cho, Y., Rangel, G., Bhardwaj, G. & Ovchinnikov, S. Protein hunter: Exploiting structure hallucination within diffusion for protein design. *bioRxiv* https://doi.org/10.1101/2025.10.10.681530 (2025).
14. Zambaldi, V. *et al.* De novo design of high-affinity protein binders with AlphaProteo. *arXiv* https://doi.org/10.48550/arXiv.2409.08022 (2024).
15. Ingraham, J. B. *et al.* Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).
16. Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
17. Shuai, R. W., Lu, T., Bhatti, S., Kouba, P. & Huang, P.-S. Ensemble-conditioned protein sequence design with Caliby. *bioRxiv* https://doi.org/10.1101/2025.09.30.679633 (2025).
18. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
19. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* https://doi.org/10.1101/2021.10.04.463034 (2021).
20. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
21. Bennett, N. R. *et al.* Improving de novo protein binder design with deep learning. *Nature Communications* **14**, 2625 (2023).
22. Overath, M. D. *et al.* Predicting experimental success in de novo binder design: A Meta-Analysis of 3,766 experimentally characterised binders. *bioRxiv* https://doi.org/10.1101/2025.08.14.670059 (2025).
23. Chevalier, A. *et al.* Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017).
24. Cao, L. *et al.* Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).
25. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).

26. M. Bran, A. *et al.* Augmenting large language models with chemistry tools. *Nature Machine Intelligence* **6**, 525–535 (2024).
27. Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature* **646**, 716–723 (2025).
28. Gottweis, J. *et al.* Accelerating scientific discovery with Co-Scientist. *Nature* **655**, 487–496 (2026).
29. Ghafarollahi, A. & Buehler, M. J. ProtAgents: Protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery* **3**, 1389–1409 (2024).
30. Adaptyv Bio. Can LLMs design proteins? How AI agents tied 10 human teams at designing TREM2 binders. https://www.adaptyvbio.com/blog/agents-vs-humans (2026).
31. Adaptyv Bio & muni. Adaptyv x muni hackathon: AI agents vs humans (ProteinBase collection). https://proteinbase.com/collections/adaptyv-x-muni-hackathon-ai-agents-vs-humans (2026).
32. muni. Muni Proteina-Complexa auto research (Proteinbase collection muni-proteina-complex-auto-research). https://proteinbase.com/collections/muni-proteina-complex-auto-research (2026).
33. Ge, F. *et al.* AutoBinder agent: An MCP-based agent for end-to-end protein binder design. *arXiv* https://doi.org/10.48550/arXiv.2602.00019 (2026).
34. Liu, G., He, M., Sun, L., Chen, F. & Zhang, Y. Open Intelligence Hub: An autonomous LLM-agent platform for computational binder design and conjugation-aware prioritization of antibody-drug conjugates. *bioRxiv* https://doi.org/10.64898/2026.04.21.719907 (2026).
35. Latent Labs Team. Latent-Y: A lab-validated autonomous agent for de novo drug design. *arXiv* https://doi.org/10.48550/arXiv.2603.29727 (2026).
36. Boltz. Driving the Boltz API with agents and integration with Claude Science. https://boltz.bio/boltz-claude (2026).
37. Candido, S. *et al.* Language modeling materializes a world model of protein biology. *bioRxiv* https://doi.org/10.64898/2026.06.03.729735 (2026).
38. Biohub. ESMFold2 and ESMFold2-Fast: Model weights and the `esm` inference package. https://huggingface.co/biohub/ESMFold2 (2026).
39. ByteDance AML AI4Science Team *et al.* Protenix - advancing structure prediction through a comprehensive AlphaFold3 reproduction. *bioRxiv* https://doi.org/10.1101/2025.01.08.631967 (2025).
40. Zhang, Y. *et al.* Protenix-v2: Broadening the reach of structure prediction and biomolecular design. *bioRxiv* https://doi.org/10.64898/2026.04.10.717613 (2026).
41. Dunbrack, R. L. Rēs ipSAE loquuntur: What's wrong with AlphaFold's ipTM score and how to fix it. *bioRxiv* https://doi.org/10.1101/2025.02.10.637595 (2025).
42. Basu, S. & Wallner, B. DockQ: A quality measure for Protein-Protein docking models. *PLOS ONE* **11**, e0161879 (2016).
43. Mirdita, M. *et al.* ColabFold: Making protein folding accessible to all. *Nature Methods* **19**, 679–682 (2022).
44. Wohlwend, J. *et al.* Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv* https://doi.org/10.1101/2024.11.19.624167 (2024).
45. The UniProt Consortium. UniProt: The universal protein knowledgebase in 2025. *Nucleic Acids Research* **53**, D609–D617 (2024).
46. Adaptyv Bio. Introducing BenchBB and the community paper of the Protein Design Competition. https://www.adaptyvbio.com/blog/benchbb/ (2025).
47. Adaptyv Bio. Protein optimization 101: Insights from the literature. https://www.adaptyvbio.com/blog/po101 (2024).
48. Adaptyv Bio. Protein optimization 102: Lessons from the protein design competition (EGFR round 1 results). https://www.adaptyvbio.com/blog/po102/ (2024).
49. Adaptyv Bio. Protein optimization 103: Racing to the top 100. https://www.adaptyvbio.com/blog/po103 (2024).
50. Adaptyv Bio. Protein design competition: Has binder design been solved? https://www.adaptyvbio.com/blog/po104/ (2025).
51. Cotet, T.-S. *et al.* Crowdsourced protein design: Lessons from the Adaptyv EGFR binder competition. *bioRxiv* https://doi.org/10.1101/2025.04.17.648362 (2025).

52. Adaptyv Bio. Adaptyv EGFR competition round 1 (Proteinbase collection adaptyv-egfr-competition-round-1). https://proteinbase.com/collections/adaptyv-egfr-competition-round-1 (2025).
53. Adaptyv Bio. Adaptyv EGFR binder design competition rounds 1 and 2 (collections). https://proteinbase.com/collections/adaptyv-egfr-competition-round-2 (2025).
54. Adaptyv Bio. Case study: Designing modular protein sensors using BindCraft. https://www.adaptyvbio.com/blog/mbp (2025).
55. bioArena & Adaptyv Bio. bioArena x Adaptyv hackathon: AI agents for TREM2 binder design (ProteinBase competition page). https://proteinbase.com/competitions/bioarena-adaptyv-trem2 (2026).
56. EPFL & Adaptyv Bio. EPFL Protein Design Week x Adaptyv: Myostatin (GDF-8) binder design. https://proteinbase.com/competitions/epfl-adaptyv-myostatin (2026).
57. GEM Workshop & Adaptyv Bio. GEM x Adaptyv Bio 2026 RBX1 binder design competition (announcement page). https://www.gembio.ai/gemadaptyv2026 (2026).
58. GEM Workshop & Adaptyv Bio. GEM x Adaptyv: RBX1 binder design competition, results. https://proteinbase.com/collections/gem-x-adaptyv-rbx1-binder-design-competition-results (2026).
59. GEM Workshop & Adaptyv Bio. GEM x Adaptyv: RBX1 binder design competition (ProteinBase competition page). https://proteinbase.com/competitions/gem-adaptyv-rbx1 (2026).
60. Adaptyv Bio. What happened in the Nipah protein design competition so far? https://www.adaptyvbio.com/blog/nipah-submissions/ (2025).
61. Adaptyv Bio. Nipah binder competition (competition page, ProteinBase). https://proteinbase.com/competitions/adaptyv-nipah-competition (2025).
62. Adaptyv Bio. Nipah competition results. https://proteinbase.com/collections/nipah-binder-competition-results (2026).
63. Adaptyv Bio. RFdiffusion re-validation collection. https://proteinbase.com/collections/rfdiffusion-re-validation (2025).
64. Adaptyv Bio. BindCraft v1 re-validation collection. https://proteinbase.com/collections/bindcraft1-revalidation (2025).
65. Adaptyv Bio. BoltzGen release designs collection. https://proteinbase.com/collections/boltzgen-release (2025).
66. Adaptyv Bio. ProteinBase: The home of protein design data. https://proteinbase.com/ (2026).
67. Google DeepMind. AlphaProteo generates novel proteins for biology and health research. https://deepmind.google/blog/alphaproteo-generates-novel-proteins-for-biology-and-health-research/ (2024).
68. Suzek, B. E. *et al.* UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2014).
69. Passaro, S. *et al.* Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv* https://doi.org/10.1101/2025.06.14.659707 (2025).
70. Chai Discovery *et al.* Chai-1: Decoding the molecular interactions of life. *bioRxiv* https://doi.org/10.1101/2024.10.10.615955 (2024).
71. The OpenFold3 Team. OpenFold3-preview. Software release, Zenodo https://doi.org/10.5281/zenodo.19485145 (2026).
72. Corley, N. *et al.* Accelerating biomolecular modeling with AtomWorks and RF3. *bioRxiv* https://doi.org/10.1101/2025.08.14.670328 (2025).
73. Aureka AI OpenDDE Project. Folding, reasoning, and scaling with open-source drug discovery engine. *arXiv* https://doi.org/10.48550/arXiv.2607.03787 (2026).
74. Ovchinnikov, S. *et al.* AlphaFold3-architecture JAX inference code with ported OpenFold3 weights. https://github.com/sokrypton/alphafold3 (2026).
75. Escalante Bio. Mosaic: Composable protein design by gradient optimisation through structure predictors. https://github.com/escalante-bio/mosaic (2025).
76. Fang, M. *et al.* HalluDesign: Protein optimization and de novo design via iterative structure hallucination and sequence design. *bioRxiv* https://doi.org/10.1101/2025.11.08.686881 (2025).
77. Vázquez Torres, S. *et al.* De novo design of high-affinity binders of bioactive helical peptides. *Nature* **626**, 435–442 (2024).

78. Goverde, C. A. *et al.* Computational design of soluble and functional membrane protein analogues. *Nature* **631**, 449–458 (2024).
79. EvolutionaryScale. ESM cambrian: Revealing the mysteries of proteins with unsupervised learning. https://www.evolutionaryscale.ai/blog/esm-cambrian (2024).
80. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
81. Mirabello, C., Wallner, B., Nystedt, B., Azinas, S. & Carroni, M. Unmasking AlphaFold to integrate experiments and predictions in multimeric complexes. *Nature Communications* **15**, 8724 (2024).
82. Dunbrack Lab. IPSAE: Code for the ipSAE interface score. https://github.com/DunbrackLab/IPSAE (2025).
83. Mirabello, C. & Wallner, B. DockQ v2: Improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics* **40**, btae586 (2024).
84. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**, 1026–1028 (2017).
85. Zhang, Y. & Skolnick, J. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* **33**, 2302–2309 (2005).
86. Kempen, M. van *et al.* Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* **42**, 243–246 (2024).
87. Google DeepMind. AlphaFold 3 inference pipeline (source code and model parameters on request). (2024).
88. Leman, J. K. *et al.* Macromolecular modeling and design in Rosetta: Recent methods and frameworks. *Nature Methods* **17**, 665–680 (2020).
89. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).
90. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
91. Karlsson, R., Katsamba, P. S., Nordin, H., Pol, E. & Myszka, D. G. Analyzing a kinetic titration series using affinity biosensors. *Analytical Biochemistry* **349**, 136–147 (2006).
92. Concepcion, J. *et al.* Label-Free Detection of Biomolecular Interactions Using BioLayer Interferometry for Kinetic Characterization. *Combinatorial Chemistry & High Throughput Screening* **12**, 791–800 (2009).
93. Mantel, N. & Haenszel, W. Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *JNCI: Journal of the National Cancer Institute* **22**, 719–748 (1959).
94. Mantel, N. Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association* **58**, 690–700 (1963).
95. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).
96. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
97. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
98. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**, 90–95 (2007).
99. Eck, M. J. & Sprang, S. R. The structure of tumor necrosis factor-α at 2.6 Å resolution. *Journal of Biological Chemistry* **264**, 17595–17605 (1989).
100. Muller, Y. A. *et al.* Vascular endothelial growth factor: Crystal structure and functional mapping of the kinase domain receptor binding site. *Proceedings of the National Academy of Sciences* **94**, 7192–7197 (1997).
101. Cotton, T. R. *et al.* Structure of the human myostatin precursor and determinants of growth factor latency. *The EMBO Journal* **37**, 367–383 (2018).
102. Walker, R. G. *et al.* Molecular characterization of latent GDF8 reveals mechanisms of activation. *Proceedings of the National Academy of Sciences* **115**, (2018).
103. Zak, K. M. *et al.* Structure of the complex of human programmed death 1, PD-1, and its ligand PD-L1. *Structure* **23**, 2341–2348 (2015).
104. Kober, D. L. *et al.* Neurodegenerative disease mutations in TREM2 reveal a functional surface and distinct loss-of-function mechanisms. *eLife* **5**, e20391 (2016).

105. McElroy, C. A., Dohm, J. A. & Walsh, S. T. R. Structural and biophysical studies of the human IL-7/IL-7Rα complex. *Structure* **17**, 54–65 (2009).
106. Wiesmann, C., Ultsch, M. H., Bass, S. H. & Vos, A. M. de. Crystal structure of nerve growth factor in complex with the ligand-binding domain of the TrkA receptor. *Nature* **401**, 184–188 (1999).
107. Ogiso, H. *et al.* Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. *Cell* **110**, 775–787 (2002).
108. Bowden, T. A. *et al.* Structural basis of Nipah and Hendra virus attachment to their cell-surface receptor ephrin-B2. *Nature Structural & Molecular Biology* **15**, 567–572 (2008).
109. Xu, K. *et al.* Host cell recognition by the henipaviruses: Crystal structures of the Nipah g attachment glycoprotein and its complex with ephrin-B3. *Proceedings of the National Academy of Sciences* **105**, 9953–9958 (2008).
110. Kvansakul, M. *et al.* Structural basis for apoptosis inhibition by Epstein-Barr virus BHRF1. *PLoS Pathogens* **6**, e1001236 (2010).
111. Zheng, N. *et al.* Structure of the Cul1–Rbx1–Skp1–F-box^{Skp2} SCF ubiquitin ligase complex. *Nature* **416**, 703–709 (2002).
112. Jiang, F., Zhou, K., Ma, L., Gressel, S. & Doudna, J. A. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* **348**, 1477–1481 (2015).
113. Niesen, F. H. *et al.* High-Affinity inhibitors of human NAD+-dependent 15-Hydroxyprostaglandin dehydrogenase: Mechanisms of inhibition and Structure-Activity relationships. *PLoS ONE* **5**, e13719 (2010).
114. Pacesa, M. & Correia, B. E. De novo designed beta barrel fold BBF-14. Protein Data Bank entry 9HAG. (2024) doi:10.2210/pdb9hag/pdb.
115. Quiocho, F. A., Spurlino, J. C. & Rodseth, L. E. Extensive features of tight oligosaccharide binding revealed in high-resolution structures of the maltodextrin transport/chemosensory receptor. *Structure* **5**, 997–1015 (1997).
116. Mukai, Y. *et al.* Solution of the structure of the TNF-TNFR2 complex. *Science Signaling* **3**, (2010).
117. Sudom, A. *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).
118. Wehrman, T. *et al.* Structural and mechanistic insights into nerve growth factor interactions with the TrkA and p75 receptors. *Neuron* **53**, 25–38 (2007).
119. Wiesmann, C. *et al.* Crystal structure at 1.7 Å resolution of VEGF in complex with domain 2 of the Flt-1 receptor. *Cell* **91**, 695–704 (1997).
120. Brozzo, M. S. *et al.* Thermodynamic and structural description of allosterically regulated VEGFR-2 dimerization. *Blood* **119**, 1781–1788 (2012).
121. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).
122. Adaptyv Bio & EPFL Protein Design Week. GDF-8 challenge results (ProteinBase collection, 100 designs from EPFL Protein Design Week tested against the myostatin (GDF-8) latent complex). https://proteinbase.com/collections/gdf-8-challenge-results (2026).
123. Adaptyv Bio & Berlin Bio. Berlin Bio x Adaptyv: 15-PGDH binder design competition (ProteinBase collection, 117 designs from the Berlin Bio x AI hackathon tested against 15-PGDH). https://proteinbase.com/collections/berlin-bio-x-adaptyv-15-pgdh-binder-design-competition (2026).
124. bioArena, Adaptyv Bio & muni. bioArena hackathon: In Silico to In Vitro (TREM2 binder design, San Francisco, February 2026). https://luma.com/a6t92ohv (2026).
125. Monaco, C., Nanchahal, J., Taylor, P. & Feldmann, M. Anti-TNF therapy: Past, present and future. *International Immunology* **27**, 55–62 (2015).
126. Dou, J. *et al.* De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485–491 (2018).
127. Kim, D. E. *et al.* De novo design of small beta barrel proteins. *Proceedings of the National Academy of Sciences* **120**, e2207974120 (2023).
128. Sappington, I. *et al.* Improved protein binder design using β-pairing targeted RFdiffusion. *Nature Communications* **17**, 1101 (2026).
129. Zhou, J. *et al.* De novo design protein binders for MBP and GST tags. *Biochemical and Biophysical Research Communications* **748**, 151322 (2025).