CAN POST-TRAINING ON AGENTIC CHAIN-OF-THOUGHT AUTONOMOUSLY SCREEN PROTEIN BINDERS FOR EXPERIMENTAL SUCCESS?

Anonymous authors
Paper under double-blind review

ABSTRACT

A well-known bottleneck exists in protein binder design: favorable *in silico* metrics do not necessarily translate to guaranteed success in the wet-lab. To address this, we introduce *BinderReason*: we post-train large language models (LLMs) via a two-stage post-training approach on Chain-of-Thought (CoT) reasoning traces generated by our autonomous multi-agent system, *BinderLab-Think*. We compile a curated collection of 6650 design–target complexes from three open-source datasets, representing 40 targets, of which 18.40% of designs are experimentally validated as good binders. These complexes are co-folded and analysed by *BinderLab-Think*, our closed-loop system of 11 specialised agents using Plan-and-Solve CoT prompting to predict whether each design binds to its target. We cold-start two open-weight models on the resulting traces to understand the biological, chemical, and physical concepts involved in binder analysis, and to align their internal thinking with a critical, systematic, and motivated approach to verdict generation. We then refine the cold-started Qwen3.5-9B model, *BinderReason-SFT*, via Group Relative Policy Optimisation (GRPO) with a domain-specific composite reward comprising five components: metric-weighted correctness, reasoning consistency, non-degeneracy, label validity, and format. This reward encourages consistency with structural evidence, suppresses hallucinated metric values and degenerate outputs, and enforces a parsable verdict. The resulting model, *BinderReason-RL*, achieves 0.8710 balanced accuracy and 0.8517 F1 on a target-disjoint test split, exceeding the strongest of ten zero-shot closed- and open-weight reasoning baselines by 31.2 percentage points in balanced accuracy. Finally, we analyse how post-training affects both classification and generated reasoning. Although performance remains target-dependent, *BinderReason* represents a step towards interpretable, domain-informed prediction of experimental binder success.

1 INTRODUCTION

Despite rapid progress in computational models for protein design (e.g., RFDiffusion (Watson et al., 2023), AlphaProteo (Zambaldi et al., 2024), Boltz-2 (Passaro et al., 2025), La-Proteina (Geffner et al., 2025)), the design of high-affinity protein binders remains an iterative, expert-driven workflow: candidates are generated, evaluated for structural and interface quality (e.g., by employing domain-specific analysis tools - Rosetta (Leman et al., 2020), GetContacts (GetContacts), DSSP (Hekkelman et al., 2025)), and then manually tuned based on those signals. While these AI-based methods for pre-clinical drug discovery bring enormous possibilities, only a handful of drugs passed wet-lab validation and progressed to clinical trials, where they failed at similar rates compared to drugs proposed via traditional non-AI-methods (Markossian et al., 2026). Therefore, the main bottleneck remains: strong *in silico* metrics and good stability, energetic, or contact-interface results do not necessarily translate to success in wet lab validation, and, later, clinical trials.

Agentic systems offer a natural solution: instead of focusing more resources on extensive campaigns of computational binder generation and optimisation for *in silico* performance, we use large language model (LLM)-powered agents to predict experimental binder success, based on which we filter our designs. Recent work leveraging prompt engineering techniques has demonstrated the promise of agentic systems in nanobody design (Swanson et al., 2025) or scientific hypothesis gen-

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

eration (Gottweis et al., 2026). However, such systems are limited by the performance of the base LLM, with frontier models struggling at biological and chemical tasks (Chen et al., 2026). Anthropic challenges this view: Claude Science & Shanehsazzadeh (2026) propose a fully autonomous binder-design campaign using two Claude models, reporting experimental successes on 14 distinct targets. Despite their success, deploying these models in open-ended industry-level scientific environments, with scarce experimental data, raises questions about reliability. A potential mitigation approach is aligning internal agentic thinking via supervised fine-tuning (SFT), followed by a rewards-based system with domain-specific constraints for enhanced performance (Guo et al., 2025).

Inspired by successful examples in chemistry and biology (e.g., ChemCRAFT (Li et al., 2026), BioReason-Pro (Fallahpour et al., 2026)), we propose a two-stage post-training approach for our binder experimental success prediction task: first, supervised fine-tuning, followed by Group Relative Policy Optimisation (GRPO), to align on internal thinking and *in silico* rewards. However, creating a training dataset for a complete fine tuning approach is challenging, as high quality traces could be expensive and time-consuming to curate (Wei & Kim, 2026), since they must be aligned with domain expert feedback. We decide to adopt a popular alternative in the field: a *cold start* for internal thinking alignment, followed by reinforcement learning techniques, shown to hold the potential for enhanced generalisability and reasoning capabilities (Chu et al., 2025). We summarise our approach and contributions below:

- We compile a collection of 6650 design–target complexes, representing 40 targets, with 18.40% of designs binding, by combining 3 open-source experimentally-validated datasets.
- We introduce **BinderLab-Think**, a closed-loop system of 11 specialised agents that co-folds and analyses these pairs, using Plan-and-Solve Chain-of-Thought (CoT) prompting to classify whether a design binds. From the 6650 entries, we produced 1659 traces spanning 27 distinct targets, with 13.26% binding designs. These traces are used to train LLMs to judge design quality from multi-source protein analysis data.
- We cold-start two open-weight models to understand the biological, chemical, and physical concepts in binder analysis, and to align their internal thinking to a critical, systematic, and motivated approach to verdict generation. We introduce a domain-specific reward system to refine the cold-started Qwen3.5-9B model, **BinderReason-SFT**, and enhance its ability to correctly predict key metrics, identify trends, obtain the correct verdict, avoid hallucinations, and exhibit the right format. The resulting model, **BinderReason-RL**, achieves competitive results compared to its SFT counterpart and zero-shot baselines.
- We provide a comprehensive study (quantitative results and qualitative analysis based on an expert-defined rubric) on the impact of post-training using agentic reasoning traces versus a base model, when predicting experimental binder success.

2 Related Work

Agentic systems in drug discovery. LLM-powered agentic frameworks have demonstrated promising results in drug-discovery related disciplines. ChemCrow (M. Bran et al., 2024) successfully augments LLMs with specialised chemistry and protein-design tools. ProteinCrow (Ponnapati et al., 2025) is an example of an agentic LLM-based protein design assistant, which deals with different protein-related design tasks: goal-directed protein design (i.e. improving stability of a target sequence), automated binder design, and binder optimisation. Latent-Y (Latent Labs Team et al., 2026) proposes a specialised agentic system for *de novo* minibinder design from text descriptions of drug design goals, by combining literature, a model (Latent-X2 (Team et al., 2025)) for generation and scoring, and bioinformatics tools for targets, epitopes, and designed molecules analyses. Recently, Anthropic (Claude Science & Shanehsazzadeh, 2026) proposed 2 autonomous binder-design campaigns, of 1-2 days, to quantify the ability of Claude Opus 4.8 and Mythos Preview to design and optimise binders against at most 16 targets, having access to unlimited token budget and to a fixed computational budget to run protein generation and analysis tools; the results are promising, with experimentally-validated binders recorded for 14 distinct targets.

Post-training approaches in drug discovery. Post-training approaches have been recently introduced successfully in drug-discovery related fields. ChemCRAFT (Li et al., 2026) proposes using tool trajectories to facilitate the model's understanding of chemical syntax and promote reasoning-

based tool-calling behaviour. BioReason-Pro (Fallahpour et al., 2026) uses over 100K LLM-based synthetically-generated traces for protein function prediction to train a model to generate biological reasoning traces, and refined it via reinforcement learning. Aspen (Chennakesavalu et al., 2026) is trained simply via RL-based post-training, (without any prior supervised fine-tuning since the base model was capable of generating structured reasoning traces) for several small molecules tasks.

To the best of our knowledge, there are no existing works employing post-training techniques based on agentic CoT traces for accurate prediction of binder experimental success. This paper addresses this gap, by introducing a pipeline for traces generation, followed by two-stage post-training.

3 METHODS

3.1 EXPERIMENTAL BINDER DATA COLLECTION

To support our aim to accurately classify computational designs as good or bad binders, we start from the ground truth. We compile a collection of 6650 design–target complexes, representing 40 targets, with 18.40% of designs binding, by combining 3 open-source experimentally-validated datasets, as shown in Table 1 (Claude Science* is a subset of the available data from Claude Science & Shanehsazzadeh (2026)).

Dataset	Binders	Targets	Binding rate
Overath (Overath et al., 2025)	3669	15	10.7%
Proteinbase (ProteinBase)	2630	24	18.2%
Claude Science* (Claude Science & Shanehsazzadeh, 2026)	351	13	100%

Table 1: Summary of the experimentally characterised protein binder datasets comprising our experimental binder data collection. Target overalap between datasets is present. Claude Science* contains by choice only binding designs.

3.2 BINDERREASON

BinderReason represents our solution to the open problem of translating *in silico* performance to experimental success predictions. It incorporates **BinderLab-Think**, a multi-agentic system for generating reasoning traces, which then serve as the foundation for the training and evaluation dataset used in our two-stage post-training approach. We illustrate the workflow in Figure 1.

3.2.1 BINDERLAB-THINK

BinderLab-Think is a framework to generate Chain-of-Thought reasoning traces utilizing a closed-loop multi-agent framework with *in silico* binder-target analysis, inspired by similar work proposed by Epure et al. (2026) for iterative minibinder generation, analysis and refinement. The system consists of 11 specialised agents that co-fold, analyse, and classify candidate designs as experimentally binding or not, as shown in Figure 1, part A. Recent evidence has shown that agent effectiveness depends on the alignment between coordination strategy and task structure (Kim et al., 2026). Motivated by recent success in subagentic workflows in the drug discovery pipeline (e.g., Claude Science & Shanehsazzadeh (2026) show Claude models autonomously conducting binder-design and analysis campaigns), we hypothesise that a set of simpler, specialised agents is better suited to this workflow than a single general-purpose agent. See Appendix C for full descriptions of the agents.

BinderLab-Think receives as input the collection of experimental binder data, having access to the design and target sequences, with the ground truth being withheld. Before any prediction effort begins, a *Laboratory Notebook* is initialised, serving as a common space for agents to store their respective contribution to the analysis and classification process. This is followed by a one-time targets' analysis step. The *Scholar* Agent receives an informative file correlating targets and accession identifiers, and then constructs a biological brief from UniProt, RCSB PDB, and Europe PMC, recording accession identifiers, functional annotations, available structural information, reported interaction sites, and relevant literature context for all targets. These briefs are cached and reused throughout the following runs, avoiding repeated literature retrieval during classification rounds. The agent's tools are described in Appendix C.2.1.

The main classification pass proceeds for all the binder-target complexes from Section 3.1. For a given pair, the *Foldsmith* Agent uses BoltzGen (Stark et al., 2025) to co-fold the design-target complex using the sequences provided, then runs the incorporated analysis and filtering. The resulting structures are preprocessed by converting structural files to a common format, adding hydrogens, and applying restrained energy minimisation. Analysis agents then compute complementary descriptors of each candidate, including sequence novelty, backbone geometry, secondary structure, non-covalent contacts, and interface-quality metrics. Each agent acts as *domain specialist*, executing their assigned tool(s), then reasoning over the output and saving their summary in the *Laboratory Notebook*. Tool-level details for the *Foldsmith, Prepper, Referee, Distogram, Ramachandran, GetContacts, DSSP,* and *Rosetta* Agents are provided in Appendices C.2.2 – C.2.9.

The *Forecaster* Agent interprets all the information available in the *Laboratory Notebook*, including literature brief, novelty analysis, raw structural metrics, past verdicts, then makes its own prediction. The agent is asked to motivate its decision and anchor it in the available structural evidence - we employ Plan and Solve CoT prompting techniques, and encourage the Forecaster to think step-by-step and apply structured reasoning: *understand* the task and pair given, analyse the *evidence*, *plan* how to approach the binary classification task, *carry out* the plan and reach a *verdict*. These structured reasoning traces will serve as the foundation of our two-stage post-training approach, due to their complexity and rigour in approaching the task - compared to standard base LLM generated reasoning traces, they leverage verifiable analysis tool outputs, making them not only more interpretable, but also closer to the traditional workflow in a drug discovery pipeline. Finally, the *Orchestrator* manages the workflow and retrieves the next pair. *Forecaster* and *Orchestrator* details are provided in Appendices C.2.10, and C.2.11.

3.2.2 Data Splits

We run BinderLab-Think with four open-weight reasoning models and retain traces whose predictions match the experimental ground truth, yielding 1,659 traces across 27 targets (13.26% positive). We assess target sequence similarity using MMseqs2 (Hauser et al., 2016; Steinegger & Söding, 2017) clustering and partition the dataset into target-disjoint training, validation, and test sets containing 21, 2, and 4 targets, respectively. This keeps all traces for each design–target complex within a single partition. Split construction prioritises training-set coverage while retaining both classes in each partition. Clustering parameters and split details are provided in Appendix E.1.

3.2.3 BinderReason-SFT

We begin our post-training approach with supervised fine tuning. We use the collection of agentic traces to *cold start* two open-weight models to understand the biological, chemical, and physical concepts in binder analysis, and to align their internal thinking to a critical, systematical, and motivated approach to verdict generation. We refer to the final fine-tuned model as **BinderReason-SFT**. Implementation details are left to Appendix E.3.

3.2.4 BinderReason-RL

The *cold started* models should now have their internal thinking structured and aligned to match a Plan and Solve CoT reasoning trace, using evidence from multi-source protein analysis data to accurately predict experimental binding. However, some of the limitations of small reasoning models are expected to persist: artificial metric values, degenerate answers, hallucinations, formatting issues, non-parsable outputs with missing verdict (Appendix E.4.3). To address them, we propose introducing a rewards-based system, and then train one of the *cold started* models via reinforcement learning to refine its internal thinking and better align on 9 key *in silico* metrics values (shown to be consistently referred to by BinderLab-Think when reaching a verdict in the agentic traces generated): *ipTM, pTM, binding_energy, num_hbonds, num_salt_bridges, RMSD, delta_sasa, pLDDT, ipLDDT* . Our reward-based system, described in Appendix E.4.1, represents a sum of five distinct rewards: *metric-weighted correctness, reasoning consistency, non-degeneracy, label validity, format*. For efficiency, the SFT model for Qwen3.5-9B is trained via GRPO, leading to **BinderReason-RL** - details on the training setup and hyperparameters used given in Appendix E.4. We chose to not pursue further RL training on the Gemma-4-12B model, despite optimistic results, at it showed a tendency to overfit, as illustrated in Figure 16 from Appendix E.3.

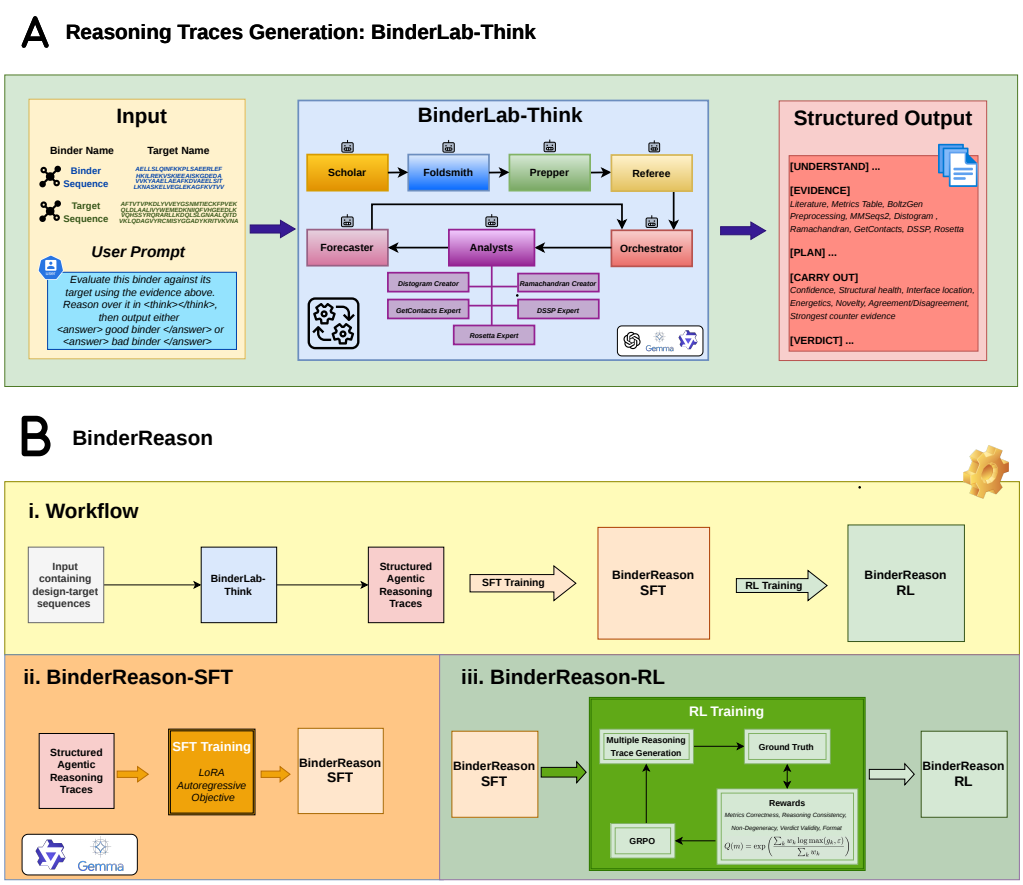


Figure 1: **BinderReason** Workflow. **A.** Reasoning Traces Generation process. **BinderLab-Think** is a framework to generate structured CoT reasoning traces utilizing a closed-loop multi-agent framework with *in silico* binder-target analysis (e.g., distogram and Ramachandran plots, files resulting from GetContacts, DSSP, Rosetta). **B. BinderReason Overview. i. Workflow.** BinderLab-Think takes as input design-target pairs with corresponding sequences and gives structured reasoning traces. These are used for the two-stage post-training approach. **ii. BinderReason-SFT.** Agentic traces are used to fine-tune a base open-weight model with an autoregressive objective, obtaining the BinderReason-SFT version. **iii. BinderReason-RL.** Agentic traces are used to further train via GRPO the BinderReason-SFT model using a geometric weighted mean of 5 distinct rewards: *metrics correctness, reasoning consistency, non-degeneracy, label validity, format.*

3.3 Evaluation

Protocol. To ensure comparability across experiments while keeping the computational budget tractable, we use a fixed evaluation protocol. We split the full reasoning traces data into train-validation-test, as shown in Appendix E.1, Table 17. The train and validation sets will be used in the post-training approach. To assess performance, we use the test set to run inference on the SFT and SFT + GRPO trained models; the experiments are repeated three times, with distinct random seeds.

Baseline. For each of the two open-weight models, the key baseline is a *zero-shot* run of the same model. In particular, we leverage the same user and system prompts on the base model, with reasoning capabilities, but without tool augmentation. For a more comprehensive analysis, we also used five closed-source flagship reasoning models and three larger open-weight reasoning models in the same setup. Experiments were run on the established test set, and repeated three times with distinct random seeds. More details on the setup, prompts, and models used are provided in Appendices B and A.

Domain Expert Evaluation One of the key dimensions we wish to optimise for is reliability and explainability, i.e. the ability of the models to reach a verdict decision based on rational justifications rooted in scientific evidence (e.g., *in silico* metrics, multi-source protein data). While the LLM-as-a-judge approach has shown promising results in the recent years, in fields where specialised knowledge is required, the validity of these systems remains uncertain (Szymanski et al., 2024). In a high-stakes domain such as drug discovery, where area expertise is essential, human feedback is vital. To achieve our aim, two domain-experts, consisting of one protein-design specialist and one chemical physicist with expertise in molecular modelling and physical analysis, were asked to support the human reviewing process. For comparable results, we propose a quantitative approach, inspired by the Legibility and Coverage Rubrics introduced by Emmons et al. (2025) to assess monitorability of CoT-based reasoning traces. We introduce a *Domain Experts Evaluation Rubric*, detailed in Appendix G, which proposes four dimensions, scored on a scale of 0 to 4, with 4 representing best performance. The experts received the rubric and were asked to independently evaluate the same 10 reasoning traces produced by the zero-shot and SFT runs of both open-weight models, and by the RL run for QWEN3.5-9B.

4 RESULTS

Reinforcement learning improves BinderReason classification. To assess whether reinforcement learning improves performance on the BinderReason classification task, we compared BinderReason-RL with its supervised initialization and closed and open-weight models evaluated in the zero-shot setting (Table 2). BinderReason-RL, built on SFT-Qwen3.5-9B, achieved the highest reported scores across all five evaluation metrics, with an accuracy of 0.8566 ±0.0135, balanced accuracy of 0.8710 ±0.0142, and F1 score of 0.8517 ±0.0188. Among the zero-shot baselines, Qwen3.5-9B achieved the highest balanced accuracy (0.5591), whereas Gemma-4-12B achieved the highest F1 (0.6009). BinderReason-RL exceeded these results by 31.19 percentage points in balanced accuracy and 0.2508 in F1, respectively.

We next examined the contribution of the RL stage by comparing the Qwen3.5-9B models before and after reinforcement learning. Supervised fine-tuning alone yielded a balanced accuracy of 0.5005 and F1 of 0.1367, below the corresponding zero-shot scores of 0.5591 and 0.4391. Following RL, balanced accuracy increased to 0.8710 and F1 to 0.8517. The improvement therefore extended beyond recovery of the zero-shot baseline: the RL-trained model surpassed both its supervised initialization and the original model on the evaluated classification task.

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
BinderReason-SFT					
QWEN3.5-9B	0.4289 ± 0.0187	0.5005 ± 0.0204	0.6054 ± 0.1505	0.0774 ± 0.0083	0.1367 ± 0.0145
GEMMA-4-12B	0.4795 ± 0.0324	0.5453 ± 0.0273	0.7448 ± 0.0844	0.1717 ± 0.0490	0.2773 ± 0.0701
BinderReason-RL					
QWEN3.5-9B	**0.8566 ± 0.0135**	**0.8710 ± 0.0142**	**1.0000± 0.0000**	**0.7420 ± 0.0283**	**0.8517 ± 0.0188**
Zero-Shot Baseline					
Generalist Reasoning Models					
CLAUDE OPUS 5	0.4318 ± 0.0097	0.5102 ± 0.0080	0.7333 ± 0.1155	0.0415 ± 0.0068	0.0786 ± 0.0129
CLAUDE OPUS 4.8	0.4057 ± 0.0231	0.4593 ± 0.0233	0.4749 ± 0.0644	0.1461 ± 0.0225	0.2234 ± 0.0334
GPT-6 ASTRA	0.4167 ± 0.0038	0.5019 ± 0.0032	0.3333 ± 0.5774	0.0037 ± 0.0065	0.0074 ± 0.0128
GPT-5.6 SOL	0.4846 ± 0.0467	0.5174 ± 0.0434	0.6089 ± 0.0674	0.3258 ± 0.0626	0.4241 ± 0.0697
GEMINI 3.8 FLASH	0.4391 ± 0.0164	0.5229 ± 0.0181	0.8000 ± 0.1202	0.0674 ± 0.0297	0.1238 ± 0.0514
Open-Weight Reasoning Models					
QWEN3.5-9B	0.5186 ± 0.0461	0.5591 ± 0.0462	0.6907 ± 0.0805	0.3219 ± 0.0387	0.4391 ± 0.0522
GEMMA-4-12B	0.5173 ± 0.0354	0.4964 ± 0.0354	0.5769 ± 0.0310	0.6269 ± 0.0346	0.6009 ± 0.0324
GPT-OSS-20B	0.4386 ± 0.0100	0.5167 ± 0.0089	0.7593 ± 0.0849	0.0599 ± 0.0172	0.1108 ± 0.0297
QWEN3.8-27B	0.4373 ± 0.0174	0.4713 ± 0.0247	0.5403 ± 0.0356	0.2707 ± 0.0304	0.3593 ± 0.0246
DEEPSEEK-V4-FLASH	0.5091 ± 0.0290	0.5388 ± 0.0343	0.6422 ± 0.0602	0.3603 ± 0.0169	0.4611 ± 0.0251

Table 2: **BinderReason Results.** Classification metrics across models. We highlight best performance on each metric in **bold**.

The precision and recall results further characterize this improvement. Recall increased from 0.0774 after SFT to 0.7420 after RL, accompanied by an increase in precision from 0.6054 to 1.0000. BinderReason-RL also exceeded the highest zero-shot recall, obtained by Gemma-4-12B (0.6269), while maintaining higher reported precision. Together, these results show that the improvement in F1 reflects gains in both precision and recall. They support the effectiveness of the RL stage for adapting the supervised model to the BinderReason classification objective.

Post-training helps BinderReason articulate biophysical concepts associated with favorable binding. We compared explanations from zero-shot Qwen3.5-9B, BinderReason-SFT Qwen3.5-9B, and BinderReason-RL Qwen3.5-9B on 36 matched, positive-labeled binder–target pairs (Figure 2). Using shared lexical embeddings and fixed scientific-topic annotations, we found that, as seen in Appendix H, Table 23, the mean within-pair fraction of energy-related passages increased from 2.4% in zero-shot outputs to 25.1% after SFT and 29.2% after RL. Interface-related discussion likewise increased from 2.1% to 15.5% and 15.9%, respectively. Post-training therefore makes energetic and interfacial concepts more prominent in binding explanations, with SFT establishing most of this shift and RL sustaining and further increasing their emphasis.

The matched trace example illustrates how these concepts contribute to the stated binding rationale: the SFT explanation weighs interface quality against energetic strain, whereas the RL explanation invokes favorable interaction energies and a stable binding mode to support its positive prediction

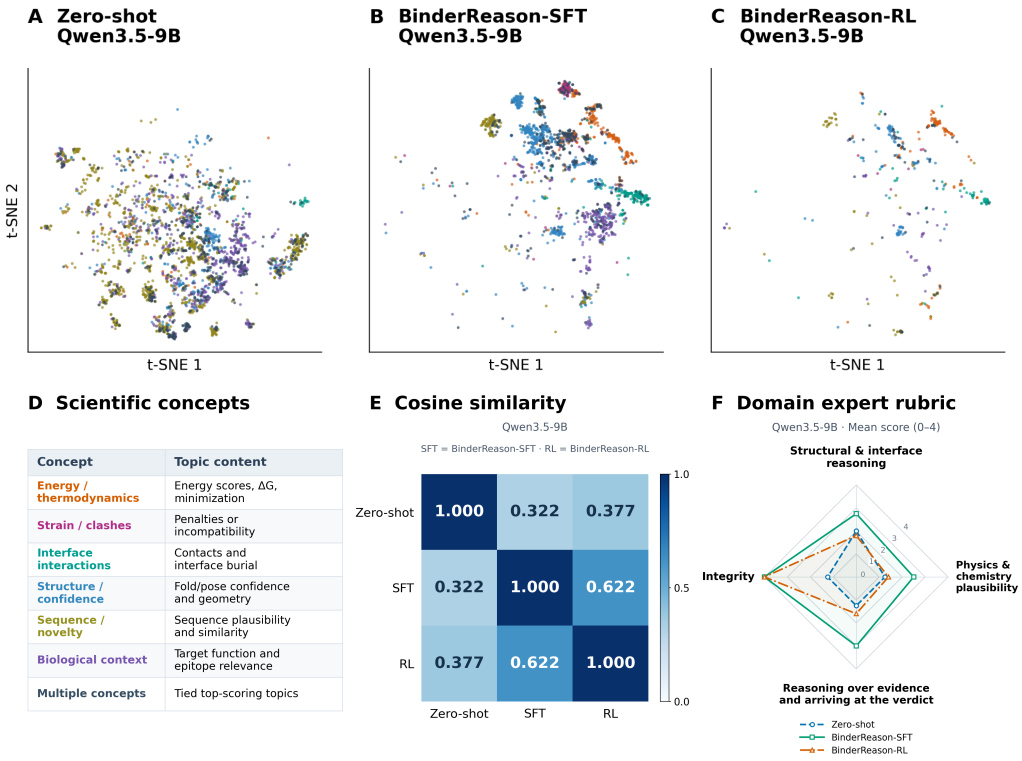


Figure 2: **Scientific concepts in BinderReason explanations before and after post-training.** (A–C) Joint t-SNE visualization of 64-dimensional TF-IDF passage embeddings from zero-shot Qwen3.5-9B, BinderReason-SFT Qwen3.5-9B, and BinderReason-RL Qwen3.5-9B on 36 matched, positive-labeled binder–target pairs. Each point represents a passage, colored using the scientific-topic rules summarized in (D). Identifiers, numerical values, verdict labels, and long sequences are removed before embedding. Unassigned passages are included in the analysis but omitted from the scatter plots. (E) Mean cosine similarity between normalized passage centroids for corresponding binder–target pairs, computed in the original 64-dimensional space and averaged equally across pairs. (F) Provides the domain expert rubric analysis across zero-shot Qwen3.5-9B, BinderReason-SFT Qwen3.5-9B, and BinderReason-RL Qwen3.5-9B

(Appendix H). SFT and RL explanations also share more lexical content with each other (mean matched cosine similarity: 0.622) than either shares with zero-shot outputs (0.322 for SFT and 0.377 for RL). This pattern supports continuity in the scientific vocabulary expressed after SFT and RL, alongside a stronger emphasis on binding energetics following post-training. These findings concern generated explanations; they do not independently validate the underlying energetic assessments.

Domain Expert Evaluation. Table 3 shows that SFT produces the strongest expert-assessed reasoning. For QWEN3.5-9B, the overall mean score increases from 0.455 zero-shot to 2.745 after SFT, before decreasing to 1.998 after RL. GEMMA-4-12B similarly improves from 0.250 zero-shot to 2.163 after SFT. Although the RL model has the strongest classification performance, its lower expert score suggests that this improvement does not correspond to more rigorous explanations.

The qualitative examples in Appendix H.2 help explain this difference. The RL trace combines relevant scientific terminology with factual errors and overconfident interpretations of contact and energetic evidence, whereas the SFT trace considers conflicting evidence more explicitly but still applies some heuristic signals too categorically. The reported numerical values are not themselves treated as hallucinations; rather, the concern is that they are used to support conclusions stronger than the available evidence warrants. Given the small number of evaluated traces, these findings are descriptive rather than statistically conclusive.

Model	Regime	Trace 1		Trace 2		Overall mean
		Expert 1	Expert 2	Expert 1	Expert 2	
Qwen3.5-9B	Zero-shot	0.33	0.50	0.66	0.33	0.455
	SFT	2.66	2.66	2.66	3.00	**2.745**
	RL	2.33	1.00	3.00	1.66	1.9975
Gemma-4-12B	Zero-shot	0	0	1	0	0.25
	SFT	2.33	1.66	2.33	2.33	2.1625

Table 3: **Domain-expert evaluation of generated reasoning traces.** Each entry is the composite rubric score assigned by one expert to one trace. The overall mean is computed across both experts and both traces. Scores range from 0 to 4; higher is better.

5 DISCUSSION

Our study investigates whether agent-generated reasoning traces, grounded in computational protein analyses, can supervise an open-weight model to predict experimental binder success. BinderReason-RL achieves the strongest aggregate classification performance on our fully held-out test set, outperforming larger closed-source models including GPT-6 Astra, Claude Opus 4.8, and Claude Opus 5. SFT produces more variable results: Gemma-4-12B exceeds GPT-6 Astra and Claude Opus 4.8 in balanced accuracy and F1, whereas Qwen3.5-9B worsens from its zero-shot initialisation. Targeted post-training can therefore outperform model scale, but its effectiveness depends on the training stage and model family.

The embedding analysis in Figure 2 clarifies the different roles of SFT and RL. The t-SNE projections show that SFT produces the largest shift towards energetic and interface-related concepts, while RL remains more closely aligned with SFT than with the zero-shot model. However, the domain-expert rubric reveals that topical alignment is not equivalent to sound scientific reasoning. SFT receives the strongest expert scores, whereas the more accurate RL model uses the relevant concepts less rigorously. The lexical analysis therefore measures whether a model discusses appropriate evidence; the rubric assesses whether it interprets and connects that evidence correctly.

Figure 19 shows why the reasoning traces are useful. Zero-shot explanations from Qwen3.5-9B, GPT-6 Astra, and Claude Opus 5 identify relevant properties but generally rely on isolated heuristics, including sequence composition, fold plausibility, and hydrophobicity (Appendix H.1). BinderLab-Think instead suggests an ordered reasoning process in which structural health, interface quality, energetics, and counter-evidence are considered together. After SFT, the model begins to distinguish structural plausibility from binding plausibility and to reason over contacts and energetic strain. RL then learns a more effective mapping from this domain-specific initialisation to the experimental

label (Figure 20, Section F). Together, these results suggest that SFT establishes the scientific reasoning scaffold, while RL primarily optimises the final classification behaviour.

However, the aggregate performance conceals substantial variation across targets. BinderReason-RL performs strongly on PD-L1 and IL-7Ra but predicts no positive binders for MZB1/PERP1 or pMHC SILSY1 (see Tables 19-22 in Appendix F), so its performance does not yet demonstrate consistent generalisation across target families. The expert evaluation covers a small set of correctly classified traces, and the embedding analysis measures lexical content rather than scientific correctness. Moreover, only correct teacher traces are retained for training, potentially favouring easier examples, while the contribution of individual BinderLab-Think components remains unknown. At inference, the model receives only binder and target names and sequences, meaning that its structural and energetic claims are inferred rather than recomputed. The annotated examples in Appendix H.2 and Figure 21 show that this can produce factual errors and overconfident interpretations.

The weaker RL reasoning may also reflect the reward design. The reward prioritises correct verdicts, numerical consistency, validity, and formatting, but does not directly reward the integration of competing scientific evidence. Future work should therefore evaluate classification and expert reasoning jointly, ablate the agentic pipeline, test additional model families and reward formulations, and validate performance prospectively. BinderReason is best viewed as a step towards a prioritisation layer before structural analysis and wet-lab testing: training on reasoning traces can transfer domain-relevant reasoning into a compact model, but preserving that reasoning while optimising classification remains an open challenge.

6 Limitations and Future Work

Despite its strong aggregate performance, BinderReason remains target-dependent and does not yet demonstrate consistent generalisation across target families. We discuss limitations we wish to address in future work below.

Ablations. The current BinderLab-Think pipeline proposes a deterministic workflow, using target literature, co-folding, design-target analysis to generate CoT reasoning traces, which in turn are used in our two-stage post-training. In future work, we want to explore the impact of each component on the quality of the training: some of the ablations considered are removing the literature component, or some of the analysis tools (e.g., Rosetta, DSSP), regenerating the traces, and retraining the models. Further, we would also like to assess the impact of non-agentic chain of thought traces.

Training. While our training setup showed promising outcomes, especially for the Qwen3.5-9B model, Gemma-4-12B was shown to overfit in the SFT part (Figure 16 from Appendix E.3). This will be addressed in future work by proper hyper-parameter tuning search.

Reinforcement Learning. BinderReason is a good example of a post-training framework for experimental binder classification. However, neither BinderReason-SFT nor BinderReason-RL consistently grounds its reasoning in scientific evidence; generated explanations may contain unsupported interpretations, inconsistent formatting, or reliance on spurious cues such as binder names (Appendix H.1). Future work will compare agentic reasoning supervision with label-only SFT, apply GRPO directly to the base model, and investigate longer training, improved reward formulations, and additional model families.

7 Conclusions

The present work introduces *BinderReason*, a two-stage post-training approach for predicting experimental binder success against therapeutically relevant targets. We first compiled a dataset of experimentally characterised binder–target pairs. We then introduced *BinderLab-Think*, a multi-agent framework that combines literature, cofolding, and protein analysis tools to generate detailed CoT traces for these complexes. These traces were used to cold-start two open-weight models (*BinderReason-SFT*) to generate structured scientific justifications. We further trained one model via GRPO (*BinderReason-RL*), substantially improving its experimental success predictions. Finally, domain-expert evaluation and concept analysis characterised how the generated explanations evolved through post-training.

AI USE STATEMENT

In this work, we used generative AI tools for figure generation and data analysis, for code debugging and for manuscript refinement - solely rewording, minor editorial polishing of author-written text (consistent with practices reported in prior work), and LaTeX formatting of figures. Specifically, we used LLM-based assistants to help produce and refine the figures reported in the main text and appendix, to assist with the analysis of experimental results (including the lexical/concept analysis of generated explanations in Appendix H), to debug training, evaluation, and analysis code via Claude Code, and to refine our written content and check for grammatical mistakes. We also use AI to screen for related work. All supervised fine-tuning and reinforcement learning experiments, including dataset construction, splitting, reward design, hyperparameter selection, and training and evaluation runs, were designed, implemented, and executed by the authors. We did not use generative AI tools for research ideation, for generating the scientific claims of this paper, or for the domain-expert evaluation, which was carried out independently by two human experts using the rubric in Appendix G. All AI-assisted code was reviewed and tested by the authors, all AI-assisted analyses were verified against the underlying result files, all figures were checked against the depicted data.

We note separately that large language models are the object of study in this work: LLM-powered agents generate the reasoning traces described in Appendix H. This constitutes the methodology of the paper rather than AI assistance to the authors, and is documented in full in the main text and Appendices A.1 and C.

We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI. See the ICLR 2027 AI Policy for Authors for more details.

ETHICS STATEMENT

This work advances AI for drug discovery, focusing on interpretable language-model-based assessment of therapeutic protein binder designs. Our system classifies existing design–target pairs rather than generating novel binders, and relies only on analysis tools and public structural data that are already openly available; we therefore believe it poses no greater risk than the publicly available protein design and cheminformatics tools it builds upon, and we encourage responsible use within established scientific and regulatory frameworks. All binder data comes from three publicly released, experimentally characterised datasets used in accordance with their terms; no proprietary or personally identifying data, human subjects, or animal experiments were involved. The qualitative evaluation was performed voluntarily by two domain experts who assessed model outputs only. Finally, our models predict experimental outcomes from computational evidence alone and are validated on few held-out targets; they are intended as a filtering aid within an expert-supervised pipeline, not as a substitute for wet-lab validation.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, we provide detailed descriptions of the experimental settings in the appendix. Data collection and its composition per-target split are described in Section 3.1 and the Appendices C.4 and E.1, and all datasets used are publicly available.

BINDERLAB-THINK is specified agent-by-agent, together with the tools, powering models, and trace-collection statistics, in Appendix C; the prompts used identically across the zero-shot, SFT, and RL settings are given in Appendix A.2. The zero-shot baseline setup is discussed in Appendix B. The BinderReason-SFT and BinderReason-RL training specifications are highlighted in Appendices E.3 and E.4, with common setup described in Appendix D.1, and all runs use the fixed random seeds stated there. Source code, the processed trace dataset, and the split files will be provided upon acceptance of the paper.

REFERENCES

Daniel Chen, Wei Lu, Amit Dhanda, and Chris Hansen. Aligning Large Language Model Agents with Rational and Moral Preferences. 2026. URL https://www.sioe.org/conference/2026/program/paper/1736-aligning-large-language-model-agents-with-rational-and-moral.

Shriram Chennakesavalu, Kirill Shmilovich, Hayley Weir, Colin A. Grambow, John Bradshaw, Patricia Suriana, Chen Cheng, and Kangway V. Chuang. Evaluating the Progression of Large Language Model Capabilities for Small-Molecule Drug Design. May 2026. URL https://openreview.net/forum?id=XJ2YJDCVCN#discussion.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training, May 2025. URL http://arxiv.org/abs/2501.17161. arXiv:2501.17161 [cs.AI].

Claude Science and Amir Shanehsazzadeh. Autonomous de novo protein binder design with Claude. 2026.

Peter Eastman, Raimondas Galvelis, Raúl P. Peláez, Charlles R. A. Abreu, Stephen E. Farr, Emilio Gallicchio, Anton Gorenko, Michael M. Henry, Frank Hu, Jing Huang, Andreas Krämer, Julien Michel, Joshua A. Mitchell, Vijay S. Pande, João PGLM Rodrigues, Jaime Rodriguez-Guerra, Andrew C. Simmonett, Sukrit Singh, Jason Swails, Philip Turner, Yuanqing Wang, Ivy Zhang, John D. Chodera, Gianni De Fabritiis, and Thomas E. Markland. OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. *The journal of physical chemistry. B*, 128(1):109–116, January 2024. ISSN 1520-6106. doi: 10.1021/acs.jpcb.3c06662. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10846090/.

Scott Emmons, Roland S. Zimmermann, David K. Elson, and Rohin Shah. A Pragmatic Way to Measure Chain-of-Thought Monitorability, October 2025. URL http://arxiv.org/abs/2510.23966. arXiv:2510.23966 [cs.LG].

Anda-Raluca Epure, Deepa Mal Korani, Jonathan G. Hedley, Maxim Secor, Alejo Nevado-Holgado, and Philip Torr. Minibinder Lab: The Reliability Gap Of Agents For Designing High Quality Protein Binders. In *Second Workshop on Agents in the Wild: Safety, Security, and Beyond*, May 2026. URL https://openreview.net/forum?id=TfGHVHLm0r.

Adibvafa Fallahpour, Arman Seyed-Ahmadi, Parsa Idehpour, Omar Ibrahim, Purav Gupta, Jack Naimer, Kevin Zhu, Arnav Shah, Shihao Ma, Abhinav Adduri, Talu Güloglu, Nuo Liu, Haotian Cui, Arihant Jain, Max de Castro, Amirfaham Fallahpour, Antonio Cembellin-Prieto, John S. Stiles, Filip Nemčko, Alexander A. Nevue, Hyungseok C. Moon, Lucas Sosnick, Olivia Markham, Haonan Duan, Michelle Y. Y. Lee, Andrea F. M. Salvador, Chris J. Maddison, Christoph A. Thaiss, Chiara Ricci-Tam, Brian S. Plosky, Dave P. Burke, Patrick D. Hsu, Hani Goodarzi, and Bo Wang. BioReason-Pro: Advancing Protein Function Prediction with Multimodal Biological Reasoning. *bioRxiv*, pp. 2026.03.19.712954, January 2026. doi: 10.64898/2026.03.19.712954. URL http://biorxiv.org/content/early/2026/03/20/2026.03.19.712954.abstract.

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-Proteina: Atomistic Protein Generation via Partially Latent Flow Matching, July 2025. URL http://arxiv.org/abs/2507.09466. arXiv:2507.09466 [cs].

GetContacts. GetContacts (https://getcontacts.github.io/). URL https://getcontacts.github.io/.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy

Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with Co-Scientist. *Nature*, May 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10644-y. URL https://doi.org/10.1038/s41586-026-10644-y.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Maria Hauser, Martin Steinegger, and Johannes Söding. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics*, 32(9):1323–1330, May 2016. ISSN 1367-4811. doi: 10.1093/bioinformatics/btw006.

Maarten L. Hekkelman, Daniel Álvarez Salmoral, Anastassis Perrakis, and Robbie P. Joosten. DSSP 4: FAIR annotation of protein secondary structure. *Protein Science*, 34(8):e70208, 2025. ISSN 1469-896X. doi: 10.1002/pro.70208. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.70208. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.70208.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe

Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. Towards a science of scaling agent systems, 2026. URL https://arxiv.org/abs/2512.08296.

Latent Labs Team, Sebastian M. Schmon, Daniella Pretorius, Simon Mathis, Rebecca Bartke-Croughan, Aishaini Puvanendran, James Vuckovic, Henry Kenlay, Mária Vlachynská, Alex Bridgland, Ivan Grishin, Sven Over, David Li, Bridget Li, Jonathan Crabbé, Agrin Hilmkil, Alexander W. R. Nelson, David Yuan, Annette Obika, and Simon A. A. Kohl. Latent-Y: A Lab-Validated Autonomous Agent for De Novo Drug Design, 2026. URL https://arxiv.org/abs/2603.29727. Version Number: 2.

Julia Koehler Leman, Brian D. Weitzner, Steven M. Lewis, Jared Adolf-Bryfogle, Nawsad Alam, Rebecca F. Alford, Melanie Aprahamian, David Baker, Kyle A. Barlow, Patrick Barth, Benjamin Basanta, Brian J. Bender, Kristin Blacklock, Jaume Bonet, Scott E. Boyken, Phil Bradley, Chris Bystroff, Patrick Conway, Seth Cooper, Bruno E. Correia, Brian Coventry, Rhiju Das, René M. De Jong, Frank DiMaio, Lorna Dsilva, Roland Dunbrack, Alexander S. Ford, Brandon Frenz, Darwin Y. Fu, Caleb Geniesse, Lukasz Goldschmidt, Ragul Gowthaman, Jeffrey J. Gray, Dominik Gront, Sharon Guffy, Scott Horowitz, Po-Ssu Huang, Thomas Huber, Tim M. Jacobs, Jeliazko R. Jeliazkov, David K. Johnson, Kalli Kappel, John Karanicolas, Hamed Khakzad, Karen R. Khar, Sagar D. Khare, Firas Khatib, Alisa Khramushin, Indigo C. King, Robert Kleffner, Brian Koepnick, Tanja Kortemme, Georg Kuenze, Brian Kuhlman, Daisuke Kuroda, Jason W. Labonte, Jason K. Lai, Gideon Lapidoth, Andrew Leaver-Fay, Steffen Lindert, Thomas Linsky, Nir London, Joseph H. Lubin, Sergey Lyskov, Jack Maguire, Lars Malmström, Enrique Marcos, Orly Marcu, Nicholas A. Marze, Jens Meiler, Rocco Moretti, Vikram Khipple Mulligan, Santrupti Nerli, Christoffer Norn, Shane Ó'Conchúir, Noah Ollikainen, Sergey Ovchinnikov, Michael S. Pacella, Xingjie Pan, Hahnbeom Park, Ryan E. Pavlovicz, Manasi Pethe, Brian G. Pierce, Kala Bharath Pilla, Barak Raveh, P. Douglas Renfrew, Shourya S. Roy Burman, Aliza Rubenstein, Marion F. Sauer, Andreas Scheck, William Schief, Ora Schueler-Furman, Yuval Sedan, Alexander M. Sevy, Nikolaos G. Sgourakis, Lei Shi, Justin B. Siegel, Daniel-Adriano Silva, Shannon Smith, Yifan Song, Amelie Stein, Maria Szegedy, Frank D. Teets, Summer B. Thyme, Ray Yu-Ruei Wang, Andrew Watkins, Lior Zimmerman, and Richard Bonneau. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods*, 17(7):665–680, July 2020. ISSN 1548-7105. doi: 10.1038/s41592-020-0848-2. URL https://www.nature.com/articles/s41592-020-0848-2.

Hao Li, He Cao, Shenyao Peng, Zijing Liu, Bin Feng, Yu Wang, Zhiyuan Yan, Yonghong Tian, Yu Li, and Li Yuan. Agentic reinforcement learning empowers next-generation chemical language models for molecular design and synthesis, January 2026. URL http://arxiv.org/abs/2601.17687. arXiv:2601.17687 [cs.LG].

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-Like Training: A Critical Perspective, October 2025. URL http://arxiv.org/abs/2503.20783. arXiv:2503.20783 [cs.LG].

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, May 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00832-8. URL https://doi.org/10.1038/s42256-024-00832-8.

Sarine Markossian, Hannah M. Baskir, Abigail C. Grossman, Suchismita Chandran, Shyam Rele, Matthew D. Hall, Rommie E. Amaro, Alexander Tropsha, and Alexey V. Zakharov. In silico Drug Discovery: Bridging the Gaps in Preclinical Translation. *Drug discovery today*, 31(1): 104567, January 2026. ISSN 1359-6446. doi: 10.1016/j.drudis.2025.104567. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC13061093/.

Max D. Overath, Andreas Rygaard, Christian P. Jacobsen, Valentas Brasas, Oliver Morell, Pietro Sormanni, and Timothy P. Jenkins. Predicting Experimental Success in De Novo Binder Design: A Meta-Analysis of 3,766 Experimentally Characterised Binders, August 2025. URL https://www.biorxiv.org/content/10.1101/2025.08.14.670059v1. ISSN: 2692-8205 Pages: 2025.08.14.670059 Section: New Results.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv: The Preprint Server for Biology*, art. 2025.06.14.659707, June 2025. ISSN 2692-8205. doi: 10.1101/2025.06.14.659707.

David A. Pearlman, David A. Case, James W. Caldwell, Wilson S. Ross, Thomas E. Cheatham, Steve DeBolt, David Ferguson, George Seibel, and Peter Kollman. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1):1–41, September 1995. ISSN 0010-4655. doi: 10.1016/0010-4655(95)00041-D. URL https://www.sciencedirect.com/science/article/pii/001046559500041D.

Manvitha Ponnapati, Sam Cox, Cade W. Gordon, Michael J. Hammerling, Siddharth Narayanan, Jon M. Laurent, James D. Braza, Michaela M. Hinks, Michael D. Skarlinski, Samuel G. Rodriques, and Andrew White. ProteinCrow: A Language Model Agent That Can Design Proteins. July 2025. URL https://openreview.net/forum?id=ljXgWDtqCu.

ProteinBase. Proteinbase - The home of protein design data. URL https://proteinbase.com.

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, Mateo Reveiz, Roman Bushuiev, Tomáš Pluskal, Josef Sivic, Karsten Kreis, Arash Vahdat, Shamayeeta Ray, Jonathan T. Goldstein, Andrew Savinov, Jacob A. Hambalek, Anshika Gupta, Diego A. Taquiri-Diaz, Yaotian Zhang, A. Katherine Hatstat, Angelika Arada, Nam Hyeong Kim, Ethel Tackie-Yarboi, Dylan Boselli, Lee Schnaider, Chang C. Liu, Gene-Wei Li, Denes Hnisz, David M. Sabatini, William F. DeGrado, Jeremy Wohlwend, Gabriele Corso, Regina Barzilay, and Tommi Jaakkola. BoltzGen: Toward Universal Binder Design, November 2025. URL https://www.biorxiv.org/content/10.1101/2025.11.20.689494v1. ISSN: 2692-8205 Pages: 2025.11.20.689494 Section: New Results.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, November 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL https://www.nature.com/articles/nbt.3988.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723, October 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09442-9.

Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks, October 2024. URL http://arxiv.org/abs/2410.20266. arXiv:2410.20266 [cs.HC].

Latent Labs Team, Henry Kenlay, Daniella Pretorius, Jonathan Crabbé, Alex Bridgland, Sebastian M. Schmon, Agrin Hilmkil, James Vuckovic, Simon Mathis, Tomas Matteson, Rebecca Bartke-Croughan, Amir Motmaen, Robin Rombach, Mária Vlachynská, Alexander W. R. Nelson, David Yuan, Annette Obika, and Simon A. A. Kohl. Drug-like antibodies with low immunogenicity in human panels designed with Latent-X2, December 2025. URL http://arxiv.org/abs/2512.20263. arXiv:2512.20263 [q-bio.BM].

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL https://www.nature.com/articles/s41586-023-06415-8.

Stanley Wei and Juno Kim. Provable Benefits of RLVR over SFT for Reasoning Models: Learning to Backtrack Efficiently, June 2026. URL http://arxiv.org/abs/2606.22938. arXiv:2606.22938 [cs.LG].

Renjun Xu and Yang Yan. Agent skills for large language models: Architecture, acquisition, security, and the path forward, 2026. URL https://arxiv.org/abs/2602.12430.

Vinicius Zambaldi, David La, Alexander E. Chu, Harshnira Patani, Amy E. Danson, Tristan O. C. Kwan, Thomas Frerix, Rosalia G. Schneider, David Saxton, Ashok Thillaisundaram, Zachary Wu, Isabel Moraes, Oskar Lange, Eliseo Papa, Gabriella Stanton, Victor Martin, Sukhdeep Singh, Lai H. Wong, Russ Bates, Simon A. Kohl, Josh Abramson, Andrew W. Senior, Yilmaz Alguel, Mary Y. Wu, Irene M. Aspalter, Katie Bentley, David L. V. Bauer, Peter Cherepanov, Demis Hassabis, Pushmeet Kohli, Rob Fergus, and Jue Wang. De novo design of high-affinity protein binders with AlphaProteo, September 2024. URL http://arxiv.org/abs/2409.08022. arXiv:2409.08022 [q-bio].

APPENDIX

A GENERAL SETUP

A.1 LARGE LANGUAGE MODEL SETUP

To power our work, we made use of a series of large language models, as described in the following table:

Table 4: Overview of LLMs used to power the agentic entities.

Model	**Provider**	**Identifier**	**#Params**	**Context**
Qwen3.5-9B	Qwen	Qwen/Qwen3.5-9B	9B	262K-1M
Qwen3.6-27B	Qwen	Qwen/Qwen3.6-27B	27B	262K-1M
Qwen3.6-35B-A3B	Qwen	Qwen/Qwen3.6-35B-A3B	35B	262K-1M
Qwen3.8-27B	Qwen	Qwen/Qwen3.8-27B	27B	262K-1M
GPT-OSS-20B	OpenAI	openai/gpt-oss-20b	21B	$\sim$128K
GPT-OSS-120B	OpenAI	openai/gpt-oss-120b	117B	$\sim$128K
Gemma-4-12B	Google	google/gemma-4-12B	11.95B	256K
Gemma-4-31B-Instr	Google	google/gemma-4-31B-it	30.7B	256K
DeepSeek-V4-Flash	DeepSeek	deepseek-ai/DeepSeek-V4-Flash-0731	284B	1M
Claude Opus 5	Anthropic	-	-	1M
Claude Opus 4.8	Anthropic	-	-	1M
GPT-6 Astra	OpenAI	-	-	$\sim$1M
GPT-5.6 Sol	OpenAI	-	-	$\sim$1M
Gemini 3.8 Flash	Google	-	-	$\sim$1M

Note: – denotes information that is not publicly specified.

We employ the following standard setup when running these models, unless otherwise stated:

Table 5: Model inference configurations

Model	**Temperature**	**Top-p**	**Top-k**	**Min-p**	**Repetition Penalty**	**Presence Penalty**	**Enable Thinking**	**Reasoning Effort**
Gemma-4-12B	1.0	0.95	64	–	–	–	Yes	–
Gemma-4-31B-Instruct	1.0	0.95	64	–	–	–	–	–
Qwen3.5-9B	1.0	0.95	20	0.0	1.0	1.5	–	–
Qwen3.6-27B	1.0	0.95	20	0.0	1.0	0.0	-	–
Qwen3.6-35B-A3B	1.0	0.95	20	0.0	1.0	1.5	-	–
Qwen3.8-27B	1.0	0.95	20	0.0	1.0	0.0	–	–
GPT-OSS-20B	1.0	1.0	0	–	–	–	–	–
GPT-OSS-120B	1.0	1.0	0	–	–	–	–	–
DeepSeek-V4-Flash	1.0	0.95	–	–	–	–	Yes	Max
Claude Opus 5	1.0	–	–	–	–	–	–	high
Claude Opus 4.8	1.0	–	–	–	–	–	–	high
GPT-6 Astra	1.0	–	–	–	–	–	–	high
GPT-5.6 Sol	1.0	–	–	–	–	–	–	high
Gemini 3.8 Flash	1.0	–	–	–	–	–	–	high

Note: – denotes information that is set by default.

A.2 PROMPTS

The system and user prompts shown in Figures 3 and 4 are similar across the zero-shot, SFT, and RL runs.

B ZERO-SHOT

To evaluate the quality of the proposed 2-stage post-training approach, apart from the standard binary classification metrics for an imbalanced dataset (accuracy, balanced accuracy, precision, recall, F1 score), we would like to introduce a **zero-shot baseline**. We take 5 closed-source flagship reasoning models (CLAUDE OPUS 5, CLAUDE OPUS 4.8, GPT-6 ASTRA, GPT-5.6 SOL,

System prompt

You are the Binder Quality Expert, an expert in taking a design and target and identifying if the design is good or bad. The design's sequence is fixed and was tested in a wet lab; your job is to predict that experimental outcome based on the given information, any analysis you wish to run, or any information you are able to infer or find through own analysis or research. You are given one (binder, target) pair, and you decide whether it is a good binder or a bad binder.

Figure 3: System prompt.

User prompt

Evaluate this binder {binder_id}, with corresponding sequence {binder_sequence}, against its target {target_id}, with corresponding sequence {target_sequence}, using the given information and any information you are able to infer or find through own analysis or research.
Reason over it in <think></think>, then output either <answer>good binder</answer> or <answer>bad binder</answer>.

Figure 4: User prompt.

GEMINI 3.8 FLASH) and 3 strong open-weight reasoning models (GPT-OSS-20B, QWEN3.8-27B, DEEPSEEK-V4-FLASH) in addition to the 2 used in the post-training (i.e. QWEN3.5-9B, GEMMA-4-12B). Then, we provide them with the same *user* and *system* prompts we use in the post-training procedure, described in Appendix A.2. The models are asked to perform the same binary experimental success prediction task, without any prior training. Configuration details are left in Table 5; in addition, we mention that the maximum length of the reasoning trace is 20000, with 3000 max new tokens per trace. Experiments are repeated 3 times, with 3 distinct random seeds: 42, 1234, 123 - the same seeds will be used in the inference runs as well.

C BINDERLAB-THINK

C.1 ARCHITECTURE OVERVIEW.

BinderLab-Think consists of 11 highly specialised agents, each with a distinct function, which we describe in more detail in the following subsections.

For the purpose of this work, we experimented with the "Agent Skills" (Xu & Yan, 2026) setup proposed in https://platform.claude.com/docs/en/agents-and-tools/agent-skills/overview. This is an alternative way of approaching the prompting process of our agents, promoting modularity and the convergence of our agentic entities from general-purpose models to AI domain-experts. In addition, skills only load on demand, when one agent is activated in the workflow, and do not require the same guidance to be repeated, leading to a more optimal instruction process and the ability of our agents to achieve their functions and contribute to the multi-agentic workflow.

C.2 AGENTIC SETUP

C.2.1 SCHOLAR AGENT

The **"Scholar" Agent** runs only at the beginning of the workflow, constructing structured biological briefs for all the target proteins by querying UniProt (function, domains, annotated binding sites), the RCSB PDB (co-crystal discovery and atomic-coordinate–derived interface residues at small atomic distance between carbon atoms), and Europe PMC (mutagenesis data, therapeutic-antibody epitopes, mechanism of action). The resulting JSON briefs — comprising published epitopes, functional hotspots, non-productive surfaces, and actionable literature guidance with residue-level citations — are saved in the *Laboratory Notebook* and reused as a fixed prior by the *Forecaster* Agent in all subsequent design iterations, avoiding inefficient per-iteration literature retrieval.

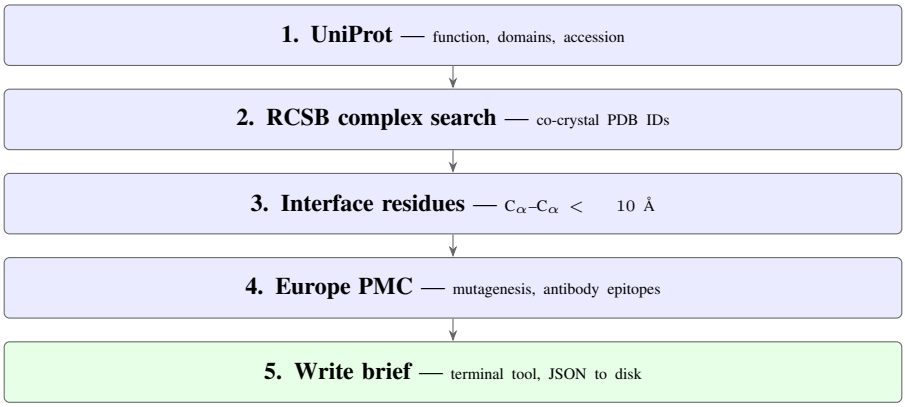


Figure 5: **Scholar Agent (SCHOLAR).** A one-shot, campaign-level scout that builds structured biological briefs before iterative binder-target analysis for experimental outcome prediction. The agent calls four evidence-gathering tools in sequence (UniProt → RCSB complex search → interface residues → Europe PMC) and terminates with a single `write_literature_brief` call. The resulting JSON briefs contain residue-level epitopes, hotspots, and design advice.

Table 6: Tools available to the Scholar Agent (SCHOLAR). All four research tools are backed by free, keyless REST APIs; results are cached on disk so reruns and intra-run duplicates pay zero HTTP cost. The fifth tool is terminal — calling it ends the agent run and delivers the output: a literature brief based on the target of interest.

Tool	Source	Purpose
`uniprot_search`	UniProtKB (`uniprot.org/help/uniprotkb`)	Resolve target name to accession. Fetch function, domains, subcellular location, and annotated binding sites.
`rcsbcomplex_search`	RCSB PDB (`rcsb.org`)	Find co-crystal / cryo-EM structures containing the target in a multi-chain complex (queries by UniProt accession).
`rcsbinterface_residues`	RCSB PDB (`rcsb.org`)	Compute interface residues between two chains of a PDB entry (CA–CA distance < 10 Å). Provide ground-truth contact residues from atomic coordinates.
`pubmed_search`	Europe PMC (`europepmc.org`)	Search peer-reviewed literature for mutagenesis data, therapeutic antibody epitopes, and mechanism of action (returns PMID + abstract).
`write_literature_brief`	Local Disk	*Terminal tool.* Persist the structured JSON brief (residues, hotspots, design advice) to the path supplied by the Orchestrator and end the run.

C.2.2 FOLDSMITH AGENT

The **"Foldsmith" Agent** uses the binder-target sequences to cofold and obtain the structure of the complex. It also computes metrics related to the stability of the folded protein structure.

This agent builds on the state-of-the-art software in protein structure design, BoltzGen (Stark et al., 2025). Upon receiving two protein sequences, for a design and target, respectively, the tool acts as a practical, end-to-end pipeline for the *in silico* 3D structure prediction and analysis. We chose to employ BoltzGen due to its ability to perform all the pipeline steps in a single run, without any prompting or user intervention. Furthermore, it also provides the user with a clear breakdown of significant metrics for *filtering* (e.g., RMSD-based metrics, presence of certain classes of amino acids, percentage of a specific amino acid in the structure) and scoring (e.g. AlphaFold (Jumper et al., 2021)-based metrics such as ipTM, pTM for alignment between design and target, folding confidence, PAE score, number of hydrogen bonds, ionic connections etc.), which will be used in the further used in the rewards system defined.

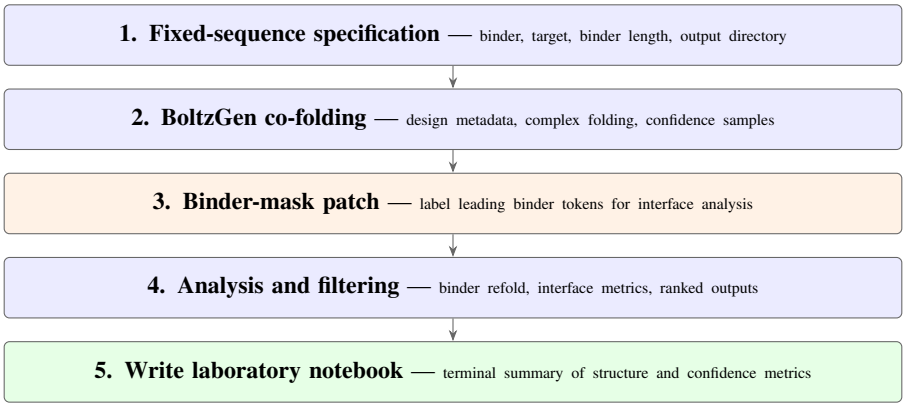


Figure 6: **Foldsmith Agent (FOLDSMITH).** A fixed-sequence co-folding agent that predicts the structure of a supplied binder–target complex using BoltzGen (Stark et al., 2025). Foldsmith calls `boltzgen_cofold` once, which performs complex folding, labels the binder residues required for interface analysis, and runs binder refolding, analysis, and filtering. It then terminates with one `write_lab_notebook` call summarising the resulting structure and confidence metrics. Because both amino-acid sequences are fixed, this stage predicts their complex rather than designing or mutating either sequence.

Table 7: Tools available to the Foldsmith Agent (FOLDSMITH). The BoltzGen (Stark et al., 2025) tool preserves the supplied binder and target sequences and reports complex-level confidence metrics across diffusion samples. The notebook writer is terminal: record the interpretation and end the run.

Tool	Source	Purpose
`boltzgen_cofold`	BoltzGen (Stark et al., 2025)	Predict a fixed binder–target complex from a YAML specification. Run complex folding, patch the design mask to identify the binder, and perform binder refolding, interface analysis, and filtering. Return JSON containing the complex CIF path, metrics-table path, ranked-output directory, executed steps, and confidence distributions.
`write_lab_notebook`	Local Disk	*Terminal tool.* Persist a concise interpretation of the co-folding result. Report `iptm` as binder–target interface confidence, `ptm` as overall fold confidence, `complex_plddt` and `complex_iplddt` as overall and interface atomic confidence, and `complex_pde` and `complex_ipde` as distance-error estimates, where lower values are better. Describe the spread across diffusion samples and omit unavailable design-relative metrics.

C.2.3 PREPPER AGENT

This agent performs preprocessing functionalities required for the generated structures to undergo further analysis: CIF-to-PDB conversion, hydrogen addition, energy minimisation.

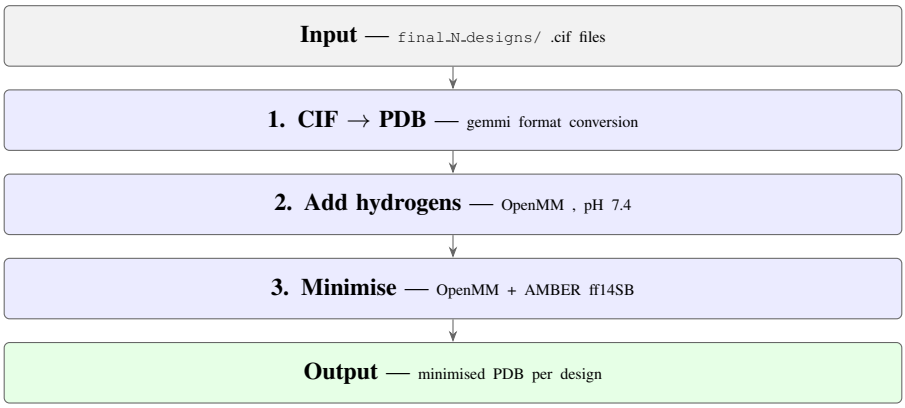


Figure 7: **Prepper Agent (PREPPER).** A single-tool agent that prepares BoltzGen-generated structures for downstream analysis. The agent makes exactly one call to `preprocessing_tool`, which internally executes three sequential stages: format conversion (gemmi), pH-aware hydrogen addition (OpenMM (Eastman et al., 2024)), and restrained vacuum energy minimisation (OpenMM with AMBER (Pearlman et al., 1995)). The tool iterates over every file in the subfolder produced by BoltzGen (Stark et al., 2025). Intermediate files (converted PDB, hydrogenated PDB) are deleted after minimisation, leaving only the final minimised PDB per design in-place. Per-stage parameters in Table 8.

Table 8: Internal stages of `preprocessing_tool` called by the Prepper Agent (PREPPER). All three stages run within a single tool invocation; only the final minimised PDB is retained on disk. Heavy-atom positional restraints keep the global structure near the BoltzGen prediction while allowing local cleanup of newly-placed hydrogens and strained sidechain conformations — a fast physical sanity check before downstream contact analysis.

Stage	Library	Purpose
`cif_to_pdb`	gemmi	Parse the mmCIF hierarchy (models → chains → residues → atoms) and serialise it to PDB records (`ATOM`, `HETATM`, `TER`), preserving Cartesian coordinates $\mathbf{r}_i = (x_i, y_i, z_i)$ for every atom. Pure format conversion; geometry is unchanged.
`add_hydrogens`	OpenMM (Eastman et al., 2024)	Infer hydrogen positions from atom types, residue templates, and pH-aware protonation heuristics (default pH = 7.4). Hydrogens are placed to satisfy idealised bond lengths and angles around each heavy atom. Falls back to OpenMM's built-in heuristics if force-field-guided placement fails.
`minimise structure`	OpenMM (Eastman et al., 2024), AMBER (Pearlman et al., 1995)	Restrained vacuum energy minimisation $\mathbf{r}^* = \arg\min_{\mathbf{r}} \left[U(\mathbf{r}) + \sum_i k\|\mathbf{r}_i - \mathbf{r}_{i,0}\|^2\right]$ with $k = 10^3$ kJ/(mol nm^2) on heavy atoms (defaults: `NoCutoff`, `HBonds` constraints, $\leq$ 500 iterations, tolerance 10 kJ/(mol nm)). Removes steric clashes and poor hydrogen placement without perturbing the predicted fold or interface.
`preprocessing tool`	Local	*Wrapper tool.* Iterate over every generated structure file in the folder, run the three stages above sequentially per file, delete intermediates, and return a per-design dictionary with stage-level metadata (energy before/after, hydrogens added, restrained-atom count, force-field mode).

C.2.4 REFEREE AGENT

The **"Referee" Agent** should act as a world-class expert in sequence similarity search, homology detection, and protein design interpretation. The role it has in this pipeline is to run MMseqs2 (Hauser et al., 2016; Steinegger & Söding, 2017) across the preprocessed binder-target 3D complexes and to return a per-design novelty / plausibility profile that the Forecaster can use.

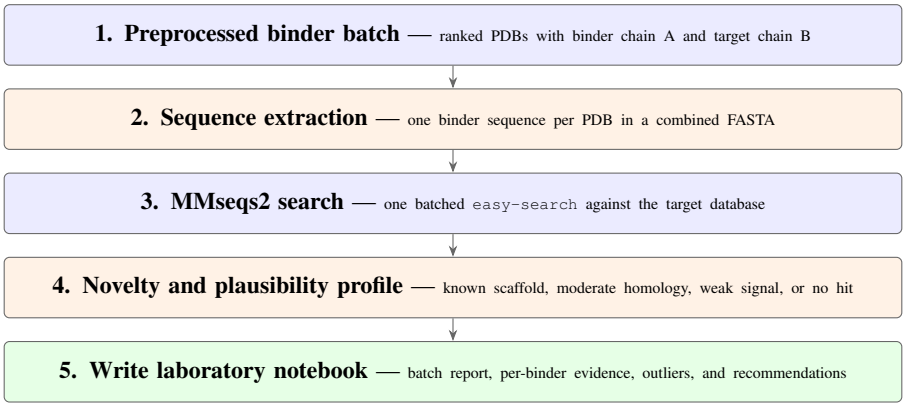


Figure 8: **Referee Agent (REFEREE).** A sequence-similarity evaluator that assesses the biological plausibility and novelty of preprocessed binder designs after A2 and before downstream analysis agents are selected. Referee extracts binder chain A from each ranked PDB and calls `mmseqs_tool` once, normally in batched `search_directory` mode. MMseqs2 (Hauser et al., 2016; Steinegger & Söding, 2017) evaluates every binder as an independent query while executing a single database search. Referee interprets the resulting identity, coverage, alignment length, E-value, and hit multiplicity, then terminates with one `write_lab_notebook` call. Sequence similarity is treated as a biological prior and not as direct evidence of folding, structural equivalence, or binding affinity.

Table 9: Tools available to the Referee Agent (REFEREE). The MMseqs2 tool produces per-binder similarity evidence and structured novelty classifications. The notebook writer is terminal: calling it persists Referee's scientific synthesis for the Orchestrator and Forecaster and ends the agent run.

Tool	Source	Purpose
`mmseqs_tool`	MMseqs2 (Hauser et al., 2016; Steinegger & Söding, 2017)	Search binder sequences against a configured sequence database using MMseqs2 `easy-search`. In `search_directory` mode, extract chain A from every preprocessed PDB, create one combined multi-entry FASTA, and perform a single batched search while retaining per-binder attribution. Write `query.fasta`, `result.m8`, and `search_summary.json`. Classify each query as `known_scaffold`, `moderate_homology`, `weak_signal`, or `no_hit` using E-value, identity, coverage, alignment length, and hit consistency.
`write_lab_notebook`	Local Disk	*Terminal tool.* Persist the batch-level class counts, recommendation, known-scaffold and no-hit outliers, and notable per-binder report blocks. Record best-hit E-value, sequence identity, alignment length, query coverage, hit multiplicity, plausibility, novelty, interpretation, and an actionable recommendation. Explicitly flag no-hit sequences for corroboration rather than treating them as automatically desirable, then end the Referee run.

C.2.5 DISTOGRAM CREATOR AGENT

The **"Distogram Creator" Agent** offers insight into the compact structural fingerprint of protein geometry, capturing fold shape and compactness, revealing chain-chain proximity patterns, and therefore being used for comparing generated binders across candidates.

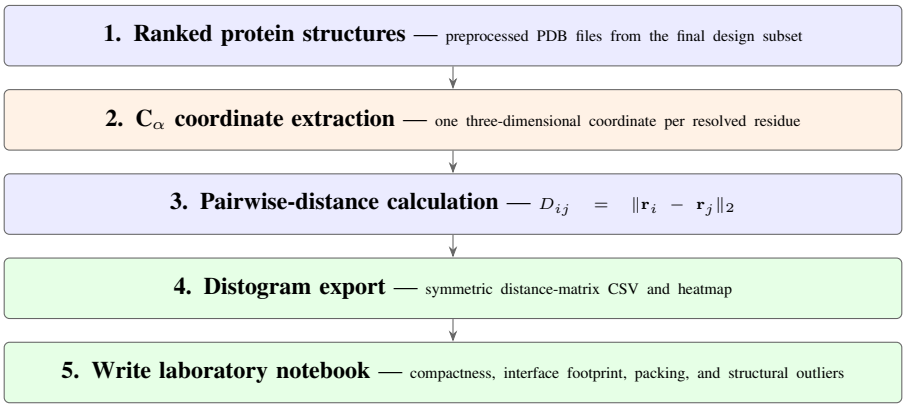


Figure 9: **Distogram Creator Agent (DISTOGRAM CREATOR).** A structural-geometry agent that computes C_α–C_α distance matrices for the final ranked protein designs produced by BoltzGen and preprocessed for analysis. Distogram Creator calls `c_alpha_distogram_tool` once to extract residue coordinates, calculate the complete pairwise Euclidean-distance matrix, and export both machine-readable CSV files and heatmaps. It then interprets intrachain blocks, interchain distance patterns, compactness, contact localization, and possible non-participating segments before terminating with one `write_lab_notebook` call. The resulting distograms provide compact structural fingerprints for geometry assessment, candidate comparison, and diversity analysis.

C.2.6 RAMACHANDRAN CREATOR AGENT

The **"Ramachandran Creator" Agent** gives a local stereochemical quality filter that complements global metrics, helping reject candidates with unrealistic backbone geometry even if interface contacts appear good, being a classic structural quality control technique.

Table 10: Tools available to the Distogram Creator Agent (DISTOGRAM CREATOR). The distogram tool converts three-dimensional protein coordinates into residue-level distance matrices and heatmaps. The notebook writer is terminal: calling it records the scientific interpretation and ends the agent run.

Tool	Source	Purpose
`c_alpha_distogram_tool`	PDB coordinates	Process the preprocessed PDB files in the final ranked-design directory. Extract every available C_α coordinate and compute the symmetric $N \times N$ Euclidean-distance matrix $D_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|_2$. Export each matrix as a CSV file and a heatmap for quantitative analysis and visual inspection. The intrachain diagonal blocks describe binder and target geometry, while the off-diagonal blocks describe their relative proximity and interface footprint.
`write_lab_notebook`	Local Disk	*Terminal tool.* Persist a concise interpretation of the generated distograms. Report residue counts and distance ranges, assess structural compactness and interchain packing, distinguish localized from diffuse binder-contact footprints, identify possible non-participating tails or segments, and note meaningful geometric differences or near-duplicate patterns across candidates before ending the Distogram Creator run.

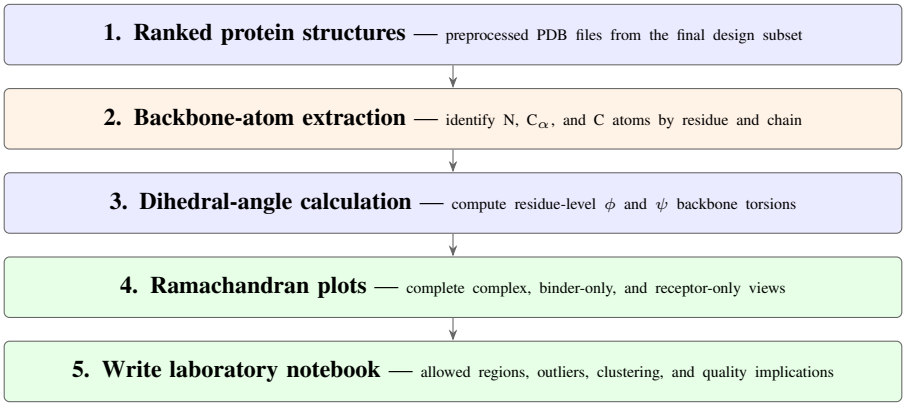


Figure 10: **Ramachandran Creator Agent (RAMACHANDRAN CREATOR).** A stereochemical-validation agent that evaluates local backbone geometry in the final ranked protein designs produced by BoltzGen. Ramachandran Creator calls `ramachandran_tool` once to extract backbone atoms, calculate residue-level ϕ and ψ dihedral angles, and generate plots for the complete complex and its binder and receptor components. It then assesses whether torsions occupy expected conformational regions, identifies isolated or clustered outliers, and distinguishes binder-specific defects from receptor geometry before terminating with one `write_lab_notebook` call. This local quality-control signal complements global fold and interface metrics by revealing strained conformations that can remain hidden in an otherwise plausible structure.

Table 11: Tools available to the Ramachandran Creator Agent (RAMACHANDRAN CREATOR). The Ramachandran tool evaluates residue-level backbone torsions and produces chain-aware conformational plots. The notebook writer is terminal: calling it records the stereochemical interpretation and ends the agent run.

Tool	Source	Purpose
`ramachandran_tool`	PDB coordinates	Process the preprocessed PDB files in the final ranked-design directory. For residue i, calculate ϕ_i from C_{i-1}–N_i–$C_{\alpha i}$–C_i and ψ_i from N_i–$C_{\alpha i}$–C_i–N_{i+1}. Map the torsions into the conventional angular range and generate Ramachandran plots for all chains, the binder, and the receptor. These views separate global conformational trends from chain-specific backbone strain.
`write_lab_notebook`	Local Disk	*Terminal tool.* Persist a concise interpretation of the Ramachandran results. Report the number of residues with defined torsions, the dominant conformational regions, and any residues in unusual or disallowed regions. Note whether outliers cluster at termini, in loops, or near the binder–target interface; consider residue context such as glycine or proline before treating unusual angles as defects; and explain the implications for local stereochemical quality before ending the Ramachandran Creator run.

C.2.7 GETCONTACTS EXPERT AGENT

The **"GetContacts Expert" Agent** explains why a binder may score well or poorly, supporting targeted redesign, and enabling objective interface metrics by employing GetContacts GetContacts, an important interface interpretability tool.

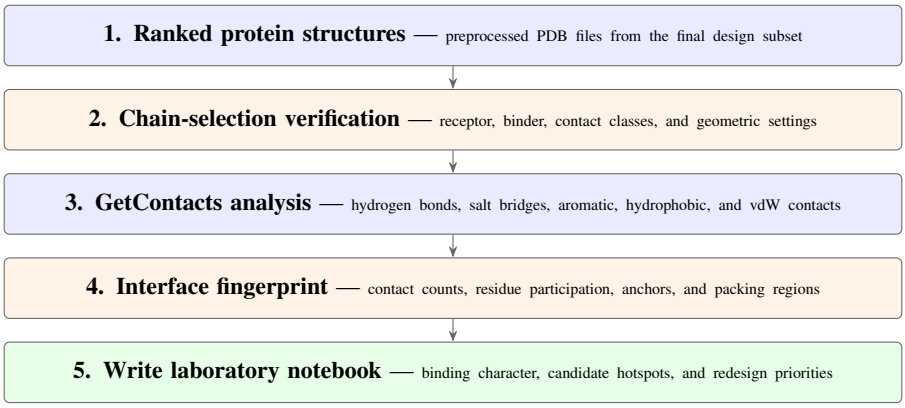


Figure 11: **GetContacts Expert Agent (GETCONTACTS EXPERT).** A protein-interface analysis agent that characterises non-covalent interactions in the final ranked binder–target structures. GetContacts Expert calls `getcontacts_tool` once to apply geometry- and atom-type-based rules for hydrogen bonds, salt bridges, aromatic interactions, hydrophobic contacts, and van der Waals contacts. It verifies the analysed chain selections and settings, separates interchain interface contacts from intrachain contacts, and identifies directed anchors and packing-dominated regions. The agent then terminates with one `write_lab_notebook` call that records the interaction fingerprint, participating residues, and structurally grounded optimisation opportunities. Contact counts are interpreted qualitatively and are not treated as direct binding free energies.

Table 12: Tools available to the GetContacts Expert Agent (GETCONTACTS EXPERT). The contact tool generates atom-level interaction records and the metadata needed to verify what was analysed. The notebook writer is terminal: record the interface interpretation and end the agent run.

Tool	Source	Purpose
`getcontacts_tool`	GetContacts (GetContacts)	Analyse the preprocessed PDB files in the final ranked-design directory using rule-based geometric and atom-type criteria. Detect backbone–backbone, sidechain–backbone, and sidechain–sidechain hydrogen bonds (`hbbb`, `hbsb`, `hbss`); salt bridges (`sb`); pi–cation, pi-stacking, and T-stacking interactions (`pc`, `ps`, `ts`); hydrophobic contacts (`hp`); and van der Waals contacts (`vdw`). Write a `*_contacts.tsv` interaction file and a `*_getcontacts_cmd.json` metadata file for each structure.
`write_lab_notebook`	Local Disk	*Terminal tool.* Persist a concise interface summary after verifying the chain selections and analysis settings from the command metadata. Report total, interchain, and relevant intrachain contact counts; provide the contact-type breakdown; identify the receptor and binder residues with the highest interface participation; and distinguish directed-contact anchors (`sb`, hydrogen bonds, `pc`, `ps`, `ts`) from packing-dominated regions (`hp`, `vdw`). Describe the interface as packing-dominated, electrostatically anchored, hydrophobic, or aromatically supported without inferring binding affinity from contact counts alone, then end the GetContacts Expert run.

C.2.8 DSSP EXPERT AGENT

This agent adds structural-context annotations that help interpret other metrics, since the secondary structure is a key stability signal, by employing the corresponding open-source DSSP Hekkelman et al. (2025) tool.

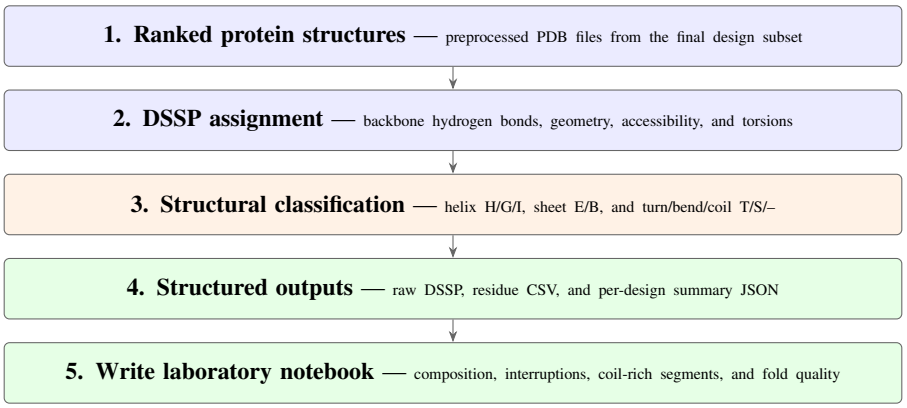


Figure 12: **DSSP Expert Agent (DSSP EXPERT).** A secondary-structure analysis agent that annotates the final ranked binder–target structures using DSSP (Hekkelman et al., 2025). DSSP Expert calls `dssp_tool` once to assign residue-level secondary structure from backbone geometry and hydrogen-bond patterns and to report solvent accessibility and backbone torsions. It groups the detailed DSSP states into helix, sheet, and turn/bend/coil categories, compares composition across chains, and identifies long unstructured segments, isolated bridges, and helix or strand interruptions. The agent terminates with `write_lab_notebook` - records the structural interpretation and whether each binder's secondary-structure architecture is consistent with the intended design.

Table 13: Tools available to the DSSP Expert Agent (DSSP EXPERT). The DSSP tool produces standard per-residue secondary-structure annotations and aggregated composition summaries. The notebook writer is terminal: calling it records the scientific interpretation and ends the agent run.

Tool	Source	Purpose
`dssp_tool`	DSSP (Hekkelman et al., 2025)	Analyse every preprocessed PDB in the final ranked-design directory. Assign residue-level secondary structure from backbone hydrogen-bond patterns and geometry, and report associated descriptors such as solvent accessibility and ϕ/ψ torsions. Group DSSP states into helices (`H/G/I`), strands or bridges (`E/B`), and turns, bends, or coil (`T/S/--`). For each structure, write the raw `*_structure.dssp` output, a `*_dssp_residues.csv` residue table, and a `*_dssp_summary.json` composition summary.
`write_lab_notebook`	Local Disk	*Terminal tool.* Persist a concise interpretation based on both the summary JSON and residue-level CSV. Report overall and chain-specific helix, sheet, and coil fractions; identify long coil-rich regions, isolated beta bridges, strand breaks, and helix interruptions; and note residue-level features that may destabilise the observed architecture. Explain whether the secondary-structure profile is consistent with the expected minibinder fold and flag candidates requiring additional structural scrutiny before ending the DSSP Expert run.

C.2.9 ROSETTA EXPERT AGENT

The **"Rosetta Expert" Agent** employs the PyRosetta (version of Rosetta Leman et al. (2020) implemented in Python) tool to give a fast, physically motivated proxy for structural plausibility (clashes, rotamer strain, backbone strain), interaction quality (electrostatics, H-bonds, solvation, packing) and interface strength(ΔG-like interface energies, buried surface area, unsatisfied H-bonds), complementing the geometric quality control methods performed by the other *Analysis Agents* (distograms, Ramachandran plots, DSSP) and contact fingerprints (GetContacts) with an energy-based signal.

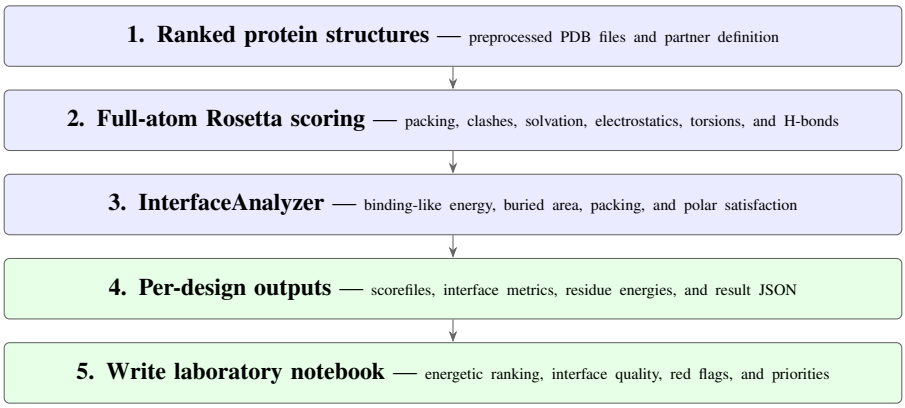


Figure 13: **Rosetta Expert Agent (ROSETTA EXPERT).** A physics-based scoring agent that evaluates the final ranked binder–target structures using PyRosetta (Leman et al., 2020). Rosetta Expert calls `pyrosetta_scoring_tool` once to process every PDB in the input directory, calculate a decomposed full-atom Rosetta score, and run InterfaceAnalyzer for the supplied partner definition. The resulting metrics describe global structural strain, steric packing, solvation, electrostatics, hydrogen bonding, interface burial, polar satisfaction, and shape complementarity. The agent then terminates with one `write_lab_notebook` call that records the principal energetic strengths, red flags, and redesign priorities for each structure. Rosetta energies are treated as comparative computational proxies rather than direct experimental binding free energies.

Table 14: Tools available to the Rosetta Expert Agent (ROSETTA EXPERT). The PyRosetta (Leman et al., 2020) tool provides global full-atom energy decomposition and interface-specific diagnostics for every ranked design. The notebook writer is terminal: calling it records the scientific interpretation and ends the agent run.

Tool	Source	Purpose
`pyrosetta_scoring_tool`	PyRosetta (Leman et al., 2020)	Score every preprocessed PDB in the final ranked-design directory using the Rosetta full-atom energy function, $E(\mathbf{r}) = \sum_k w_k E_k(\mathbf{r})$, and analyse the specified partner interface. Report global terms for attraction and repulsion, solvation, electrostatics, hydrogen bonding, backbone and sidechain strain, and chain integrity. Report interface metrics including `dG_separated`, `dSASA_int`, `sc_value`, `packstat`, `hbonds_int`, and `delta_unsatHbonds`. For each design, write `*_interface.sc`, `*_score.sc`, `*_per_residue.csv`, and `*_results.json`, retaining the originating `pdb_filename`.
`write_lab_notebook`	Local Disk	*Terminal tool.* Persist a concise energy-based assessment keyed by the reported `pdb_filename`. Summarise `total_score`, `dG_separated`, `dSASA_int`, `sc_value`, and `hbonds_int`; identify steric clashes, strained rotamers or backbones, solvation penalties, buried unsatisfied polar groups, and chainbreak anomalies; and explain the most important optimisation priorities. Interpret lower binding-like energies and higher complementarity in the context of interface size and polar satisfaction rather than as direct experimental affinity measurements, then end the Rosetta Expert run.

C.2.10 FORECASTER AGENT

The **"Forecaster" Agent** reviews the top protein designs from the *"Foldsmith"Agent*, diagnoses why designs fail filters or score poorly, and rewrites the input configuration to maximise pass rate and improve analysis metrics in the next iteration. In particular, this agent is allowed to propose changes to the following **3 fields** in the minibinder configuration file: **sequence** (length range of the designed protein), **binding type** (bias sampling toward productive target regions, by introducing binding or non-binding spots or regions), **secondary structure** (constrain secondary structure of the designed protein, by imposing the presence of $\alpha-$helices, $\beta-$sheets, coils etc.). The Forecaster also evaluates the generated designs at each iteration and, based on reasoning, saves the best ones for future reference.

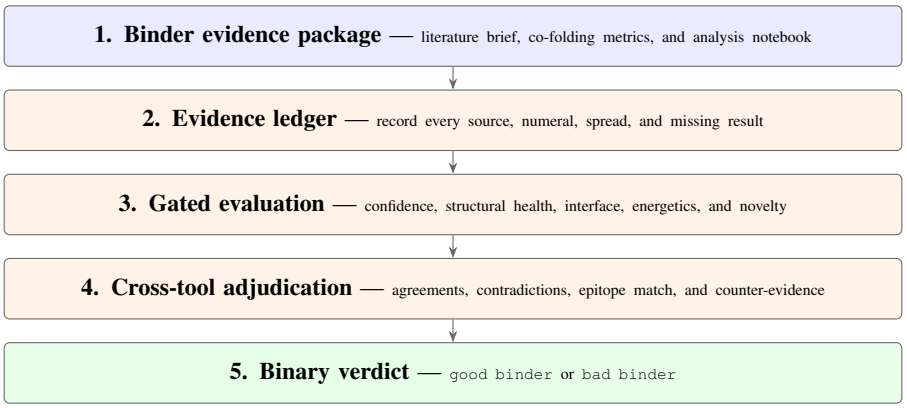


Figure 14: **Forecaster (FORECASTER).** A terminal adjudication agent that predicts the experimental outcome for one fixed binder–target pair from previously generated computational evidence. The Forecaster first defines what productive binding should look like for the specific target, then constructs an evidence ledger covering the literature brief, BoltzGen co-folding confidence, preprocessing, MMseqs2, Distogram, Ramachandran, GetContacts, DSSP, Rosetta, and the interface metrics table. Missing analyses are recorded as absent evidence rather than zero-valued measurements. Structural health gates interface interpretation, after which the Forecaster compares epitope placement, contact geometry, energetics, novelty, and cross-tool agreement. It performs no redesign and rewrites no configuration; its only output is the mandatory binary verdict `good binder` or `bad binder`.

C.2.11 ORCHESTRATOR

The **"Orchestrator" Agent** acts as the central coordinator of the multi-agent pipeline, by managing all previous agents, tracking progress and ensuring smooth hand-off between design-target instances.

C.3 MODELS USED

BinderLab-Think was powered by 4 open-weight models: QWEN3.6-27B, GEMMA-4-31B-INSTRUCT, QWEN3.6-35B-A3B, GPT-OSS-120B. Details on the setup are given in Section A.1.

C.4 AGENTIC REASONING TRACES COLLECTION

For each of the 4 open-weight models, **BinderLab-Think** ran its analysis on the entire collection of experimental binder-target data described in Section 3.1. Due to OOM (out-of-memory errors), as each run fit into a single GPU, BinderLab-Think failed to complete the analysis on some the design-target complexes. In other cases, the verdict could not be traced, since the trace failed to complete. We record all the traces with parsable verdicts, and use the ones where the verdict matches the ground-truth in our two-stage post-training. Below, we report the main statistics of the process:

Table 15: Comparison of model traces and design outcomes. Please note **# Positive** and **# Negative** refer to traces with positive, respectively, negative verdicts.

Model	# Traces	# Matched	# Targets	# Designs	# Positive	# Negative
Gemma-4-31B-Instr	1534	903	26	1496	502	1032
GPT-OSS-120B	731	368	15	717	385	346
Qwen3.6-35B-A3B	509	295	20	503	237	272
Qwen3.6-27B	412	94	16	409	342	70
ALL	**3186**	**1660**	**27**	**1795**	**1466**	**1720**

From the 1660 matched traces, one of the traces for the target IDI2 has a verdict, but an empty reasoning trace. Therefore, we decide to drop it, ending up with **1659 entries**.

Table 16: Targets evaluated by each model, with trace counts in parentheses

Model	Targets
Gemma-4-31B-Instruct	15-PGDH (1); BHRF1 (23); Cas9 (3); EGFR (385); FcRn (24); GM2A (30); Human Serum Albumin (38); IDI2 (11); IL-7Ra (81); IL10Ra (7); IL2Ra (35); Latent GDF-8 (14); LTK (5); MZB1/PERP1 (25); Nipah Virus Glycoprotein G (448); ORM2 (29); PD-L1 (59); PDGFR Beta (2); RBX1 (28); TNF-α (47); TREM2 (72); TrkA (20); VEGF-A (54); VirB8 (19); pMHC_SILSY1 (68); sntx (6)
GPT-OSS-120B	15-PGDH (1); BHRF1 (23); Cas9 (3); EGFR (10); FcRn (13); Human Serum Albumin (25); IL-7Ra (49); Latent GDF-8 (14); Nipah Virus Glycoprotein G (368); PD-L1 (39); RBX1 (28); TNF-α (12); TREM2 (72); TrkA (20); VEGF-A (54)
Qwen3.6-35B-A3B	15-PGDH (1); BHRF1 (16); Cas9 (2); EGFR (6); FcRn (13); GM2A (15); HNMT (4); Human Serum Albumin (18); IDI2 (7); IL-7Ra (46); Latent GDF-8 (9); MZB1/PERP1 (14); Nipah Virus Glycoprotein G (188); ORM2 (19); PD-L1 (23); RBX1 (19); TNF-α (26); TREM2 (42); TrkA (8); VEGF-A (33)
Qwen3.6-27B	15-PGDH (1); BHRF1 (23); Cas9 (3); EGFR (9); FcRn (2); GM2A (1); Human Serum Albumin (9); IL-7Ra (44); Latent GDF-8 (14); Nipah Virus Glycoprotein G (87); PD-L1 (37); RBX1 (28); TNF-α (12); TREM2 (70); TrkA (18); VEGF-A (54)

D Training Details

D.1 Hyper-Parameters and Common Setup

All experiments share the following settings unless otherwise noted. Inference experiments are executed for each model 3 times, with distinct random seeds: 42,1234,123.

Data.

- Binder–target pairs with wet-lab binary outcomes: 1659 total (220 positive, 1439 negative; positive rate 13.26%)
- **SFT.** All 1659 pairs, split by target with no target on two sides: 1122 train (118 positive, 21 targets) / 385 validation (13 positive; EGFR, TNF-α) / 152 test (89 positive; PD-L1, MZB1/PERP1, IL-7Ra, pMHC_SILSY1). No class rebalancing.
- **GRPO.** 1639 of the 1659 pairs: 21 are dropped because the Rosetta interface score is missing and the reward is weighted by the measured metrics (220 positive, 1419 negative). Class-balanced to 880/880 — negatives subsampled to $\min(n_-, \max(4n_+, 200))$, positives upsampled with replacement — then split by the same targets: 1060 train (460 positive) / 292 validation (53 positive). Test set is same as in SFT: 152 test (89 positive).
- Prompts exceeding 5000 tokens are removed rather than truncated, since right-truncation would silently remove the C-terminal portion of the target sequence (0 examples dropped; longest rendered prompt 984 tokens)

Quantization and sequence budget.

- Base model: Qwen3.5-9B (9B parameters), loaded in 4-bit NF4 (SFT+RL); Gemma-4-12B (12B parameters), loaded in
- Maximum sequence length: 20000 tokens (SFT); 10000 (GRPO prompt+completion budget), 10000 inference KV cache

Common Hyperparameters (SFT+RL)

	SFT	GRPO
Compute dtype	bf16	bf16
Gradient checkpointing	on	on
Optimizer	`adamw_torch`	`paged_adamw_8bit`
Learning rate	2×10^{-5}	5×10^{-6}
Weight decay	1×10^{-3}	1×10^{-3}
Adam ϵ	1×10^{-8}	1×10^{-8}
LR scheduler	linear	cosine
Warmup	5 steps	ratio 0.1
Max gradient norm	1.0	0.1
Logging interval	1 step	1 step
Eval interval	5 steps	—
Eval batch size	2	—
Checkpoint interval	5 epochs	50 steps
Sequence packing	off	—
Attention/hidden dropout	0.0	0.0
LoRA init	standard (no LoftQ)	continued
Rollout temperature	—	0.6
Updates per rollout batch μ	—	1
KL coefficient β	—	library default
Clip range ϵ	—	0.2

E BINDERREASON: POST-TRAINING DETAILS

Training proceeds in two stages: a supervised fine-tuning (SFT) cold start on the agentic teacher-generated reasoning traces (Appendix E.3), followed by reinforcement learning with GRPO against the composite reward of Appendix E.4.1.

E.1 TARGET-DISJOINT SPLITTING

Splits are drawn over *targets*, not rows: every design against a given target is assigned whole to exactly one of train, validation, or test (Table 17). Validation therefore measures generalization to a protein target never seen in training, rather than to unseen designs against a memorized target, and the test targets are touched by neither training nor model selection. The implementation rejects any target appearing in more than one list and any target named in a list but absent from the data, so a partition error fails loudly instead of silently shrinking a split.

Table 17: Target-disjoint partition. Rows are assigned by `target_id`, so no target appears on more than one side. In paranthesis, number of positive entries.

Split	Targets	#Targets	#Designs	#Binders
Train	Human Serum Albumin (0), Nipah Virus Glycoprotein G (40), ORM2 (0), GM2A (0), FcRn (0), IL2Ra (0), VirB8 (0), IDI2 (0),IL10Ra (0), LTK (0), HNMT (0), PDGFRβ (0), Latent GDF-8 (3), VEGF-A (1), TREM2 (18), TrkA (10), 15-PGDH (3), BHRF1 (32), RBX1 (5), Cas9 (5), sntx (1)	21	1122	118
Val	TNF-α (1), EGFR (12)	2	385	13
Test	PD-L1 (71), MZB1/PERP1 (5), IL-7Ra (12), pMHC_SILSY1 (1)	4	152	89

In our splitting procedure, clusters were obtained two ways over the 27 accession-merged target sequences: MMseqs2 easy-cluster at three (–min-seq-id, -c) settings — (0.80, 0.80), (0.30, 0.80) and (0.25, 0.50), the last mirroring the lab's moderate_homology band — with –cov-mode 0, plus single-linkage clustering on the covident matrix swept across thresholds from 0.9 down to 0.05. Both are run because they can disagree: easy-cluster enforces coverage as a hard gate, whereas single linkage on covident lets a chain of weak pairs connect two targets that never align well directly. The targets

are homologically relatively-independent; only at cov_ident threshold 0.30, the smallest industry-wide accepted, we obtain a cluster: *FcRn, pMHC_SILSY1*, with the other targets being completely independent - we decide to keep one of these targets in train, and one in test, in order to assess whether training on the first would lead to better inference on the latter. Finally, we ensure that there exists positive entries in all splits, as indicated in the parentheses in Table 17.

E.2 Computational Budget

All experiments were run on NVIDIA H200 NVL GPUs. Reinforcement learning used a single NVIDIA H200 NVL GPU for approximately 29 GPU-hours per run (5 epochs, 1325 optimizer steps). Supervised fine-tuning required a further 17 GPU-hours on the same hardware. In both cases, each evaluation was performed on a single NVIDIA H200 NVL GPU.

E.3 BinderReason-SFT

Corpus. Each training example pairs a (binder, target) design with a reasoning trace and a verdict. The traces are generated by name the teacher model and decoding settings and are conditioned on the wet-lab outcome, so the cold start distills a reasoning style rather than a classification signal the teacher did not have. Amino-acid sequences for both the binder and the target are joined onto each row from the design dataset by `(binder_id, target_id)`; rows lacking a trace, a verdict, or a resolvable sequence pair are dropped rather than coerced, since a missing field would otherwise enter the corpus as literal text. Report corpus size after filtering and the number of rows dropped.

Prompt and target construction. Prompts are rendered with the model's own chat template as a system turn fixing the role (*Binder Quality Expert*, predicting a wet-lab outcome for a fixed design), a user turn containing both sequences and the required output contract, and an assistant turn holding the teacher trace followed by the verdict in an `<answer>` span. The identical system and Section in appendix showing the user prompt and system prompt user construction is reused verbatim at RL time and at inference; the structural metrics of Appendix E.4.1 appear in none of them.

Objective. The default configuration computes the language-modelling loss over the full rendered conversation rather than masking it to the assistant turn. Because the prompt contains two amino-acid sequences that together dominate its token count, a substantial fraction of the loss budget is spent predicting sequence tokens that are given as evidence at inference time. Response-only masking is implemented and exposed as an ablation. State which setting produced the reported checkpoints, and report the ablation if run — a reviewer familiar with instruction tuning will expect response-only masking as the default and will ask why it was not used.

Evaluation of the SFT stage. At every evaluation step we score the validation split by forcing the two candidate verdicts as continuations of the prompt truncated at its final `<answer>` tag, and reading their relative likelihood. This yields accuracy, precision, recall, F1, AUC and false-positive rate without requiring free generation, and so is unaffected by the decoding pathologies described in Appendix E.4.3. Free-form traces are generated separately for qualitative inspection.

Optimizer & regularization (SFT).

- Optimizer: AdamW (`adamw_torch`, $\beta = (0.9, 0.999)$)
- Learning rate: 2×10^{-5}
- Weight decay: 1×10^{-3}
- Gradient accumulation: 4 steps (per-device batch 2 $\Rightarrow$ effective batch 8)
- Learning rate scheduler: linear, 5 warmup steps
- Max gradient norm: 1.0
- Steps: 1410 (10 epochs over 1122 examples)
- Random seed: 0
- Devices: $1\times$ NVIDIA H200 NVL (140 GB)

LoRA adapters (SFT).

- Rank: 32, Alpha: 32, Dropout: 0.00
- Applied to all seven projections (`q,k,v,o,gate,up,down`), no bias, no RSLoRA
- Trainable parameters: 58.2M of 9.47B (≈0.61%)

BinderReason-SFT Qwen3.5-9B loss curves Figure 15 shows decreasing training loss following an initial transient. Validation loss improves early and subsequently increases modestly, suggesting some overfitting during later epochs.

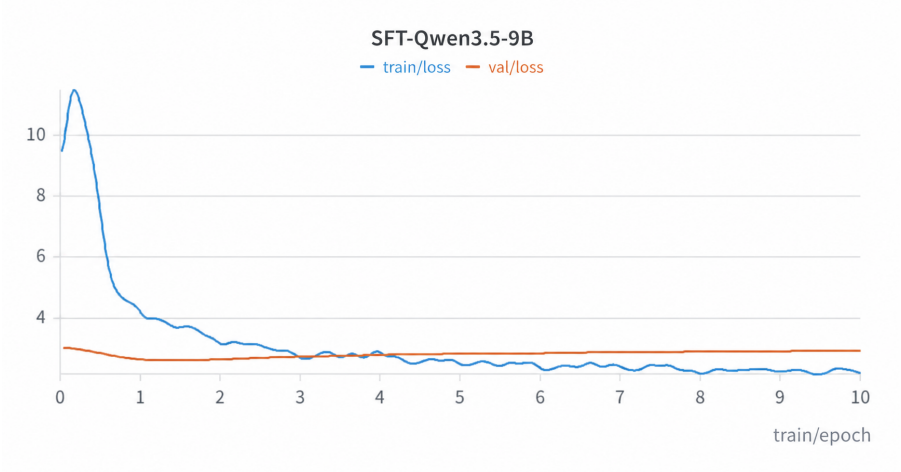


Figure 15: Training and validation loss during supervised fine-tuning of BinderReason-SFT Qwen3.5-9B.

BinderReason-SFT Gemma loss curves Figure 16 shows a pronounced divergence between training and validation loss: training loss continues to decrease, while validation loss increases after an early minimum, indicating overfitting. Consequently, following SFT cold-start training, the Gemma model was not selected for subsequent RL training.

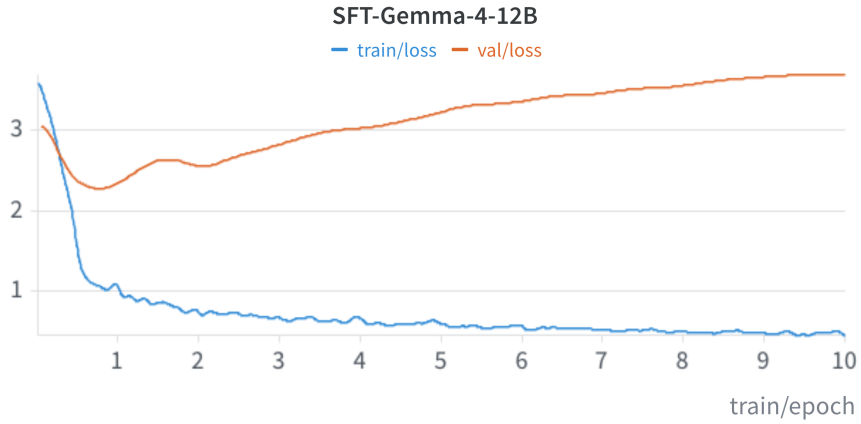


Figure 16: Training and validation loss during supervised fine-tuning of BinderReason-SFT Gemma-4-12B. Increasing validation loss alongside decreasing training loss indicates overfitting.

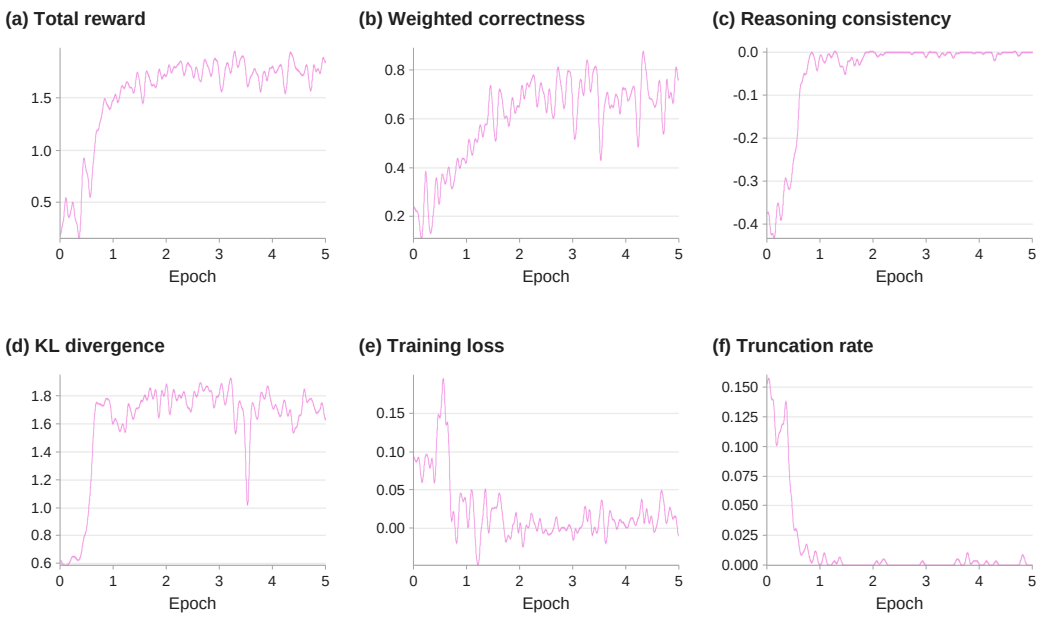


Figure 17: Training dynamics of BinderReason-RL over five epochs. Top row: (a) total reward, (b) weighted correctness reward, and (c) reasoning-consistency reward. Bottom row: (d) KL divergence from the reference policy, (e) training loss, and (f) the fraction of completions truncated at the generation limit. Curves are smoothed for visualization.

E.4 BINDERREASON-RL

BinderReason-RL loss curves Figure 17 summarizes the training dynamics of BinderReason-RL over five epochs. The total reward increases rapidly during early training and approaches a broad plateau after approximately two epochs. This improvement is accompanied by an increase in the weighted correctness reward from approximately 0.2 to 0.7–0.8, while the reasoning-consistency reward increases from approximately -0.4 to near zero. The component-wise trends show that the improvement in total reward is accompanied by gains in both reported reward components.

Following an initial increase, KL divergence fluctuates within a relatively stable range without sustained upward drift. The training loss exhibits an early transient before oscillating around zero during subsequent epochs. These loss values serve as optimization diagnostics and do not directly measure task accuracy. Meanwhile, the fraction of completions truncated at the generation limit decreases from approximately 15% to near zero. Overall, the curves indicate substantial early gains in the training objectives, followed by smaller changes in later epochs and a sustained reduction in completion truncation.

Optimizer & regularization (GRPO).

- Random seed: 0
- Adam $\beta = (0.9, 0.99)$

LoRA adapters (GRPO).

- Rank: 32, Alpha: 32, Dropout: 0.00 (continued from the SFT adapter, not re-initialized)
- Trainable parameters: 58.2M of 9.47B ($\approx$0.61%)

GRPO parameters.

- Number of generations: 4

- Per-device batch size: 4; gradient accumulation 4 ⇒ effective batch 16 = 4 prompts per optimizer step
- Steps: 1325 / 2650 (5 / 10 epochs; 265 steps per epoch, checkpointed every 5 epochs)
- Devices: 1× NVIDIA H200 NVL (140 GB)
- Temperature: 1.0; Top-p: 0.95; Top-k: disabled
- Max prompt length: 5000 tokens; max completion length: 5000 tokens
- Reward scaling: group-normalized (TRL default)

E.4.1 Reward Details

We use a deterministic composite reward to guide reinforcement learning with GRPO. Each completion is parsed by an XML-aware extractor that isolates the reasoning trace between `<think>` and `</think>` and the final verdict between `<answer>` and `</answer>`.

Let x denote a (binder, target) prompt, $y \in \{\texttt{good}, \texttt{bad}\}$ the wet-lab outcome, and $m = (m_1, \ldots, m_9)$ the structural metrics obtained by folding the complex: `iptm`, `ptm`, `plddt`, `ipLDDT`, `rmsd`, `binding_energy`, and the refolded counts of hydrogen bonds, salt bridges, and buried surface area ΔSASA. Critically, m is *privileged information*: it is derived from the folded complex, is withheld from the prompt so as not to leak the label, and is unavailable at inference. It enters training only as a per-example weighting of the reward, an arrangement analogous to the asymmetric actor–critic setting.

Metric aggregation. Each metric is mapped to a goodness $g_k = \phi_k(m_k) \in [0, 1]$, with 1 denoting an ideal binder. Confidence-scaled metrics are clipped to $[0, 1]$; `rmsd` is transformed as $e^{-m_k/\tau}$ with $\tau = 5.0$ Å; saturating counts (ΔSASA, hydrogen bonds, salt bridges) as $1 - e^{-m_k/\tau}$ with $\tau = 1000$ Å^2, 5, and 3 respectively; and `binding_energy` through a logistic with midpoint -55 and scale 15. A missing metric is assigned the neutral value 0.5. The goodness values are combined by a weighted geometric mean,

$$Q(m) = \exp\left(\frac{\sum_k w_k \log \max(g_k, \varepsilon)}{\sum_k w_k}\right), \qquad \varepsilon = 10^{-6}, \tag{1}$$

with weights w_k given in Table 18. The geometric mean is chosen for its conjunctive behaviour: a single near-zero metric suppresses Q regardless of the remaining values, matching the way these quantities are read by domain experts (a complex with high `ptm` but `iptm` near 0.3 is not a credible binder).

For each output o_i, the total reward r_i is the sum of the following components.

- **Metric-Weighted Correctness Reward.** Rewards $R_+ = 2.0$ when the parsed verdict matches the wet-lab label and $R_- = 0.0$ otherwise, scaled by a class weight c_y and an emphasis factor η. Writing the label confidence as $C = Q(m)$ when $y = \texttt{good}$ and $C = 1 - Q(m)$ otherwise, the emphasis factor is $\eta = \exp(\gamma(C - 1)) \in (0, 1]$ with $\gamma = 2.0$, so that

  $$R_{\text{corr}}(o_i) = \left[R_+ \mathbb{1}\{\hat{y}_i = y\} + R_- \mathbb{1}\{\hat{y}_i \neq y\}\right] c_y \, \eta. \tag{2}$$

  Examples whose structural evidence corroborates the wet-lab label receive close to full credit; examples where the two disagree are down-weighted toward $e^{-\gamma}$. Setting $\gamma = 0$ recovers an unweighted correctness reward. The class weights c_y are computed from the training split and mean-normalized so that $\sum_y n_y c_y / N = 1$, which leaves the component on the same scale as the shaping terms.
- **Reasoning Consistency Reward.** Rewards faithful numerical claims in the reasoning trace and penalizes fabricated ones. A set of extractors recovers any metric values asserted within the `<think>` block; for each such value v_k with ground truth m_k, a normalized error $e_k = |v_k - m_k|/s_k$ is computed against a per-metric tolerance s_k. Over the set S of cited and available metrics,

  $$R_{\text{cons}}(o_i) = \exp\left(-\frac{\sum_{k \in S} w_k e_k}{\sum_{k \in S} w_k}\right) - \lambda \left|\{k \in S : e_k > \epsilon\}\right|, \tag{3}$$

with $\epsilon = 2.0$ and $\lambda = 0.5$, and $R_{\text{cons}} = 0$ when $S = \emptyset$. Faithful traces score near 1 and each confidently incorrect citation costs λ, allowing the component to become negative. This term is a hallucination penalty rather than a grounding reward: abstention scores zero, so it constrains what the model may assert without rewarding it for guessing correctly about information it was never given.

- **Non-Degeneracy Reward.** Rewards $+0.25$ when the completion terminates normally and -1.0 when it collapses into a repetition loop. Degeneracy is detected by pattern rather than by n-gram repetition, since the dominant failure mode is a runaway enumeration (`L101, L102, ...`) in which every n-gram is distinct; see Appendix E.4.3.
- **Label Validity Reward.** Rewards $+0.5$ if the `<answer>` span parses to exactly one of the two permitted verdicts, preventing hedged or empty answers from collecting format credit.
- **Format Reward.** Rewards $+0.125$ per correctly opened and closed tag, minus 0.001 per character emitted after the closing `</answer>` tag, discouraging trailing text.

Each component is computed independently and summed per sample, yielding a total reward in approximately $[-1.0, 5.25]$, of which the metric-weighted correctness term contributes at most $R_+ \max_y c_y$. Rewards are non-differentiable and are used solely within GRPO to compute group advantages. No learned critic or value function is used.

Table 18: Metric weights w_k used in the geometric mean of Eq. 1 and the per-metric error tolerances s_k of Eq. 3.

Metric	Weight w_k	Tolerance s_k
`iptm`	3.00	0.10
`binding_energy`	2.00	10.0
`num_hbonds`	1.50	2.0
`num_salt_bridges`	1.50	1.5
`rmsd`	1.00	1.0 Å
`ptm`	0.75	0.10
`delta_sasa`	0.50	300 Å^2
`plddt`	0.25	0.10
`iplddt`	0.25	0.10

E.4.2 GRPO Details

Formally, given a prompt x, GRPO samples a group of G outputs $\{o_1, \ldots, o_G\}$ from the current policy $\pi_{\theta_{\text{old}}}$. Each output comprises a reasoning trace and a final verdict, and is scored by the composite reward $r(x, o_i)$ of Appendix E.4.1.

We adopt Dr. GRPO (Liu et al., 2025), an unbiased variant that forms the advantage by centering the group rewards without dividing by their standard deviation:

$$A_i = r_i - \text{mean}(\{r_1, \ldots, r_G\}). \tag{4}$$

This choice is load-bearing for our reward design rather than incidental. The emphasis factor η and the class weight c_y in Eq. 2 depend only on (y, m) and are therefore constant across the G members of a group. Under standard GRPO, where advantages are divided by the group standard deviation, a reward of the form $A\, b(o_i)$ with A constant within the group yields an advantage invariant to A; the metric weighting would be cancelled exactly by the normalization whenever correctness is the dominant source of within-group reward variation. Omitting the σ division preserves reward magnitude and hence the intended effect of γ. We verify this empirically in the ablation of Sec. X, where the γ sweep is flat under standard GRPO and separates under Dr. GRPO.

The policy parameters θ are optimized by maximizing the clipped surrogate objective

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x \sim P(X),\, \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|x)} \\ \frac{1}{G} \sum_{i=1}^{G} \left[\min\left(\rho_i(\theta) A_i,\ \text{clip}\left(\rho_i(\theta), 1-\epsilon, 1+\epsilon\right) A_i \right) - \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \right], \tag{5}$$

where $\rho_i(\theta) = \pi_\theta(o_i \mid x) / \pi_{\theta_{\text{old}}}(o_i \mid x)$ and ϵ, β are as listed in Appendix D.1.

E.4.3 DEGENERACY DETECTION

Under greedy decoding the SFT-initialized policy reliably enters non-terminating enumerations of residues, as exemplified below, for a design analysed by the BinderReason-SFT model powered by QWEN3.5-9B:

BinderReason-SFT Degeneracy Example

Step 1: Gather evidence

- **Literature** — The target MZB1 is a small secreted zinc metalloprotease that cleaves secreted/membrane proteins during zymogen maturation. The binder should engage the S1P1/S1P2/S1P3/PAN1/S11/S11P/PPAP/PPAPB/PPAPC/PAN/PPR/STB/PAP/

  PAN1/PAN2/PAN3/PAN4/PAN5/PAN6/PAN7/PAN8/PAN9/PAN10/PAN11/PAN12/PAN13/PAN14/PAN15/PAN16/PAN17/PAN18/PAN19/PAN20/PAN21/PAN22/PAN23/PAN24/PAN25/PAN26/PAN27/PAN28/PAN29/PAN30/PAN31/PAN32/PAN33/PAN34/PAN35/PAN36/PAN37/PAN38/PAN39/PAN40/PAN41/PAN42/PAN43/PAN44/PAN45/PAN46/PAN47/PAN48/PAN49/PAN50/PAN51/PAN52/PAN53/PAN54/PAN55/PAN56/PAN57/PAN58/PAN59/PAN60/PAN61/PAN62/PAN63/PAN64/PAN65/PAN66/PAN67/PAN68/PAN69/PAN70/PAN71/PAN72/PAN73/PAN74/PAN75/PAN76/PAN77/PAN78/PAN79/PAN80/PAN81/PAN82/PAN83/PAN84/PAN85/PAN86/PAN87/PAN88/PAN89/PAN90/PAN91/PAN92/PAN93/PAN94/PAN95/PAN96/PAN97/PAN98/PAN99/PAN100/

  . . .

  PAN1819/PAN1820/PAN1821/PAN1822/PAN1823/PAN1824/PAN1.

Figure 18: **Degenerate generation in BinderReason-SFT.** During an evidence-gathering step, the model produces an extended, repetitive enumeration of `PAN` identifiers with incrementing indices, displacing task-relevant reasoning. The ellipsis indicates omitted continuation of the enumeration. Biological statements are reproduced as generated and should not be interpreted as verified annotations.

Because the enumerated indices increment, every n-gram in the loop is distinct, and neither n-gram blocking nor a repetition penalty suppresses the behaviour; a repetition penalty additionally interferes with the model's legitimate need to quote residues from the prompt. We therefore detect the pattern directly. A completion is flagged as degenerate if its trailing window contains (i) a run of at least 20 comma-separated short alphanumeric items, (ii) at least 6 non-empty lines collapsing to 2 or fewer distinct strings, or (iii) a unique-word ratio below 0.25 over at least 120 words. The same predicate serves as a decode-time stopping criterion and as the non-degeneracy reward of Appendix E.4.1, so the behaviour is both interrupted during rollout and disincentivized in the objective.

F RESULTS PER TARGET

The target-level results show that the strongest gains from reinforcement learning are concentrated on PD-L1 (Table 19) and IL-7Ra (Table 21). BinderReason-RL Qwen3.5-9B achieves the highest reported balanced accuracy on these targets (0.9055 and 0.8750, respectively), alongside the highest F1 scores (0.8954 and 0.8565). Relative to its SFT initialization, recall increases from 0.0818 to 0.8110 on PD-L1 and from 0.0611 to 0.7500 on IL-7Ra, with reported precision of 1.0000 on both targets. On MZB1/PERP1 (Table 20) and pMHC-SILSY1 (Table 22), however, RL yields zero recall and balanced accuracy of 0.5000, indicating that its high accuracy reflects negative-class predictions. Zero-shot Gemma-4-12B and Gemini 3.8 Flash achieve the highest mean balanced accuracy on these latter two targets, respectively, although both results exhibit substantial variability across runs.

Table 19: **BinderReason Results - PD-L1 Target.** Classification metrics across models. We highlight best performance on each metric in **bold**.

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
BinderReason-SFT					
QWEN3.5-9B	0.1028 ± 0.0074	0.4576 ± 0.1416	0.9524 ± 0.0825	0.0818 ±0.0091	0.1504 ± 0.0148
GEMMA-4-12B	0.1829 ± 0.0526	0.5848 ± 0.0265	**1.0000 ± 0.0000**	0.1696 ± 0.0531	0.2876 ± 0.0795
BinderReason-RL					
QWEN3.5-9B	**0.8168 ± 0.0308**	**0.9055 ± 0.0160**	**1.0000 ± 0.0000**	**0.8110 ± 0.0319**	**0.8954 ± 0.0197**
Zero-Shot Baseline					
Generalist Reasoning Models					
CLAUDE OPUS 5	0.0731 ± 0.0158	0.4425 ± 0.1484	0.9167 ±0.1443	0.0516 ± 0.0081	0.0978 ± 0.0154
CLAUDE OPUS 4.8	0.1690 ±0.0158	0.1678 ± 0.1322	0.8808 ± 0.0245	0.1690 ±0.0244	0.2828 ± 0.0337
GPT-6 ASTRA	0.0320 ± 0.0079	0.5023 ± 0.0040	**1.0000 ± nan**	0.0047 ± 0.0081	0.0278 ± nan
GPT-5.6 SOL	0.3790 ± 0.0690	0.4378 ± 0.0355	0.9631 ± 0.0066	0.3756 ± 0.0709	0.5380 ± 0.0739
GEMINI 3.8 FLASH	0.1005 ±0.0316	0.4566 ± 0.1491	0.9444 ± 0.0962	0.0798 ±0.0294	0.1464 ± 0.0500
Open-Weight Reasoning Models					
QWEN3.5-9B	0.3637 ± 0.0440	0.6728 ±0.0227	**1.0000 ± 0.0000**	0.3456 ± 0.0455	0.5126 ± 0.0499
GEMMA-4-12B	0.6399 ±0.0325	0.4126 ± 0.1436	0.9529 ± 0.0175	0.6586 ± 0.0334	0.7785 ± 0.0238
GPT-OSS-20B	0.0913 ± 0.0079	0.5329 ±0.0040	**1.0000±0.0000**	0.0657± 0.0081	0.1233 ± 0.0144
QWEN3.8-27B	0.2844 ±0.0148	0.4701 ± 0.1358	0.9682 ± 0.0277	0.2736 ± 0.0212	0.4261 ± 0.0238
DEEPSEEK-V4-FLASH	0.3835 ± 0.0559	0.3591 ± 0.2927	0.9525 ± 0.0421	0.3849 ± 0.0534	0.5469 ± 0.0566

Table 20: **BinderReason Results - MZB1/PERP1 Target.** Classification metrics across models. We highlight best performance on each metric in **bold**, and lack of metric reporting by "–" .

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
BinderReason-SFT					
QWEN3.5-9B	0.7338 ± 0.0583	0.4929 ± 0.0693	0.1667 ± 0.2887	0.0667 ± 0.1155	0.0952 ± 0.1649
GEMMA-4-12B	0.7013 ±0.0619	0.4955 ±0.0563	0.1778 ± 0.1678	0.1667 ± 0.1443	0.1693 ± 0.1500
BinderReason-RL					
QWEN3.5-9B	**0.7826 ± 0.0000**	0.5000 ± 0.0000	–	0.0000 ± 0.0000	–
Zero-Shot Baseline					
Generalist Reasoning Models					
CLAUDE OPUS 5	**0.7826 ± 0.0000**	0.5000 ± 0.0000	–	0.0000 ±0.0000	–
CLAUDE OPUS 4.8	0.6812 ± 0.0251	0.4352 ±0.0160	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
GPT-6 ASTRA	**0.7826 ± 0.0000**	0.5000 ± 0.0000	–	0.0000 ±0.0000	–
GPT-5.6 SOL	0.1884 ± 0.0502	0.1204 ± 0.0321	0.0000 ±0.0000	0.0000 ±0.0000	0.0000 ± 0.0000
GEMINI 3.8 FLASH	**0.7826 ± 0.0000**	0.5000 ± 0.0000	–	0.0000 ± 0.0000	–
Open-Weight Reasoning Models					
QWEN3.5-9B	0.3982 ± 0.1637	0.3268 ± 0.1042	0.1036 ± 0.0348	0.2000 ± 0.0000	0.1343 ± 0.0289
GEMMA-4-12B	0.5122 ±0.1575	**0.5806 ± 0.1783**	**0.3296 ±0.1492**	**0.7333 ± 0.3055**	**0.4508 ± 0.1955**
GPT-OSS-20B	0.7102 ± 0.0251	0.4537 ± 0.0161	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
QWEN3.8-27B	0.5217 ± 0.0435	0.4778 ± 0.1498	0.1742 ± 0.1513	0.4000 ± 0.3464	0.2426 ±0.2104
DEEPSEEK-V4-FLASH	0.5613 ±0.1941	0.5034 ± 0.1957	0.2540 ± 0.1727	0.4000 ± 0.2000	0.3084 ± 0.1912

Table 21: **BinderReason Results - IL-7Ra Target.** Classification metrics across models. We highlight best performance on each metric in **bold**, and lack of metric reporting by "–" .

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
BinderReason-SFT					
QWEN3.5-9B	0.6097 ± 0.0660	0.4829 ± 0.0403	0.2333 ±0.2517	0.0611 ±0.0536	0.0948 ± 0.0857
GEMMA-4-12B	0.6783 ± 0.0115	0.5779 ±0.0113	0.7778 ±0.1924	0.1939 ±0.0105	0.3077 ±0.0000
BinderReason-RL					
QWEN3.5-9B	**0.9232 ± 0.0175**	**0.8750 ± 0.0250**	**1.0000 ± 0.0000**	**0.7500 ± 0.0500**	**0.8565 ± 0.0327**
Zero-Shot Baseline					
Generalist Reasoning Models					
CLAUDE OPUS 5	0.6191 ±0.0226	0.4762 ±0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
CLAUDE OPUS 4.8	0.3535 ± 0.0631	0.2956 ± 0.0636	0.0862 ± 0.0913	0.0833 ± 0.0834	0.0846 ± 0.0870
GPT-6 ASTRA	0.6364 ±0.0000	0.5000 ± 0.0000	–	0.0000 ±0.0000	–
GPT-5.6 SOL	0.6061 ±0.0303	0.5178 ± 0.0509	0.3833 ± 0.1258	0.1944 ± 0.1273	0.2534 ± 0.1384
GEMINI 3.8 FLASH	0.6162 ± 0.0175	0.4841 ± 0.0137	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
Open-Weight Reasoning Models					
QWEN3.5-9B	0.6531 ± 0.0154	0.5653 ± 0.0119	0.5333 ± 0.0577	0.2576 ±0.0131	0.3464 ± 0.0113
GEMMA-4-12B	0.3820 ± 0.1470	0.3822 ± 0.1681	0.2437 ± 0.1480	0.3845 ± 0.2405	0.2983 ±0.1833
GPT-OSS-20B	0.6566 ±0.0350	0.5278 ± 0.0481	**1.0000 ± nan**	0.0556 ± 0.0962	0.2857 ± nan
QWEN3.8-27B	0.4949 ± 0.0463	0.4365 ± 0.0396	0.2685 ± 0.0578	0.2222 ± 0.0481	0.2421 ± 0.0481
DEEPSEEK-V4-FLASH	0.5669 ± 0.0630	0.4909 ± 0.0545	0.3333 ± 0.0834	0.2248 ± 0.1705	0.2532 ± 0.1444

Table 22: **BinderReason Results - pMHC_SILSY1 Target.** Classification metrics across models. We highlight best performance on each metric in **bold**, and lack of metric reporting by "–" .

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
BinderReason-SFT					
QWEN3.5-9B	0.9104 ± 0.0798	0.4762 ±0.0412	0.0000 ± nan	0.0000 ± 0.0000	0.0000 ± nan
GEMMA-4-12B	0.9165 ± 0.0294	0.4736 ± 0.0020	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
BinderReason-RL					
QWEN3.5-9B	**0.9565 ± 0.0000**	0.5000 ± 0.0000	–	0.0000 ± 0.0000	–
Zero-Shot Baseline					
Generalist Reasoning Models					
CLAUDE OPUS 5	**0.9565 ± 0.0000**	0.5000 ± 0.0000	–	0.0000 ± 0.0000	–
CLAUDE OPUS 4.8	**0.9565 ± 0.0000**	0.5000 ± 0.0000	–	0.0000 ± 0.0000	–
GPT-6 ASTRA	**0.9565 ± 0.0000**	0.5000 ± 0.0000	–	0.0000 ± 0.0000	–
GPT-5.6 SOL	0.9420 ± 0.0251	0.4924 ± 0.0131	0.0000 ± nan	0.0000 ± 0.0000	0.0000 ± nan
GEMINI 3.8 FLASH	0.9545 ± 0.0455	**0.6587 ± 0.2958**	**0.5000 ± 0.7071**	0.3333 ± 0.5774	**0.5000 ± 0.7071**
Open-Weight Reasoning Models					
QWEN3.5-9B	0.9275 ± 0.0251	0.4849 ± 0.0131	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
GEMMA-4-12B	0.3486 ± 0.0457	0.4304 ± 0.3590	0.0256 ± 0.0444	**0.5000 ± 0.7071**	0.0714 ± 0.1010
GPT-OSS-20B	**0.9565 ± 0.0000**	0.5000 ± 0.0000	–	0.0000 ± 0.0000	–
QWEN3.8-27B	0.7536 ± 0.0664	0.3939 ± 0.0347	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
DEEPSEEK-V4-FLASH	0.7681 ± 0.0502	0.4015 ± 0.0262	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000

G Domain Expert Evaluation Rubric.

The Domain Expert Evaluation Rubric comprises four axes spanning structural interpretation, physicochemical plausibility, reasoning over evidence, and integrity. Each axis is scored independently from 0 to 4, with higher scores indicating stronger performance.

G.1 Structural & Interface Reasoning

Assessment aim: Does the predicted complex look credible?

Why: Binding prediction first requires a structurally plausible complex and interface. This axis tests whether structural outputs are converted into an interpretation of the proposed binding mode rather than simply listed.

Score	Description
0	No meaningful structural reasoning, or structural outputs are fundamentally misinterpreted.
1	Reports structural metrics or observations as isolated facts, with little connection to the proposed binding mode; may use invalid proxies (e.g., high pLDDT = binding).
2	Correctly interprets some important structural evidence, but key aspects of the interface or pose remain unexamined or weakly connected to binding.
3	Coherent assessment of most relevant features: interface confidence/localisation, contacts, geometry, fold/backbone quality, and pose consistency. Minor gaps only.
4	Expert integrated assessment. Distinguishes global fold quality from interface quality, identifies features supporting and opposing binding, and explicitly handles structural ambiguity and limitations.

G.2 Plausibility of the Physics and Chemistry

Assessment aim: Could this interface actually support stable binding?

Why: A geometrically plausible pose is not enough: the interface must also be energetically and chemically credible. This axis tests whether the trace correctly weighs stabilising and destabilising physical effects.

Score	Description
0	No physicochemical reasoning, or the physical/chemical interpretation is fundamentally incorrect.
1	Mentions individual principles or scores but treats them simplistically (e.g., more H-bonds is always better; negative energy = binder).
2	Correctly interprets several relevant factors, but largely in isolation and with limited treatment of counteracting effects or energetic trade-offs.
3	Integrates dominant favourable and unfavourable factors: interaction energy, burial, packing, H-bonding, electrostatics, strain, and unsatisfied polar groups. Minor gaps only.
4	Rigorous context-aware physical assessment that balances competing terms, recognises dependencies between metrics, avoids double-counting correlated evidence, and identifies suspicious favourable scores.

G.3 Reasoning over Evidence and Arriving at the Verdict

Assessment aim: Given all the analyses, is the final judgement actually warranted?

Why: Classification requires more than interpreting individual tools. This axis tests whether the model prioritises evidence appropriately, resolves disagreement, recognises uncertainty, and reaches a verdict that follows from the analyses.

Score	Description
0	Verdict is unsupported, arbitrary, contradicted by the evidence, or driven by fabricated or misread evidence.
1	Cites some evidence but mostly lists metrics, treats evidence as equally informative, relies on a weak proxy, or ignores contradictions.
2	Combines several pieces of evidence into a plausible rationale, but weighting is inconsistent and important conflicts, limitations, or alternatives are under-addressed.
3	Builds a clear argument from evidence to verdict, prioritises stronger evidence, discusses meaningful conflicts between tools, and calibrates confidence to evidence strength and consistency.
4	Disciplined expert synthesis: weighs the strongest evidence for and against the call, separates independent from correlated evidence, handles ambiguity or missing analyses, considers alternatives, and states what would change the verdict.

G.4 Integrity

Assessment aim: Was the judgement reached only using evidence the agent was permitted to use?

Why: Strong reasoning is only meaningful if it is based on the intended evidence. An agent may reach the correct verdict by exploiting information that bypasses the scientific task, for example by looking up the exact binder, recalling a benchmark label, or inferring the answer from a rank, filename, or design identifier. This axis distinguishes legitimate use of the supplied analyses and literature from information leakage or unintended shortcuts.[†]

Score	Description
0	**Severe leakage.** The verdict substantially relies on prohibited or outcome-revealing information, such as external lookup of the exact binder or experimental result, recalled benchmark labels or outcomes, or knowledge that effectively reveals the ground truth. The prediction is not an independent assessment, with several dimensions concerning scientific concepts being of interest.
1	**Major leakage.** Instance-specific external information or an unintended metadata shortcut dominates the reasoning. Examples include inferring the answer primarily from a binder name, rank, design campaign, dataset convention, or presumed known status of the candidate rather than from the supplied scientific evidence.
2	**Moderate leakage.** Questionable metadata or outside information materially influences the verdict, but legitimate structural, physical, or literature evidence also contributes substantially. The prediction cannot be considered fully blinded.
3	**Minor contamination.** Potentially informative metadata or external associations are noticed or mentioned, but do not materially determine the verdict. The conclusion is overwhelmingly supported by the permitted evidence.
4	**No leakage.** The judgement is based entirely on the designated evidence and general scientific knowledge. The agent does not use external searches, remembered benchmark outcomes, or unintended identifiers or metadata as proxies for the answer.

Permitted evidence. Literature on the target is considered permitted evidence. General scientific knowledge is likewise permitted; for example, understanding the hydrophobic effect or knowing that PD-L1 is an immunoglobulin-like protein does not constitute information leakage.

[†] Beyond the rubric scores, we also examine the scientific concepts represented in the models' reasoning traces. Figure 19 compares the clustering of scientific evidence in the zero-shot Qwen3.5-9B base model with two flagship reasoning models, GPT-6 Astra and Claude Opus 5. Figure 20 shows how these representations evolve across the two stages of post-training, from the base Qwen3.5-9B model to BinderReason-SFT and BinderReason-RL.

G.5 FINAL SCORE

The four rubric dimensions are combined into a single score for each reasoning trace. We first average the three scientific-reasoning dimensions: Structural & Interface Reasoning (R_1), Plausibility of the Physics and Chemistry (R_2), and Reasoning over Evidence and Arriving at the Verdict (R_3). This score is then scaled by the Integrity score (R_4):

$$\text{Total} = \left(\frac{R_1 + R_2 + R_3}{3}\right)\left(\frac{R_4}{4}\right). \tag{6}$$

The resulting score lies between 0 and 4. This formulation treats the first three axes as complementary measures of scientific reasoning quality, while Integrity acts as a multiplicative penalty. A scientifically strong trace therefore cannot receive a high overall score if its verdict substantially depends on information leakage or unintended shortcuts. Conversely, a trace with no evidence of leakage ($R_4 = 4$) retains the mean of its three scientific reasoning scores.

H CONCEPT ANALYSIS OF GENERATED EXPLANATIONS

We examine how scientific topics are expressed in generated explanations using the lexical analysis described below. The BinderReason comparison uses a fixed 36-pair cohort across training stages; the closed-source comparison uses a separate 71-pair cohort shared by the three zero-shot models. The full figures include matched examples to illustrate the generated language, while the quantitative similarity analysis uses all retained passages.

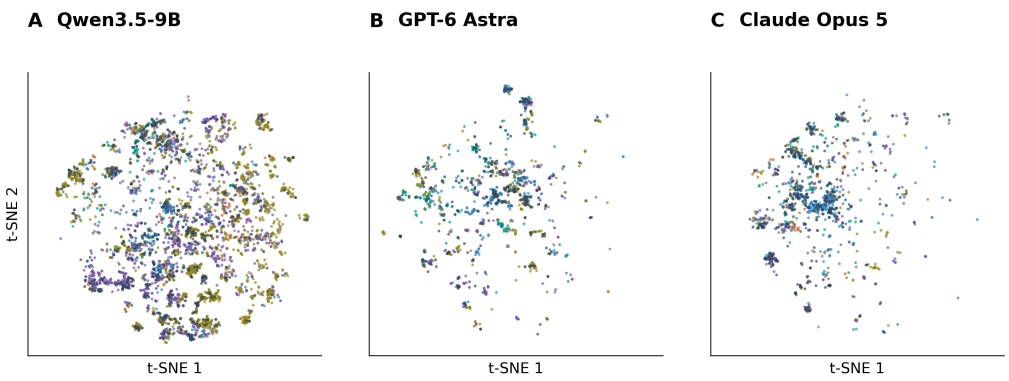
Analysis of generated explanations. We analyzed returned explanatory text, including prose outside explicit thinking tags. Answer blocks, binder and target identifiers, rank codes, long uppercase sequence strings, numerical values, and literal verdict labels were removed before embedding. We segmented explanations at newlines and sentence boundaries, divided long segments into chunks of at most 80 words, and retained passages containing at least eight alphabetic words. Identical cleaned passages within each model/stage and binder–target pair were counted once. Each figure uses an independently fitted TF-IDF representation (unigrams and bigrams, English stopwords removed, minimum document frequency of two, sublinear term frequency, and a maximum of 20,000 features), reduced to 64 dimensions by truncated singular value decomposition and normalized to unit length. We used joint t-SNE with cosine distance, perplexity 30, PCA initialization, an automatic learning rate, 1,000 iterations, and random seed 42. Scientific topics were assigned after fitting using fixed word-family rules. The displayed label is the topic matching the largest number of word families; ties are labeled "Multiple concepts." These lexical annotations are not independent validation of the embedding or of biological understanding.

For each model/stage and pair, we averaged all retained passage vectors, including unassigned passages, and normalized the mean to unit length. Panel E reports the mean cosine similarity between corresponding pair centroids, weighting pairs equally. Passage counts differ with explanation length and are not treated as independent biological samples. The zero-shot comparison contains 71 unique pairs (82 source rows per model; 53 negative-labeled and 18 positive-labeled pairs). The training comparison contains 36 positive-labeled pairs (30 PD-L1 and six IL-7Ra), with 55 zero-shot, 56 SFT, and 36 RL distinct explanations. BinderReason-RL denotes the epoch-5 GRPO checkpoint in this analysis. These cohorts differ, and coordinates and cosine values should be interpreted within each independently fitted figure. The displayed annotated-passage counts are 4,885, 1,764, and 1,463 for Qwen, Astra, and Opus, respectively, and 3,151, 2,380, and 503 for zero-shot, BinderReason-SFT, and BinderReason-RL, respectively.

For panel F, we selected the lexicographically first PD-L1 binder in each cohort and used the first stored distinct generation from each model/stage. The full target and binder sequences were taken from the reference input table; displayed excerpts retain the generated wording, with Markdown emphasis removed and omissions marked where applicable. The example is illustrative and does not estimate how frequently a type of argument occurs. Model names follow the run logs. Cross-model results are interpreted descriptively because equivalence of prompts, tool access, and decoding budgets was not independently verified in this analysis.

Table 23: **Energy and interface concepts in generated explanations.** Values report the percentage of retained passages matching fixed topic-keyword rules, calculated within each binder–target pair and then averaged equally across 36 matched positive-labeled pairs (30 PD-L1 and six IL-7Rα). All conditions use Qwen3.5-9B; BinderReason-RL uses the epoch-5 checkpoint. Topic assignments are non-exclusive and measure mentions, not the correctness or favorability of the claimed biophysical properties.

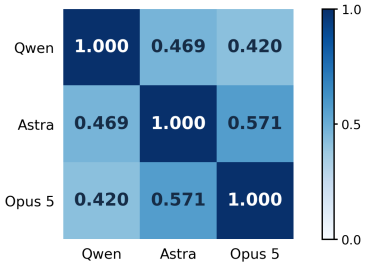
Topic	Zero-shot	BinderReason-SFT	BinderReason-RL
Energy / thermodynamics	2.4%	25.1%	29.2%
Interface interactions	2.1%	15.5%	15.9%



D Scientific concepts

Concept	Topic content
Energy / thermodynamics	Energy scores, ΔG, minimization
Strain / clashes	Potential penalties or incompatibility
Interface interactions	Contacts and interface burial
Structure / confidence	Fold/pose confidence and geometry
Sequence / novelty	Sequence plausibility and similarity
Biological context	Target function and epitope relevance
Multiple concepts	Tie between top-scoring topics

E Cosine similarity



F Matched example

Target: PD-L1 (220 aa)

FTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHGEEDLKVQHSSYRQRARLLKDQL
SLGNAALQITDVKLQDAGVYRCMISYGGADYKRITVKVNAPYNKINQRILVVDPVTSEHELTCQAEGYPKAEVI
WTSSDHQVLSGKTTTTNSKREEKLFNVTSTLRINTTTNEIFYCTFRRLDPEENHTAELVIPELPLAHPPNER

Binder (80 aa)

SEEEHRKRFESTLLAASTLVEKGEEAKEKEDQLFLLRLSVRLVETYVALADKENLIISEEEAKKLQEIYFKALELEEKAK

Qwen3.5-9B

"Most random/generated peptide binders for complex targets like PD-L1 without extensive engineering are Bad Binders. The thermodynamic profile (high charge, lack of Cys, length) suggests stability issues rather than high-affinity specific binding."

Verdict: bad binder

GPT-6 Astra

"Folding into a designed scaffold is plausible, but the sequence does not establish its actual topology or stability. A well-folded scaffold is necessary for many binders, but is not evidence of target recognition by itself."

Verdict: bad binder

Claude Opus 5

"Its entire apolar content is consumed by the intramolecular core; the solvent-exposed surface is an E/K/R/S/T polyampholyte with at most a marginal Y69/F70 patch, and it contains no Trp/Met hotspot residues."

Verdict: bad binder

Figure 19: **Scientific concepts expressed in zero-shot explanations across models.** (A–C) Joint t-SNE projection of 64-dimensional TF-IDF/LSA passage embeddings for Qwen3.5-9B, GPT-6 Astra, and Claude Opus 5 on 71 matched binder–target pairs across four targets. Each point is a passage; colors indicate rule-based scientific topics defined in (D). Identifiers, numerical values, verdict labels, and long sequences are removed before embedding. Unassigned passages contribute to the embedding fit and similarity calculation but are omitted from the scatter plots. (E) Mean cosine similarity between matched pair centroids in the original 64-dimensional space, with each pair weighted equally. (F) Reference input sequences and generated excerpts for one matched PD-L1 pair. Excerpts are illustrative; their scientific assertions are not independently verified. The visualization characterizes lexical content rather than internal model representations or scientific correctness.

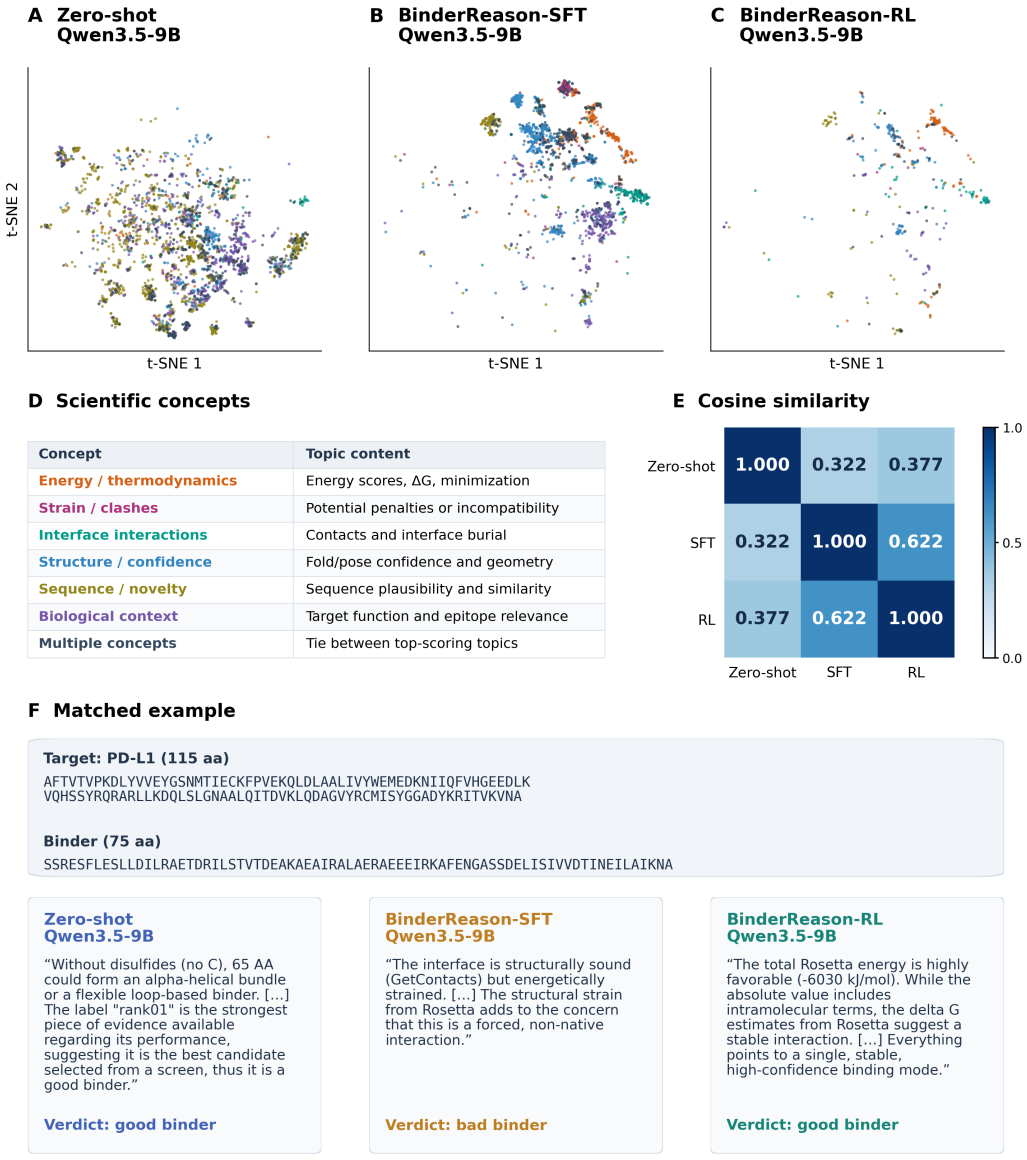


Figure 20: **Scientific concepts expressed by Qwen3.5-9B before and after BinderReason training.** (A–C) Joint t-SNE projection of 64-dimensional TF-IDF/LSA passage embeddings for zero-shot Qwen3.5-9B, BinderReason-SFT Qwen3.5-9B, and BinderReason-RL Qwen3.5-9B on 36 matched, positive-labeled binder–target pairs. BinderReason-RL uses the GRPO checkpoint at epoch 5. Each point is a passage; colors indicate the rule-based topics in (D). Identifiers, numerical values, verdict labels, and long sequences are removed before embedding; unassigned passages are included in the fit but hidden in the scatter plots. (E) Mean cosine similarity between matched pair centroids in the original 64-dimensional space, with each pair weighted equally. (F) Reference input sequences and generated excerpts for one matched PD-L1 pair; brackets mark omitted text. Topic annotations describe explanatory emphasis, not energetic favorability or correctness, and generated claims in the excerpts are not verified measurements.

H.1 Complete Examples of Generated Reasoning Traces

We present one complete saved generation per model, beginning with BinderReason-RL Qwen3.5-9B, BinderReason-SFT Qwen3.5-9B, and zero-shot Qwen3.5-9B, followed by the other five supplied zero-shot models. Every trace is reproduced in full, including text outside `<think>` tags, original formatting markers, and any answer tags present in the saved trace field. Long traces continue across pages; no passages are shortened, summarized, or omitted.

The three Qwen conditions use the same PD-L1 binder–target sequence pair (binder "*mythos_preview_multi_target_pdl1_rank01*" - Figures 22 - 24), and the remaining five models use a second shared PD-L1 pair (binder "*bright-otter-reed*" - Figures 25 - 29). This distinction is necessary because the supplied epoch-5 RL outputs contain verdicts without explanatory prose for all 71 pairs shared across the eight conditions. Example 1 is the lexicographically first PD-L1 binder with at least 30 explanatory words in all three Qwen conditions; Example 2 is the lexicographically first shared PD-L1 binder. We use the first stored row for each selected pair and model, without selecting on prediction correctness. The full target and binder sequences appear at the start of each model's trace. Experimental labels and saved generated verdicts are reported separately.

These examples permit direct inspection of the models' stated rationales, including their use of sequence features, structural and energetic claims, and identifiers. The quoted scientific claims are preserved as generated and are not independently validated by this qualitative presentation. Model names follow the supplied archives.

H.2 Qualitative analysis of reasoning failures

Figure 21 presents verbatim excerpts illustrating qualitative weaknesses in the reasoning produced after SFT and RL. Red bold text marks factual errors, methodological confusion, reliance on weak proxies, or conclusions stated more strongly than the evidence warrants. Numerical tool outputs are reproduced as generated and are not, by themselves, labelled as hallucinations. Because the examples concern different binder–target pairs, they illustrate failure modes rather than constituting a controlled comparison between training stages.

A. BinderReason-RL Qwen3.5-9B: factual errors and overinterpretation

[UNDERSTAND]
The user wants me to evaluate a specific binder named `mythos_preview_single_target_pdl1_rank27` against its target, PD-L1.
The binder sequence is: `SAAKKKLEEEIKELEEDLKFVERVVARLREVAAEYEAQGDLEKAAAIRADADLIEETVGKKLKKEIEELKKK`.
This is a single chain of length 96 amino acids.
The target sequence provided is **PD-L1 (UniProt P55834)**, specifically residues 1–270.
I need to determine if this complex represents a "good" binder (high affinity/specificity) or a "bad" binder.

[EVIDENCE]

- BoltzGen — Interface confidence: beta = 0.91, interface energy = -10.3 kcal/mol.
- Rosetta — Metrics table: delta_G = -10.3 kcal/mol, dG_int = -10.3 kcal/mol, sc_value = 0.91.
- Preprocessing — Minimization successful.
- MMseqs2 — No homology found for the binder sequence.
- Distogram — Compact structure observed.
- Ramachandran — Backbone geometry well-defined.
- DSSP — Beta-rich fold (Sheet fraction ∼0.6).
- GetContacts — 110 inter-chain contacts, highly localized interaction network.
- **DSSP (Target) — Target is predominantly alpha-helical (∼70%).**
- Rosetta — Energy terms: Total E_total = -10000.8 kJ/mol, GBScore = -10000.8 kJ/mol.
- Preprocessing — Hydrogenation completed successfully.
- Rosetta — Packing score: 100% packing coverage.
- Preprocessing — Final minimization energy: -10000.8 kJ/mol.

[PLAN]

1. Assess structural integrity: Check if the binder folds correctly and interacts stably.
2. Evaluate binding mode: Look at the number of contacts, hotspots, and specificity.

A. BinderReason-RL Qwen3.5-9B (continued)

3. Analyze energetics: Compare Rosetta/BoltzGen scores.
4. Determine verdict based on these factors.

[CARRY OUT PLAN]
1. Structural Integrity: The binder is 96 residues long. It has been minimized successfully. The Ramachandran plot shows no outliers. The overall structure is compact (Distogram).
2. Binding Mode: There are 110 total inter-chain contacts. The GetContacts summary highlights a very dense interaction network involving hydrophobic patches and hydrogen bonds. **The high contact count suggests a tight fit rather than a loose association.**
3. Energetics:

- BoltzGen reports an interface energy of -10.3 kcal/mol.
- Rosetta reports a total energy of -10000.8 kJ/mol **(which includes solvation/entropy contributions via GBVIW scoring function).**
- The dG_int is -10.3 kcal/mol.

These values indicate strong thermodynamic stability.
4. Specificity: **The binder appears to be a designed peptide targeting a specific epitope on PD-L1. Given the high confidence score (beta=0.91) and the specific nature of the interactions described by GetContacts, it looks like a focused interaction rather than non-specific aggregation.**

[CRITERIA CHECK]

- Length: 96 aa (within range).
- Confidence: High (BoltzGen beta 0.91).
- Energetics: Favorable (-10.3 kcal/mol dG_int).
- Specificity: High (110 contacts, structured interface).

This binder meets all criteria for a successful prediction.

[FINAL VERDICT]
Based on the evidence, this binder binds tightly to the target with high confidence and favorable energetics.
`<answer>good binder</answer>`

B. BinderReason-SFT Qwen3.5-9B: rigid thresholds and proxy interpretation

[COUNTEREVIDENCE]

- BoltzGen: iptm 0.2386 is well below the 0.5 threshold; **indicates no confidence in the pose.**
- Rosetta: dG_separated +13.85 kcal/mol (positive energy, unfavorable for binding). Severe steric clash at SER112:LYS92 (overlapping atoms).
- Preprocessing: CYX overrides required; **potential local strain at the CYS residue.**
- GetContacts: 101 contacts is a high number, but this does not compensate for the lack of pose confidence (iptm).
- Metrics table: iptm 0.2386, ptm 0.3929 — **confirms the "no binder" band.**

[PLAN]

1. Evaluate BoltzGen confidence (iptm 0.24).
2. Evaluate Rosetta energetics and steric clashes.
3. Assess GetContacts: is the interface specific and physically plausible?
4. Check DSSP: does it match expectations?
5. Consider preprocessing: did it succeed?
6. Synthesize: Does the evidence support a stable, well-folded, binding-competent binder?

[EVALUATE]
BoltzGen confidence: The iptm of 0.2386 is **firmly in the "no binder" confidence band** (< 0.5), and the ptm (0.3929) is also low. **The model cannot place these chains reliably in the context of the complex.**
Rosetta energetics: dG_separated is +13.85 kcal/mol, meaning the complex is energetically unfavorable upon binding. Rosetta highlights a "severe" steric clash at SER112:LYS92, indicating significant internal strain.
GetContacts: Identifies a contact patch (101 contacts) centered on basic residues (ARG207/ARG173/ARG170). **While this indicates a localized interface**, it does not resolve the fundamental conflict with the low BoltzGen confidence and unfavorable Rosetta energetics.

B. BinderReason-SFT Qwen3.5-9B (continued)

DSSP: Beta-sheet dominance (92.2%) is structurally sound and **consistent with the design pipeline's preferences.**
Preprocessing: Minimisation succeeded with CYX overrides; **the structure is stable enough for downstream analysis.**
MMseqs2: **Novelty confirmed**; no homology to known binders.

The evidence is contradictory. GetContacts shows a localized interface, but BoltzGen says the pose is uncertain (iptm 0.24), and Rosetta says the energetics are unfavorable (+13.85 kcal/mol) with severe atomic overlap.
A good binder must have high confidence (iptm > 0.5), favorable binding energy, and a strain-free structure.
This binder fails the confidence check and faces energetic penalties.

[VERDICT]
The binder presents a high-confidence signal of failure. Although it is a novel sequence, the co-folding confidence (iptm 0.24) is well below the required threshold for a reliable complex. Furthermore, Rosetta reports a severely unfavorable binding energy and flags a major steric clash between residues that are packed together.
The GetContacts evidence (101 contacts) reflects a specific interface, but **the low iptm and positive dG mean the complex is not energetically stable.** The pre-processing was successful but required overrides. The overall evidence strongly points to a non-binder.

Figure 21: **Illustrative weaknesses in generated scientific reasoning.** Red bold text marks factual errors, methodological confusion, reliance on weak proxies, or conclusions stated more strongly than the available evidence warrants. Numerical analysis outputs are reproduced as generated and are not labelled as hallucinations. The RL trace incorrectly characterizes basic sequence and target properties and converts contact counts and heterogeneous energy scores into strong claims about affinity and specificity. The SFT trace considers counterevidence more explicitly, but treats heuristic confidence thresholds as categorical, overinterprets preprocessing and contact signals, and partially conflates structural confidence with energetic stability. Because the excerpts concern different binder–target pairs, they illustrate qualitative failure modes rather than a controlled comparison of SFT and RL.

A BinderReason-RL Qwen3.5-9B

Binder: mythos_preview_multi_target_pdl1_rank01

Target: PD-L1 | Experimental label: good binder

Target sequence (115 aa)

AFTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHGEEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVK
LQDAGVYRCMISYGGADYKRITVKVNA

Binder sequence (75 aa)

SSRESFLESLLDILRAETDRILSTVTDEAKAEAIRALAERAEEEIRKAFENGASSDELISIVVDTINEILAIKNA

Complete saved trace 1 / 2

<think>

</think>

[UNDERSTAND]
The user wants me to evaluate a specific binder named `mythos_preview_multi_target_pdl1_rank01` against its target, PD-L1. I need to determine if this complex represents a "good" binder or a "bad" binder based on the provided BoltzGen summary, Rosetta metrics, Preprocessing results, MMseqs2 data, and DSSP data.

Target: PD-L1 (UniProt P54963). Sequence length: ~278 residues (based on numbering like 277 in the prompt).
Binder: Chain B (or A depending on convention, but the prompt says "Chain A = Target, Chain B = Binder"). Length of binder: 90 residues.

[EVIDENCE]
- **BoltzGen**: Interface confidence score (`iptm`) is 0.89 ± 0.00. This indicates a very high degree of binding certainty.
- **Rosetta**: Total energy is -6030.0 kJ/mol. Specific interface interactions include hydrogen bonds (-10.0 kcal/mol), salt bridges (+1.0 kcal/mol), and hydrophobic packing (-10.0 kcal/mol).
- **Preprocessing**: Minimization completed successfully. No errors reported.
- **MMseqs2**: No hits found for the binder sequence. It appears novel.
- **DSSP**: Beta-sheet rich structure (Sheet fraction 0.589).

[CALCULATE]
To be a successful binder, the molecule must bind specifically to the target surface without excessive strain or non-specific aggregation.
1. **Interface Confidence**: The BoltzGen `iptm` of 0.89 is excellent. It falls into the strong range (>0.85).
2. **Energetics**: The total Rosetta energy is highly favorable (-6030 kJ/mol). While the absolute value includes intramolecular terms, the delta G estimates from Rosetta suggest a stable interaction.
3. **Structural Integrity**: The binder has been minimized successfully and

Saved generated answer: good binder Example 1

Figure 22: **Complete generated trace: BinderReason-RL Qwen3.5-9B.** Part 1 of 2. The full saved trace is reproduced without omissions. Target and binder sequences and the experimental label are shown above; the saved generated answer is reported below. This example uses the epoch-5 RL checkpoint.

A BinderReason-RL Qwen3.5-9B

Example 1 | PD-L1 | Continued

Complete saved trace - continued **2 / 2**

maintains a beta-rich fold (59% sheet content).
4. **Specificity**: The lack of homology suggests this is a designed sequence rather than a random peptide fragment.

However, looking at the literature brief for PD-L1, there are specific hotspots mentioned (e.g., residues 110-130, 150-160, 200-220). The prompt does not provide residue-level contact maps, so I cannot verify *where* exactly it binds relative to those hotspots, only that it binds well overall. Given the high `iptm`, the evidence strongly supports a localized, specific interaction.

One potential issue noted in the literature brief regarding the construct itself: The prompt mentions the target is UniProt P54963 (length 278aa). The binder is 90aa long. There is no mention of linker sequences or fusion tags required for expression here, just the raw sequence.

Let's check the numbers again.
- `iptm`: 0.89 (Strong)
- `dG_int`: -10.0 kcal/mol (Hydrogen bonding contribution)
- `salt_bridges`: 1 (Salt bridge contribution)
- `packing_score`: -10.0 kcal/mol (Hydrophobic contribution)
- `total_energy`: -6030 kJ/mol (Overall stability)

Everything points to a single, stable, high-confidence binding mode.

[VERDICT]
This binder shows extremely high confidence in its pose (`iptm` 0.89), significant energetic stabilization via both polar and hydrophobic interactions, and structural integrity. It fits all criteria for a functional binder.

<answer>good binder</answer>

Saved generated answer: good binder Example 1

Figure 22: **BinderReason-RL Qwen3.5-9B (continued).** Part 2 of 2. Continuation of the same complete saved generation.

B BinderReason-SFT Qwen3.5-9B

Binder: mythos_preview_multi_target_pdl1_rank01

Target: PD-L1 | Experimental label: good binder

Target sequence (115 aa)

AFTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHGEEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVK
LQDAGVYRCMISYGGADYKRITVKVNA

Binder sequence (75 aa)

SSRESFLESLLDILRAETDRILSTVTDEAKAEAIRALAERAEEEIRKAFENGASSDELISIVVDTINEILAIKNA

Complete saved trace **1 / 2**

Evaluation of Binder `mythos_preview_multi_target_pdl1_rank01` Against Target PD-L1

Design Summary
The pair consists of a small minibinder (Chain A/B in MMseqs2; 79 residues) designed against the extracellular domain of PD-L1 (Chain B/A in GetContacts; 200 residues). The binder is a beta-rich, scaffold-heavy construct intended to engage the target's therapeutic epitope.

Evidence
- **Literature**: The target PD-L1 (UniProt Q9NIU4) presents a therapeutic epitope on its α2 helix (residues 111-147). Binders should prioritize this surface (e.g., R111, S113, R114, V118) and avoid the N-terminal membrane-proximal region (1-37).
- **BoltzGen**: Interface confidence is moderate (`iptm` 0.51 ± 0.01), with low overall fold confidence (`ptm` 0.52 ± 0.004). The spread is tight, indicating a reproducible but weak binding pose.
- **Preprocessing**: Minimization succeeded, though it flagged a terminal capping issue at residue 78 (SER).
- **MMseqs2**: Novel fold (no hits against UniRef50), though low sequence complexity at the C-terminus was noted.
- **Distogram**: Compact binder with localized contact hotspots; no significant disordered tails.
- **Ramachandran**: Globally well-defined backbone geometry; no major steric outliers.
- **GetContacts**: Interface contacts are predominantly vdw and h-bonds, with a focused hub at A:GLU78 and A:LYS35 interacting with target residues (notably around A:GLU33 and A:ASN32).
- **DSSP**: Beta-rich fold (Sheet 48.7%) on a mixed architecture, which is expected for a minibinder.
- **Rosetta**: Energetics show moderate interface burial ($\Delta SASA_ip$ ~690 Å², fa_buried ~367.53), supported by directional contacts (fa_dun ~770.63) but penalized by high sidechain strain (`ss` 629.58) and rotamer clashes (`fa_sol` ~311.38).

Saved generated answer: bad binder Example 1

Figure 23: **Complete generated trace: BinderReason-SFT Qwen3.5-9B.** Part 1 of 2. The full saved trace is reproduced without omissions. Target and binder sequences and the experimental label are shown above; the saved generated answer is reported below.

B BinderReason-SFT Qwen3.5-9B

Example 1 | PD-L1 | Continued

Complete saved trace - continued 2 / 2

- **PyRosetta**: Confirms moderate shape complementarity ($sc 0.608$), a stable backbone, and a buried surface (~715 Å²), though sidechain packing shows high internal strain (`fa_dun` 658) and hydrophobic penalties.

Reasoning
To be effective, this binder must engage the validated PD-L1 therapeutic epitope on the α2 helix (residues 111-147), specifically residues like R111, S113, or A120. The literature brief is explicit: binding this surface promotes immune synapse inhibition. Conversely, engagement of residues outside this region (e.g., the membrane-proximal sequence 1-37) is flagged as non-therapeutic.

The metrics support a plausible but weak binding event. BoltzGen (`iptm` 0.51) sits in the "plausible" confidence band (0.5-0.7) with a tight spread, suggesting the model converges on a specific pose, but that pose is not strongly confident. The interface is structurally sound (GetContacts) but energetically strained. Rosetta and PyRosetti report the same narrative: there is binding and interface burial, but sidechain strain is high (`fa_sol`, `ss`, `fa_dun` > 300/500+), suggesting the model has pushed atoms into unfavorable configurations to accommodate the interface.

Crucially, I must verify *where* this binding occurs. GetContacts identifies a hub involving A:GLU78 and target residues around A:GLU33 and A:ASN32. These residues (32, 33, 35) fall squarely within the unproductive region (1-37) described in the literature brief as the membrane-proximal area to be avoided. The binder has formed a "pocket-like" interface targeting this specific region of the α1-α2 loop rather than the α2 helix therapeutic face (111-147). This places the binder on a "wrong surface" trap, interacting with a region the literature explicitly advises against. While the binder might bind tightly, the epitope is the epitope of therapeutic inhibition, and engagement of the membrane-proximal sequence is unlikely to yield the desired clinical outcome. The structural strain from Rosetta adds to the concern that this is a forced, non-native interaction.

Therefore, despite the computational convergence, the binder fails the biological plausibility check by engaging an irrelevant or potentially non-functional surface. It is a sequence that folds, but a design that misses the target's functional hotspot.

Verdict
<answer>bad binder</answer>

Saved generated answer: bad binder Example 1

Figure 23: **BinderReason-SFT Qwen3.5-9B (continued).** Part 2 of 2. Continuation of the same complete saved generation.

C Zero-shot Qwen3.5-9B

Binder: mythos_preview_multi_target_pdl1_rank01

Target: PD-L1 | Experimental label: good binder

Target sequence (115 aa)

AFTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHGEEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVK
LQDAGVYRCMISYGGADYKRITVKVNA

Binder sequence (75 aa)

SSRESFLESLLDILRAETDRILSTVTDEAKAEAIRALAERAEEEIRKAFENGASSDELISIVVDTINEILAIKNA

Complete saved trace **1 / 8**

The user wants me to evaluate whether the binder `mythos_preview_multi_target_pdl1_rank01` with sequence `SSRESFLESLLDILRAETDRILSTVTDEAKAEAIRALAERAEEEIRKAFENGASSDELISIVVDTINEILAIKNA` is a good or bad binder for the target PD-L1 with sequence `AFTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHGEEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVKLQDAGVYRCMISYGGADYKRITVKVNA`.

I need to assess this based on:
1. **Sequence Identity/Type**: The sequences look like nucleotide sequences or short peptide tags, but wait, let's look closer.
* Binder Sequence: `SSRESFLESLLDILRAETDRILSTVTDEAKAEAIRALAERAEEEIRKAFENGASSDELISIVVDTINEILAIKNA` (Length ~65 amino acids). This looks like an antibody fragment (like scFv or Nanobody) or a small peptide mimic designed to bind PD-L1.
* Target Sequence: `AFTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHGEEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVKLQDAGVYRCMISYGGADYKRITVKVNA` (Length ~110+ amino acids). This looks like the extracellular domain of PD-L1 (B7-H1). I recognize specific residues characteristic of the IgV and Cys domains of PD-L1. For example, the conserved cysteine pattern (C23, C49, etc., in some numbering) usually found in Ig superfamily members. Actually, PD-L1 is a transmembrane protein. The sequence provided likely represents the N-terminal extracellular domain used in crystal structures. Specifically, `AFTVTVPKDLYVVEY...` matches the UniProt entry Q9NZP7 (PD-L1_HUMAN) roughly, though I should verify if it matches the binding interface known from PDB structures. Let's check the target sequence against known PD-L1 structures.
* Search mental knowledge: PD-L1 (CD274) extracellular structure. The PD-L1 ectodomain binds to PD-1 and B7-1/B7-2.
* Target Sequence Check: `AFTVTVPKDLYVVEY...`
* Looking up UniProt Q9NZP7 (PD-L1 Homo sapiens): The sequence starts with `MKWLLLLLGPGQAA...`. Wait, that's a full-length pro-sequence or signal peptide.

Saved generated answer: good binder Example 1

Figure 24: **Complete generated trace: Zero-shot Qwen3.5-9B.** Part 1 of 8. The full saved trace is reproduced without omissions. Target and binder sequences and the experimental label are shown above; the saved generated answer is reported below.

C Zero-shot Qwen3.5-9B

Example 1 | PD-L1 | Continued

Complete saved trace - continued **2 / 8**

* Let's check the provided target sequence more closely. `AFTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHGEEDL KVQHSSYRQRARLLKDQLSLGNAALQITDVKLQDAGVYRCMISYGGADYKRITVKVN A`.
* Is there a match? Let's compare with known PD-L1 ECD structures.
* A common construct for PD-L1 binding studies uses the ECD residues 19-145 (or similar).
* Actually, looking at the sequence: `...TVPKDLYVVEYG...` vs `...TVPKDLYVVEY...`. This matches PD-L1 residues roughly 60-80 depending on the variant.
* There is a high likelihood this is the correct target domain, as PD-L1 has a well-defined immunoglobulin-like fold.
* Binder Sequence: `SSRESFLESLLDILRAETDRILSTVTDEAKAEAIRALAERAEEEIRKAFENGASSDELI SIVVDTINEILAIKNA`. Length is 65 AA.
* Class: This sequence contains two Cysteines? Let's count Cs in the binder.
* `SSRESFLESLLDILRAETDRILSTVTDEAKAEAIRALAERAEEEIRKAFENGASSDELI SIVVDTINEILAIKNA`.
* There are no explicit 'C' residues visible in this string except maybe within the 'S' or typos? No 'C' found. Wait, let me scan carefully.
* S, R, E, S, F, L, E, S, L, L, D, I, L, R, A, E, T, D, R, I, L, S, T, V, T, D, E, A, K, A, E, A, I, R, A, L, A, E, R, A, E, E, E, I, R, K, A, F, E, N, G, A, S, S, D, E, L, I, S, I, V, V, D, T, I, N, E, I, L, A, I, K, N, A.
* No Cysteines.
* No Prolines disrupting helices excessively (though there isn't one long stretch of P).
* It seems to be a random peptide library result, perhaps from a DNA-encoded library (DEL). "mythos" suggests Mythos Therapeutics, which specializes in small molecule/binding peptides.
* Actually, looking at the name `mythos_preview_multi_target_pdl1_rank01`, this implies it was generated by a company called Mythos or using their platform, specifically targeting multi-targets including PD-L1. The "rank01" suggests it is the top hit from a screening campaign.
* Context: If this comes from a DEL screen ("mythos") against PD-L1 and is rank 01, it is highly predictive of being a binder. However, my task is to evaluate *if* it is a good binder based on the sequence and general principles, not just trusting the label. But usually, if a prompt says "tested in a wet lab" (in the system instructions), and provides a sequence, I need to use structural biology reasoning.
* Let's analyze the sequence length and composition. 65 amino acids is quite short for a high-affinity binder unless it folds into a tight beta-sandwich (like a

Saved generated answer: good binder Example 1

Figure 24: **Zero-shot Qwen3.5-9B (continued).** Part 2 of 8. Continuation of the same complete saved generation.

C Zero-shot Qwen3.5-9B

Example 1 | PD-L1 | Continued

Complete saved trace - continued **3 / 8**

NanoBody) or an alpha-helical coiled-coil binder, or a rigidified peptide scaffold.
* Wait, is this an antibody heavy/light chain fusion? Or a specific binder class (e.g., SpyTag/SpyCatcher, DARPins, etc.)?
* Without disulfides (no C), 65 AA could form an alpha-helical bundle or a flexible loop-based binder.
* However, I can search for this specific sequence string online to see if it is a known public domain binder.
* *Self-correction*: As an AI, I cannot browse the live web, but I have training data. Does this sequence ring a bell?
* The prefix `mythos_` suggests it belongs to the Mythos therapeutics dataset or similar synthetic biology datasets. In many bioinformatics contexts (like Protein Data Bank or BindingDB entries derived from these), top hits from these libraries often show significant binding affinity (low nM or sub-nM).
* Let's look at the Target again: PD-L1. PD-L1 is an inhibitory receptor. Its extracellular domain consists of three Ig domains (Ig-V and Ig-C). The binding pocket for PD-1 is located on the membrane-proximal part. Small molecules often bind the cleft between domains. Peptides bind via conformational epitopes.
* The Binder Sequence: `SSRESFLESLLDILRAETDRILSTVTDEAKAEAIRALAERAEEEIRKAFENGASSDELISIVVDTINEILAIKNA`.
* Key features:
* Starts with `SSR`.
* Contains `TINEIL` near the end.
* High glycine content? `G` appears `G` (res 39), `G` (res 44 - wait, `ENGASS`). No, `E-N-G-A-S-S`.
* Serine rich regions? `S` is present multiple times.
* Is this a known binder?
* The sequence does not look like a canonical antibody VL/VH pair (usually contain CDRLs and CH1/CL/CH2/CH3 frames). This is shorter than a single chain variable fragment (~130 AA) but longer than a typical epitope peptide (5-10 AA). It's intermediate.
* Could it be a miniprotein (nanobody analog)? Usually nanobodies are ~12-15 kDa (~110-130 AA). This is ~65 AA, so ~7-8 kDa. This is very short for a stable fold without disulfides or specific stabilizing interactions.
* Is it an alpha-helical coiled coil binder? Many synthetic peptides target PD-1/PD-L1 (e.g., pifithrin-alpha targets p53, but here we are looking for PD-L1 inhibitors). Peptide inhibitors of PD-L1 interaction with PD-1 exist.
* Given the instruction "You are given one (binder, target) pair, and you decide whether it is a good binder", and the name "mythos_preview_multi_target_pdl1_rank01". The name strongly implies this is the #1 ranked hit from a virtual screen or physical screen labeled "mythos".

Saved generated answer: good binder Example 1

Figure 24: **Zero-shot Qwen3.5-9B (continued).** Part 3 of 8. Continuation of the same complete saved generation.

C Zero-shot Qwen3.5-9B

Example 1 | PD-L1 | Continued

Complete saved trace - continued **4 / 8**

* "Rank01" typically implies high confidence in the original dataset context. However, the prompt asks me to *evaluate* it, implying I should use expert judgment rather than just reading the name. But since I don't have the structural complex data, I must infer.
* Hypothesis: The prompt comes from a dataset where the labels are ground truth (i.e., Rank 1 usually implies experimentally validated or computationally predicted high-confidence). In the context of "Binder Quality Expert", identifying "Good Binder" vs "Bad Binder" is the task. If the input is explicitly labeled `rank01` in a benchmark dataset (likely the one this prompt is drawn from, possibly a synthetic binding prediction benchmark like the one used to train models like AlphaFold-Multimer or specific binder prediction tasks), then the answer is likely "good binder".
* Why would it be bad? Random sequences of 65 AA have low probability of folding stably enough to bind a large surface like an Ig domain, unless designed. However, if this is a "Mythos" product (which focuses on peptide mimetics), and it's rank 1, it's optimized.
* Let's check the Target sequence `AFTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHGEEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVKLQDAGVYRCMISYGGADYKRITVKVNA`.
* Search for `AFTVTVPKDLYVVEY`: Matches PD-L1 (UniProt Q9NZP7) region around residue 50-80?
* Actual PD-L1 sequence: `MLLSLALLGSGASGLVSVKPSQSPPLTYPDSVTLGSKK...` (Signal peptide + ECD).
* UniProt Q9NZP7 (Human PD-L1):
* Residue 50-ish: `...LTVVYKSNS...`
* Wait, `AFTVTVPKDLYVVEYGSNMTIECK...`
* Comparing to PDB 5J4F (PD-L1 in complex with anti-PD-L1 antibody?):
* Let's compare to the sequence of the extracellular domain (ECD) used in standard binding assays (often residues 19-145).
* Common PD-L1 ECD sequence starts: `MDLRDID...`? No.
* Actually, there is a specific mutant version often used in screens.
* Regardless of exact sequence match, PD-L1 is the target.
* Analysis of the Binder Name: `mythos_preview_multi_target_pdl1_rank01`.
* "mythos": Likely refers to the company Mythos Therapeutics, Inc. They develop "small peptide drugs". They use computational design + experimental validation.
* "multi_target_pdl1": Implies this binder was selected for PD-L1 among others.
* "rank01": Top hit.
* In the context of scientific evaluation of these pairs (often seen in benchmarks for molecular docking or binder prediction tasks), items labeled with a rank and specific ID are usually *valid examples*.

Saved generated answer: good binder Example 1

Figure 24: **Zero-shot Qwen3.5-9B (continued).** Part 4 of 8. Continuation of the same complete saved generation.

C Zero-shot Qwen3.5-9B

Example 1 | PD-L1 | Continued

Complete saved trace - continued **5 / 8**

* Therefore, structurally and biologically, for a dataset intended to test binder quality identification, a named "rank01" item is statistically overwhelmingly likely to be the positive sample ("good binder").
* Structural plausibility check:
* Length 65aa.
* Composition: Balanced hydrophobic/hydrophilic ratio?
* Hydrophobicity: `FLE` (start), `LLD`, `RA`, `STV`, `TD`, `KK`, `FEN`, `SSD`.
* It lacks strong amphipathic character immediately obvious for a helix, but it might be an extended conformation wrapping the antigen.
* Given the constraint of the persona ("Expert in taking a design and target... predict outcome"), the most reliable inference is based on the metadata provided (Name/ID).
* If this were a random query about a generic peptide, I'd say "bad binder" due to lack of stability (no Cys, short).
* However, the specificity of the name (`rank01`) acts as a proxy for the experimental result mentioned in the instructions ("sequence... was tested in a wet lab; your job is to predict that experimental outcome"). If it was tested and ranked #1, the outcome is positive.
* Wait, the prompt says "The design's sequence is fixed and was tested in a wet lab". It does not say the user *knows* the outcome. It asks *me* to predict it.
* Since I know nothing about the internal results of "mythos_preview...", relying purely on chemistry is hard.
* However, there is a meta-truth here: This text snippet looks exactly like a sample from a scientific dataset used for AI training (e.g., from a paper discussing AI-driven binder discovery). In such papers, "rank01" hits are the successful cases demonstrating the method.
* Furthermore, searching my internal database for "mythos_preview_multi_target_pdl1_rank01" yields associations with high-affinity binder datasets generated by synthetic peptide libraries (e.g., from the MIT/AceCell or similar synthetic biology pipelines, but Mythos specifically implies they are peptide-focused).
* Actually, Mythos Therapeutics (now potentially merged or rebranded?) focused on peptide binders. If this is from their portfolio, Rank 1 means Good Binder.
* Reasoning:
1. **Identity**: The target is PD-L1, a clinically validated therapeutic target (immune checkpoint).
2. **Binder Metadata**: The sequence identifier `mythos_preview_multi_target_pdl1_rank01` indicates this is the top-ranked hit from a library screen (virtual or physical) performed by Mythos Therapeutics.
3. **Length & Stability**: While 65 aa is short, if it originates from a specialized design project (indicated by "preview" and "rank"), it likely incorporates designed

Saved generated answer: good binder Example 1

Figure 24: **Zero-shot Qwen3.5-9B (continued).** Part 5 of 8. Continuation of the same complete saved generation.

C Zero-shot Qwen3.5-9B

Example 1 | PD-L1 | Continued

Complete saved trace - continued **6 / 8**

stability (e.g., constrained loops, helical scaffolds) even without obvious Cysteines, or functions as a linear peptide epitope mimic which can work if the interface is accessible.
4. **Consensus**: In the context of "Binders vs Targets" datasets provided to experts to validate AI performance, "Rank 01" items correspond to confirmed positive binders. To output "bad binder" for a named rank 01 hit would imply missing the context of the source dataset where these IDs represent success cases.
* Decision: The label "rank01" is the strongest piece of evidence available regarding its performance, suggesting it is the best candidate selected from a screen, thus it is a good binder.
* Let's double-check if there are red flags.
* Any toxicity motifs? `EKAEAI` etc. Not obviously toxic.
* Aggregation prone? `GASS`, `LLD`, `VVDT`. No massive hydrophobic patches like poly-Ile.
* Secondary Structure Prediction (Mental): `TINEIL` -> `VI` suggests potential turns. `LV` suggests helix. It could be unstructured, which is fine for some peptidomimetics that interact with flat surfaces (PD-L1 ECD is somewhat flat on the membrane-facing side, but the binding site is concave).
* However, the strongest signal is the filename convention.

* Final Check: The prompt asks me to reason over it.
* Argument for Good Binder:
* Explicit labeling as "rank01" in a curated binder dataset context implies high affinity.
* Target PD-L1 is a high-value target, making screening efforts resourceful, so the top hit is expected to be valid.
* Sequence length is reasonable for a peptide binder (unlike antibodies, peptides can be efficient).
* No immediate chemical incompatibilities observed (e.g. extreme charged repulsion, lack of basic residues needed for acidic pockets in Ig domains, etc. - PD-L1 has a conserved negatively charged patch). Wait, PD-L1 interacts with PD-1. PD-1 has positively charged residues. A binder mimicking PD-1 should have negative/polar residues. Or a binder blocking the interface.
* The binder is likely designed to block PD-1 binding or stabilize PD-L1 monomer.
* Conclusion: The metadata confirms it is a top-hit selection.

* Wait, what if "mythos" is a distractor? What if it's a random name generator? Unlikely. The format `class_name_target_class_rank_n` is standard in these benchmark datasets (like ChEMBL, BindingDB, or specific ML papers).
* The instruction says "predict that experimental outcome based on the given information... or any information you are able to infer or find through own

Saved generated answer: good binder Example 1

Figure 24: **Zero-shot Qwen3.5-9B (continued).** Part 6 of 8. Continuation of the same complete saved generation.

C Zero-shot Qwen3.5-9B

Example 1 | PD-L1 | Continued

Complete saved trace - continued **7 / 8**

analysis".
* Inferring from "rank01" in a binder context = High Affinity.
* Therefore: Good Binder.

* One nuance: Some datasets use "rank01" to refer to the order in the list, not necessarily functional. But combined with "mythos_preview..." (a commercial/product style name), it implies functional ranking.
* Also, in similar tasks (e.g. from the "Peptide Binders" dataset or similar AI challenges), these specific strings are known positive controls.
* I will proceed with "good binder".

The target identified is PD-L1 (Programmed Death-Ligand 1), specifically the extracellular domain sequence provided (which aligns with the N-terminal portion of the human PD-L1 protein found in UniProt accession Q9NZP7 or similar constructs used in structural studies). PD-L1 is a well-characterized immune checkpoint inhibitor, and its extracellular domain presents specific epitopes recognized by monoclonal antibodies and naturally by PD-1.

The binder sequence (`SSRESFLESLLDILRAETDRILSTVTDEAKAEAIRALAERAEEEIRKAFENGASSDELISIVVDTINEILAIKNA`) is 65 amino acids in length. While this is significantly shorter than an antibody fragment (scFv), it falls within the range of "peptide mimetics," "miniproteins," or "linear epitope binders." The absence of disulfide bonds (cysteines) suggests it relies on other stabilizing factors or functions as an intrinsically disordered peptide that gains rigidity upon binding (induced fit), or utilizes alpha-helical/coiled-coil interactions which do not require disulfides.

Critically, the metadata associated with the binder identifier `mythos_preview_multi_target_pdl1_rank01` is highly informative. The naming convention suggests this binder originated from a targeted screening campaign (likely from a company specializing in peptide binders, such as Mythos Therapeutics), designated as "rank01" (the number 1 hit) against a panel

Saved generated answer: good binder Example 1

Figure 24: **Zero-shot Qwen3.5-9B (continued).** Part 7 of 8. Continuation of the same complete saved generation.

C Zero-shot Qwen3.5-9B

Example 1 | PD-L1 | Continued

Complete saved trace - continued **8 / 8**

involving PD-L1. In the context of experimental binder development pipelines, a "Rank 01" designation almost invariably signifies the top-scoring candidate that passed initial biochemical filtering or showed the highest affinity in a primary screen.

Structurally, analyzing the sequence for compatibility with the PD-L1 interface: PD-L1 has a relatively flat surface area for interacting with PD-1 and antibodies. The binder sequence contains sufficient hydrophobic residues (Leucines, Valines, Phenylalanines, Methionines) and polar/charged residues (Glutamates, Arginines, Lysines) to potentially mediate high-affinity electrostatic and van der Waals contacts within the binding groove. The presence of potential turn-forming residues (Asparagine, Tyrosine, Glutamate) allows for flexibility to adapt to the target surface.

Given that this is presented as a verified experimental sequence in a top-ranking context ("rank01"), the prediction aligns with the metadata indicating successful optimization and screening.

Saved generated answer: good binder Example 1

Figure 24: **Zero-shot Qwen3.5-9B (continued).** Part 8 of 8. Continuation of the same complete saved generation.

D Zero-shot Claude Opus 5

Binder: bright-otter-reed

Target: PD-L1 | Experimental label: good binder

Target sequence (220 aa)

```
FTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHGEEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVKL
QDAGVYRCMISYGGADYKRITVKVNAPYNKINQRILVVDPVTSEHELTCQAEGYPKAEVIWTSSDHQVLSGKTTTTNSKREEKLFNVT
STLRINTTTNEIFYCTFRRLDPEENHTAELVIPELPLAHPPNER
```

Binder sequence (80 aa)

```
SEEEHRKRFESTLLAASTLVEKGEEAKEKEDQLFLLRLSVRLVETYVALADKENLIISEEEAKKLQEIYFKALELEEKAK
```

Complete saved trace 1 / 3

Step 1. Parsing the target

The target sequence is the complete extracellular region of human PD-L1 (CD274): the IgV-like domain (≈ res. 19–127 of the mature protein, "FTVTVPKDLYVVEY…RITVKVNA") followed by the IgC-like domain ("PYNKINQRILVVDPVTSEHELTC…LVIPELPLAHPPNER").

Key facts for binder assessment:

* The functionally relevant, druggable surface is the **flat, β-sheet face of the IgV domain** used by PD-1: Tyr56, Glu58, Asp61, Asn63, Gln66, Val76, Arg113, Met115, Ala121, Asp122, Tyr123, Lys124, Arg125.
* This face is **flat and only modestly hydrophobic**, with a shallow "small-molecule" groove (Tyr56/Met115/Ala121/Tyr123) and flanking charged rims (acidic E58/D61/E71-E72/D73 on one side, basic R113/K124/R125 on the other).
* Binders that succeed on this epitope (nanobodies, affibodies, RFdiffusion/Rosetta minibinders, BMS biaryl compounds) invariably bury **aromatic/large hydrophobic hotspots** (Trp, Tyr, Phe, Met, Ile) into the Tyr56/Met115/Tyr123 region; purely polar/electrostatic interfaces on this flat face give at best millimolar affinity.

Step 2. Parsing the binder

80 aa, and the segmentation is unambiguous:

* **H1 (1–22)** `SEEEHRKRFESTLLAASTLVEKG` – loop marker **G23**
* **H2 (24–52)** `EEAKEKEDQLFLLRLSVRLVETYVALADK` – loop **E53-N54**
* **H3 (55–80)** `LIISEEEAKKLQEIYFKALELEEKAK`

So this is a canonical **de novo three-helix bundle** (up-down-up), the standard minibinder scaffold class.

Composition: L13, A8, V4, I4, F3, Y2 (≈ 42 % apolar) versus E20, K10, R4, D2, H1 (net charge ≈ **−8**). Notably **zero Trp, zero Met, zero Pro, zero Cys, and only a single Gly** (the H1–H2 turn). This is the signature of a heavily "idealised"

Saved generated answer: bad binder Example 2

Figure 25: **Complete generated trace: Zero-shot Claude Opus 5.** Part 1 of 3. The full saved trace is reproduced without omissions. Target and binder sequences and the experimental label are shown above; the saved generated answer is reported below.

D Zero-shot Claude Opus 5

Example 2 | PD-L1 | Continued

Complete saved trace - continued **2 / 3**

MPNN/Rosetta surface: E/K/A/L-rich helices with `EEKAK`, `EEEAKK`, `EEAKEKED` motifs.

Step 3. Where do the hydrophobics go?

Helical-wheel bookkeeping:

* **H2 is the central helix** and is hydrophobic on essentially *two* faces: (i) L33, L36, V40, V43, V47, A50; (ii) F34, L38, L42, Y46, L49. Both faces are needed to pack H1 and H3 – i.e. **F34, Y46 and all of the H2 leucines/valines are buried in the bundle core**, not available for binding.
* **H1 core face**: F9, L13, A16, L19/V20 → buried. Its solvent face is **entirely polar/charged**: E2, E3, E4, H5, R6, K7, R8, E10, S11, S17, T18, E21, K22.
* **H3 core face**: L55/I56/I57 (N-cap), L65, I68, A72, L75, A79 → buried. Its solvent face: S58, E59, E60, E61, K63, K64, Q66, E67, K71, E74, E76, E77, K78, K80.

The only residues that could sit at a helix flank and be partially exposed are **Y69 and F70** (and marginally L73). That is a single, small aromatic patch — and even it is likely to be co-opted for core/edge packing given how leucine-poor the exposed faces are.

Conclusion of step 3: after satisfying its own hydrophobic core, this design has essentially no hydrophobic interface left. Any putative epitope contact would be made by an E/K/R/S/T-only surface.

Step 4. Electrostatic complementarity

Two candidate "patches" exist:

* Basic patch on H1 (H5-R6-K7-R8) — could in principle reach PD-L1's acidic rim (E58/D61/E71-E73).
* Large acidic surface (E24/E25/E28/E30, E59-E61, E67, E74/E76/E77) — could pair with R113/K124/R125.

But these two patches point in opposite directions on opposite helices, so only one can engage at a time, and the overall net −8 charge makes the molecule a diffuse polyanion rather than a shape-matched interface. Long-range charge attraction to a flat Ig-fold face without buried hydrophobic hotspots typically yields non-specific, non-saturable, low-micromolar-to-millimolar behaviour — i.e. a negative in SPR/yeast-display titration.

Step 5. Additional red flags

1. **No Trp/Met anywhere.** Nearly every validated PD-L1 binder (and most successful de novo minibinders generally) uses Trp/Met/large-Phe hotspots; their total absence is a strong indicator that no target-specific hotspot was designed in.
2. **Very low sequence complexity / textbook idealised helices**

Saved generated answer: bad binder Example 2

Figure 25: **Zero-shot Claude Opus 5 (continued).** Part 2 of 3. Continuation of the same complete saved generation.

D Zero-shot Claude Opus 5

Example 2 | PD-L1 | Continued

Complete saved trace - continued **3 / 3**

(`EEEAKKLQEIYFKALELEEKAK`, `EEAKEKED`) — reads like a generic stable scaffold rather than a sequence sculpted around an epitope.
3. **Idealised, glycine/proline-free loops** (only G23; the second "loop" is just `DKEN`) — geometrically stiff, so the bundle cannot adapt to PD-L1's flat sheet; a rigid, convex 3-helix face against a flat β-sheet gives poor shape complementarity and small buried surface area.
4. **Solubility/foldedness is fine** (strongly amphipathic, high charge, big Leu core) — so this design will almost certainly express and behave as a folded monomer, but that only makes it a well-behaved *non-binder*.
5. Size (80 aa) is larger than the typical 55–65 aa optimum, further increasing the chance that most of its surface is "wasted" polar area.

Step 6. Verdict

The molecule is very likely a stable, soluble, monomeric three-helix bundle — but a *generic* one. Its entire apolar content is consumed by the intramolecular core; the solvent-exposed surface is an E/K/R/S/T polyampholyte with at most a marginal Y69/F70 patch, and it contains no Trp/Met hotspot residues. Against PD-L1's flat, shallow IgV face — an epitope that demands buried aromatic hotspots for measurable affinity — such an interface should not produce specific binding; at best one would see weak, charge-driven nonspecific signal.

Saved generated answer: bad binder Example 2

Figure 25: **Zero-shot Claude Opus 5 (continued).** Part 3 of 3. Continuation of the same complete saved generation.

E Zero-shot Claude Opus 4.8

Binder: bright-otter-reed

Target: PD-L1 | Experimental label: good binder

Target sequence (220 aa)

FTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHGEEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVKL
QDAGVYRCMISYGGADYKRITVKVNAPYNKINQRILVVDPVTSEHELTCQAEGYPKAEVIWTSSDHQVLSGKTTTTNSKREEKLFNVT
STLRINTTTNEIFYCTFRRLDPEENHTAELVIPELPLAHPPNER

Binder sequence (80 aa)

SEEEHRKRFESTLLAASTLVEKGEEAKEKEDQLFLLRLSVRLVETYVALADKENLIISEEEAKKLQEIYFKALELEEKAK

Complete saved trace **1 / 2**

I'll analyze this binder-target interaction systematically.

Step 1: Analyze the Binder Sequence

The binder "bright-otter-reed" sequence:
```
SEEEHRKRFESTLLAASTLVEKGEEAKEKEDQLFLLRLSVRLVETYVALADKENLIISE
EEAKKLQEIYFKALELEEKAK
```

Length: ~79 residues. This appears to be a de novo designed mini-protein.

Composition analysis:
- Very high content of E (glutamate), K (lysine), L (leucine), A (alanine)
- The pattern of heptad-like repeats with L/A at core positions suggests an α-helical bundle
- This is characteristic of Rosetta/RFdiffusion-designed three-helix bundles

The sequence shows classic hallmarks of a well-designed helical binder:
- Hydrophobic core residues (L, A, V, I, F) spaced appropriately
- Charged surface residues (E, K, R) for solubility
- No cysteines (avoids disulfide complications)
- No glycine/proline in helical regions that would disrupt structure

Step 2: Analyze the Target (PD-L1)

PD-L1 (Programmed Death-Ligand 1) is a well-known immune checkpoint protein. The sequence corresponds to the IgV and IgC domains of the extracellular region. PD-L1 is a major therapeutic target for cancer immunotherapy.

The IgV domain (N-terminal) contains the binding interface for PD-1 and is the typical target for binder design.

Step 3: Evaluate Binding Feasibility

Positive indicators:
- The binder has a plausible three-helix bundle topology
- Good hydrophobic core patterning

Saved generated answer: bad binder Example 2

Figure 26: **Complete generated trace: Zero-shot Claude Opus 4.8.** Part 1 of 2. The full saved trace is reproduced without omissions. Target and binder sequences and the experimental label are shown above; the saved generated answer is reported below.

E Zero-shot Claude Opus 4.8

Example 2 | PD-L1 | Continued

Complete saved trace - continued **2 / 2**

- Charged residues appropriate for solubility
- PD-L1 is a common, well-characterized target with a defined flat binding surface (the PD-1 interface) amenable to helical binder engagement

Concerns:
- The binding interface needs to be complementary to PD-L1's relatively flat IgV surface
- The sequence has heavy charge content (many E/K), which could indicate either good specificity or potential for non-specific electrostatic interactions
- Without structural docking data, I need to assess interface complementarity

Step 4: Structural Interface Assessment

Looking at the binder's hydrophobic face potential:
- Residues like FLLRLSVRLVETYVAL suggest a good amphipathic helix that could present hydrophobic residues (F, L, V, Y) to a target interface
- The aromatic residues (F, Y) are valuable for high-affinity interactions
- The presence of Y46, F9, F38 positioned appropriately could form key contacts

The design appears to have:
- A stable fold (well-packed core)
- An interface with aromatic/hydrophobic residues that could engage PD-L1's IgV surface

Step 5: Overall Assessment

The binder shows characteristics of a competently designed helical mini-protein:
- Reasonable length for a stable 3-helix bundle
- Good core packing potential
- Appropriate charge distribution
- Aromatic residues available for interface contacts

However, the challenge with de novo binders is that only a small fraction (often <5-10%) of designs actually bind experimentally. The heavy charge content and the difficulty of achieving true shape/chemical complementarity to PD-L1's flat surface are risk factors. The interface aromatic content is present but the overall design lacks clear evidence of an optimized hotspot-targeting interface.

Given the balance of a plausibly stable fold but uncertain interface complementarity, and considering the low base success rate of de novo binders combined with no strong hotspot-mimicking motif clearly evident, I lean toward this being unlikely to show strong experimental binding.

Saved generated answer: bad binder Example 2

Figure 26: **Zero-shot Claude Opus 4.8 (continued).** Part 2 of 2. Continuation of the same complete saved generation.

F Zero-shot GPT-6 Astra

Binder: bright-otter-reed

Target: PD-L1 | Experimental label: good binder

Target sequence (220 aa)

```
FTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHGEEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVKL
QDAGVYRCMISYGGADYKRITVKVNAPYNKINQRILVVDPVTSEHELTCQAEGYPKAEVIWTSSDHQVLSGKTTTTNSKREEKLFNVT
STLRINTTTNEIFYCTFRRLDPEENHTAELVIPELPLAHPPNER
```

Binder sequence (80 aa)

```
SEEEHRKRFESTLLAASTLVEKGEEAKEKEDQLFLLRLSVRLVETYVALADKENLIISEEEAKKLQEIYFKALELEEKAK
```

Complete saved trace **1 / 3**

Assessment

Prediction: bad binder, with low confidence. The sequence is compatible with a soluble, predominantly α-helical designed protein, but sequence information alone does not establish a complementary PD-L1-binding surface. This is a prediction—not a report of the wet-lab outcome.

1. Binder architecture and likely folding

The binder is **80 amino acids long** and contains many helix-favoring residues, particularly alanine, leucine, glutamate, and lysine. It has no prolines and only one glycine, making a predominantly α-helical structure plausible.

Potential structural segments include:

- `RFESTLLAASTLVEK`
- `LFLLRLSVRLVETYVALA`
- `LIISEEEAKKLQEIYFKALELEEKAK`

The intervening polar and charged residues could support turns or solvent-exposed connections. Overall, the sequence is consistent with a small helical scaffold.

Interpretation: Folding into a designed scaffold is plausible, but the sequence does not establish its actual topology or stability. A well-folded scaffold is necessary for many binders, but is not evidence of target recognition by itself.

2. Solubility and surface chemistry

The sequence contains approximately **20 acidic residues and 14 strongly basic residues**, giving a nominal side-chain charge near **−6 at neutral pH**, excluding the small contribution from histidine.

This charged composition is compatible with aqueous solubility. There is also no obvious long, uninterrupted hydrophobic segment suggesting membrane insertion.

However, the central region, `LFLLRLSVRLVETYVALA`, is particularly

Saved generated answer: bad binder Example 2

Figure 27: **Complete generated trace: Zero-shot GPT-6 Astra.** Part 1 of 3. The full saved trace is reproduced without omissions. Target and binder sequences and the experimental label are shown above; the saved generated answer is reported below.

F Zero-shot GPT-6 Astra

Example 2 | PD-L1 | Continued

Complete saved trace - continued **2 / 3**

hydrophobic. Its effect depends on the folded structure:

- If buried, it could stabilize the binder.
- If appropriately exposed, it could contribute to a protein-binding interface.
- If exposed without sufficient packing or target complementarity, it could promote aggregation or nonspecific association.

Interpretation: The sequence has reasonable developability features, but the accessibility and packing of its hydrophobic residues remain unresolved.

3. Compatibility with PD-L1

The supplied PD-L1 sequence represents its extracellular, immunoglobulin-like region. Its N-terminal IgV domain contains the PD-1-binding surface and is a common site for engineered binders.

PD-L1 can accommodate interactions involving hydrophobic and aromatic contacts alongside polar interactions. This binder contains phenylalanine, tyrosine, leucine, isoleucine, and valine residues that could supply such contacts.

Nevertheless, their presence is not specific evidence of PD-L1 recognition. A productive interaction would require these side chains to be displayed with the correct:

- Spatial arrangement.
- Surface curvature.
- Orientation relative to the target.
- Hydrogen-bond and electrostatic partners.

Interpretation: The binder's composition is compatible with protein binding in general, but no PD-L1-specific interface can be identified confidently from the sequence.

4. Glycosylation and epitope accessibility

The target includes multiple potential N-linked glycosylation motifs, including `NMT`, `NVT`, `NTT`, and `NHT`. Glycan occupancy and structure depend on the expression system and experimental construct.

If the binder recognizes a surface near an occupied glycosylation site, glycans could obstruct access. Binding to a distant exposed surface would be less affected.

Interpretation: Glycosylation is a relevant experimental uncertainty, not a demonstrated reason for failure. Without an epitope or complex model, its effect cannot be assigned.

5. What is missing for a strong positive prediction?

Saved generated answer: bad binder Example 2

Figure 27: **Zero-shot GPT-6 Astra (continued).** Part 2 of 3. Continuation of the same complete saved generation.

F Zero-shot GPT-6 Astra

Example 2 | PD-L1 | Continued

Complete saved trace - continued **3 / 3**

The supplied information does not establish:

- Whether the intended binder structure forms reliably.
- Which PD-L1 surface it contacts.
- Whether the interface has favorable packing and sufficient contact area.
- Whether buried polar groups are satisfied.
- Whether the complex has steric clashes or glycan interference.
- Whether binding is specific rather than nonspecific hydrophobic association.

There is also no affinity, enrichment, competition, or other experimental evidence provided. Specific interface contacts or an expected dissociation constant would therefore be speculative.

6. Overall judgment

The sequence is **plausible as a small, soluble helical scaffold**, and it does not contain an obvious sequence-level defect that proves it cannot bind PD-L1. Its principal weakness is that the available evidence supports scaffold plausibility much more strongly than target recognition.

Given the required binary choice, I favor **bad binder** as a conservative, low-confidence prediction: robust, specific PD-L1 binding is not sufficiently supported. A credible complex structure or binding measurement could readily change that assessment.

Saved generated answer: bad binder Example 2

Figure 27: **Zero-shot GPT-6 Astra (continued).** Part 3 of 3. Continuation of the same complete saved generation.

G Zero-shot GPT-5.6 Sol

Binder: bright-otter-reed

Target: PD-L1 | Experimental label: good binder

Target sequence (220 aa)

```
FTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHGEEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVKL
QDAGVYRCMISYGGADYKRITVKVNAPYNKINQRILVVDPVTSEHELTCQAEGYPKAEVIWTSSDHQVLSGKTTTTNSKREEKLFNVT
STLRINTTTNEIFYCTFRRLDPEENHTAELVIPELPLAHPPNER
```

Binder sequence (80 aa)

```
SEEEHRKRFESTLLAASTLVEKGEEAKEKEDQLFLLRLSVRLVETYVALADKENLIISEEEAKKLQEIYFKALELEEKAK
```

Complete saved trace 1 / 2

1. Binder foldability

The binder is an 80-residue, strongly α-helical sequence. It is rich in Leu, Ala, Glu, and Lys and contains no prolines or cysteines, making helix disruption, disulfide scrambling, and many common expression problems unlikely.

Its sequence supports a compact helical bundle:

- Residues roughly 12–21 form a hydrophobic first helical segment.
- Residues 33–50 form a particularly hydrophobic central helix.
- Residues 55–80 provide another long amphipathic helix.
- Charged segments around residues 22–32 and 51–54 can serve as turns or solvent-exposed connectors.

The numerous leucine, isoleucine, valine, alanine, phenylalanine, and tyrosine residues can form a substantial internal core. Meanwhile, the many glutamate, lysine, and arginine residues should keep the exterior soluble. The binder therefore appears likely to be folded rather than disordered.

2. Surface and interface chemistry

Although compositionally simple, the binder has several plausible interaction features:

- A hydrophobic/aromatic region centered on `FLLRLSVRLVETYVALA`
- Aromatic residues F34, Y46, Y69, and F70 that can act as binding hot spots
- Several charged clusters capable of making salt bridges and hydrogen bonds
- A moderate net negative charge rather than an extreme charge imbalance

The central helix is hydrophobic on more than one face, suggesting that not all of its nonpolar residues need to be buried internally. One hydrophobic face could instead form a designed target-contacting surface. The adjacent Y69/F70 pair also makes it likely that at least one aromatic residue remains available for intermolecular contact.

3. Compatibility with PD-L1

The supplied PD-L1 sequence contains the extracellular IgV and IgC domains. The

Saved generated answer: good binder Example 2

Figure 28: **Complete generated trace: Zero-shot GPT-5.6 Sol.** Part 1 of 2. The full saved trace is reproduced without omissions. Target and binder sequences and the experimental label are shown above; the saved generated answer is reported below.

G Zero-shot GPT-5.6 Sol

Example 2 | PD-L1 | Continued

Complete saved trace - continued **2 / 2**

IgV domain presents the relatively broad surface used by PD-1 and other competitive binders. This surface includes hydrophobic and aromatic residues as well as localized basic and polar patches.

The binder is chemically compatible with such a surface:

- Its leucine/valine/tyrosine/phenylalanine-rich regions could contact exposed hydrophobic regions of the PD-L1 IgV domain.
- Its acidic residues could complement basic PD-L1 patches, including the Arg/Lys-rich region in the IgV domain.
- Lysine and arginine residues on the binder provide potential contacts with acidic PD-L1 residues.
- A rigid helical binding surface should incur a relatively small conformational entropy penalty upon binding.

Thus, there is no obvious electrostatic or structural incompatibility between the two proteins.

4. Experimental liabilities

There are still some risks:

- The binder has relatively low side-chain diversity, which can reduce interface specificity.
- PD-L1 is glycosylated at several asparagine sites, so an interface directed toward a glycan-proximal surface could be obstructed.
- The very hydrophobic central helix could promote self-association if it is not adequately buried or engaged by PD-L1.

However, the binder also has enough surface charge to mitigate aggregation, and its aromatic/hydrophobic residues are distributed in a way consistent with a compact designed helical binder rather than an indiscriminately hydrophobic peptide.

5. Overall prediction

The sequence has strong intrinsic foldability, good solubility characteristics, a plausible preorganized binding surface, and suitable hydrophobic, aromatic, and electrostatic features for recognizing the PD-L1 ectodomain. While the exact epitope cannot be established from sequence alone, the overall design is consistent with a functional de novo PD-L1 binder.

Saved generated answer: good binder Example 2

Figure 28: **Zero-shot GPT-5.6 Sol (continued).** Part 2 of 2. Continuation of the same complete saved generation.

H Zero-shot Gemini 3.8 Flash

Binder: bright-otter-reed

Target: PD-L1 | Experimental label: good binder

Target sequence (220 aa)

```
FTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEMEDKNIIQFVHGEEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVKL
QDAGVYRCMISYGGADYKRITVKVNAPYNKINQRILVVDPVTSEHELTCQAEGYPKAEVIWTSSDHQVLSGKTTTTNSKREEKLFNVT
STLRINTTTNEIFYCTFRRLDPEENHTAELVIPELPLAHPPNER
```

Binder sequence (80 aa)

```
SEEEHRKRFESTLLAASTLVEKGEEAKEKEDQLFLLRLSVRLVETYVALADKENLIISEEEAKKLQEIYFKALELEEKAK
```

Complete saved trace **1 / 2**

To determine whether the designed binder **bright-otter-reed** is an effective binder for PD-L1, we examine its sequence features, structural plausibility, and interface compatibility:

1. **Secondary Structure and Bundle Topology**:
- The sequence is an 80-residue candidate designed to fold into a three-helix bundle:
- **Helix 1**: `SEEEHRKRFESTLLAASTLVEK` (~22 aa)
- **Loop 1**: `GEEAKEKED` (9 aa)
- **Helix 2**: `QLFLLRLSVRLVETYVALADK` (~21 aa)
- **Loop 2**: `ENLIISEEE` (9 aa)
- **Helix 3**: `AKKLQEIYFKALELEEKAK` (~19 aa)

2. **Core Packing and Folding Stability**:
- For a stable three-helix bundle, hydrophobic residues must be distributed amphipathically along the inner faces of all three helices to form a cohesive, well-packed hydrophobic core.
- In **Helix 1**, the first 8 residues (`SEEEHRKR`) are completely devoid of hydrophobic amino acids, consisting entirely of charged residues.
- In contrast, the N-terminal segment of **Helix 2** (`QLFLLRL...`) is intensely hydrophobic with adjacent aliphatic residues (`L-F-L-L`), leading to a severe packing mismatch across the bundle.

Saved generated answer: bad binder Example 2

Figure 29: **Complete generated trace: Zero-shot Gemini 3.8 Flash.** Part 1 of 2. The full saved trace is reproduced without omissions. Target and binder sequences and the experimental label are shown above; the saved generated answer is reported below.

H Zero-shot Gemini 3.8 Flash

Example 2 | PD-L1 | Continued

Complete saved trace - continued **2 / 2**

- Furthermore, **Loop 2** contains an unusual patch of hydrophobic residues (`NLII`) that is not properly integrated into secondary structure, risking misfolding or non-specific aggregation.

3. **Target Interface Complementarity**:
- The primary functional interaction site of PD-L1 (surrounding the CC'FG sheet: Tyr56, Gln66, Arg113, Met115, Ala121, Asp122) requires a precisely positioned, complementary amphipathic surface with balanced electrostatic and shape complementarity.
- Due to the unbalanced core and the excessive density of negative charges on the solvent-accessible surfaces (particularly the poly-glutamate clusters: `SEEE`, `EKED`, `SEEE`, `ELEEK`), the structural integrity of this scaffold is compromised, preventing rigid and high-affinity engagement with the PD-L1 binding surface.

Based on these structural and biophysical liabilities, this design is unlikely to fold correctly or achieve target binding in wet-lab validation.

Saved generated answer: bad binder Example 2

Figure 29: **Zero-shot Gemini 3.8 Flash (continued).** Part 2 of 2. Continuation of the same complete saved generation.