AbFUSE: Joint Antibody Optimization for Discriminating Similar Targets

Anonymous authors
Paper under double-blind review

Abstract

Advances in protein structure prediction enable model-guided antibody sequence optimization, yet suppressing binding to related negative targets can also reduce predicted binding to the intended target. We present AbFUSE, a training-free framework that optimizes one antibody sequence against a positive target and multiple negative targets using frozen structure and affinity predictors. A reference-score penalty limits decreases in predicted positive-target binding; an optional static local relation term compares mapped target neighborhoods. For chimeric antigen receptor (CAR) design, the objective can additionally include a sequence-level positively charged patch (PCP) term from CAR-Tonic Signal Tuner (CAR-Toner). We evaluate candidate quality on a computational benchmark and explore multi-target predictions in a case study involving STEAP family members. No updated AbFUSE benchmark candidate passed the prespecified interface-confidence gate. In an exploratory AlphaFold Server assessment, designs targeting STEAP family member 1 (STEAP1) showed lower predicted binder–target pairwise alignment error against STEAP1 than against STEAP2/3; this pattern also appeared in the baseline designs. These findings characterize the framework and its current validation limits without establishing experimental affinity or selectivity.

1 Introduction

Antibodies participate in immune defense by specifically recognizing antigens, while their binding domains are amenable to engineering, allowing this recognition ability to be linked to different effector functions and thereby providing a molecular basis for targeted therapy (1). Chimeric antigen receptor (CAR) T-cell (CAR-T) therapy is an important application: it uses an antibody-derived binding domain to recognize a cell-surface target antigen and connects recognition to T-cell activation and targeted killing through the receptor's intracellular signals (1; 2). Binding-domain forms such as single-chain variable fragments and single-domain antibodies provide different options for receptor design; their effects depend on both antigen recognition and the properties of the binding domains themselves. For example, positively charged patches (PCP) on the binding-domain surface can promote CAR aggregation and affect tonic signaling. Tonic signaling that is too weak may be unfavorable for the persistence of CAR-T cells, whereas signaling that is too strong may promote their exhaustion (3; 4). Therefore, antibody design for CAR-T needs to consider the effect of binding-domain properties on receptor behavior while maintaining the required antigen-recognition ability. This allows binding-domain sequence design to coordinate target recognition and cellular functional requirements at the molecular level.

Changing a binding-domain sequence can regulate its interactions with different proteins and its own molecular properties, so that an existing or newly designed antibody meets application requirements (3; 5). For CAR-T, these requirements include recognizing the target antigen, reducing binding to non-target proteins, and controlling surface properties that may affect tonic signaling. Because these three design requirements act on the same sequence, one substitution may simultaneously change positive-target contacts, non-target recognition, and surface charge. Therefore, task-guided antibody design needs to consider these requirements together during sequence updates, to prevent an improvement in one property from disrupting another required function.

This design problem takes a discrete amino-acid sequence as the editable object, and the design requirements involve the spatial structure formed by that sequence, its interactions with positive and

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.

negative targets, and an optional CAR-related sequence property. Structure and property prediction models provide computable ways to evaluate these requirements, connecting sequence changes to the task objectives (6; 7). With a continuous sequence representation and gradients of prediction models, the sequence can be evaluated and updated during the search (8; 9).

Joint optimization faces two immediate difficulties. Suppressing binding to related negative targets may lower predicted binding to the positive target, motivating a reference-score lower bound and, where residue correspondence is available, a local relation term. Binding scores also omit sequence properties relevant to CAR tonic signaling; an optional charged-patch objective can guide these properties while the same sequence is evaluated against all targets (3; 4).

AbFUSE keeps prediction-model parameters fixed while optimizing a shared antibody sequence against one positive and multiple negative targets. At the start of stage 3, the first hard-sequence positive-target PPAP evaluation establishes a reference score; a penalty and final-screening condition limit predicted score deterioration. The optional local relation term compares mapped positive- and negative-target neighborhoods, and a learned CAR-Toner objective can guide PCP toward a specified interval. When both hard-sequence quality gates pass, discrete candidates are reconstructed and rescored before final screening. A separate mutation-context two-site proposal module is implemented but was not enabled in the experiments reported here.

2 Related Work

2.1 Target-Conditioned De Novo Binder Design

BindCraft uses AlphaFold2 gradients to optimize binder sequences to form a predicted interface, then redesigns the sequence of residues outside the interface and predicts and screens candidates again. Its post-processing allows the formed interface to be fixed (9). Among antibody-specific methods, dynamic Multi-channel Equivariant grAph Network (dyMEAN) jointly designs complementarity-determining region (CDR) sequences and all-atom structures from the epitope and an incomplete antibody sequence (14); Germinal combines structure prediction with an antibody language model to generate epitope-directed CDRs on a specified framework (10). The main task of these works is to generate new target-conditioned binders; AbFUSE focuses on local modification of existing binding domains under positive- and negative-target and CAR-property requirements, while also supporting sequence design under a given framework.

2.2 Multi-Target and Selectivity Design

RESP2, an upgraded version of the Rapid Engineering System for Proteins (RESP), starts from existing antibodies and multi-antigen screening data, trains an affinity model with uncertainty estimation, and then uses stepwise mutation and simulated annealing to search for sequences that bind multiple antigen variants, while considering humanness and solubility during candidate screening (15). Odin-Multi separately predicts positive- and negative-target complexes for the same binder sequence, combines with weights the objectives of promoting positive-target binding and suppressing negative-target binding, and then predicts and screens discrete candidates again (11). RedNet encodes target side chains in fixed-backbone sequence design and uses contrastive decoding under positive- and negative-target conditions to improve selectivity (12). Odin-Multi's positive-target objective already includes interface contacts (11); AbFUSE additionally uses a reference positive-target score as a lower bound and can add a static local relation score when mapped target residues are available.

2.3 Binding Evaluation and CAR-Related Properties

The protein-protein affinity predictor (PPAP) combines sequence representations and interface contact-aware attention to predict protein-protein affinity and shows that its predictions can be used to enrich binder candidates (7); it provides an evaluation signal in this workflow. CAR-Tonic Signal Tuner (CAR-Toner) uses a fine-tuned Evolutionary Scale Modeling 2 (ESM2) model to predict a PCP score related to CAR tonic signaling and can guide point mutations to increase or decrease PCP; its original paper also notes that the effects of these adjustments on antigen-binding specificity and affinity still need verification (4). In the related task of enzyme design, CatESO incorporates

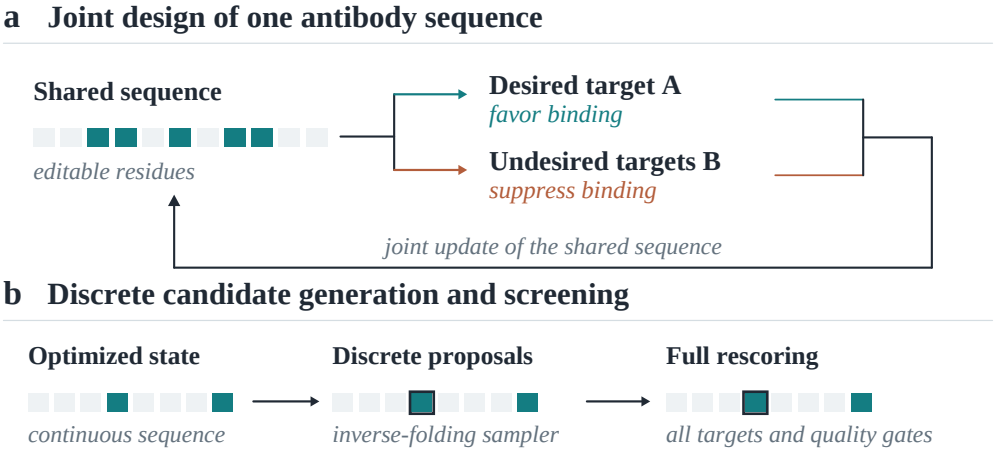

Figure 1: Overview of the AbFUSE workflow. Frozen structure and affinity predictors guide updates to one antibody sequence against the positive and negative targets. Candidate reconstruction and full hard rescoring occur only after both hard-sequence quality gates pass (Appendix Algorithm 1); the two-site proposal was disabled in the reported runs.

substrate-conditioned predicted k_{cat} into differentiable sequence optimization (13). AbFUSE includes PPAP predictions for positive and negative targets and deviation of PCP from an interval in the same sequence objective; these predictions are design proxies, and the actual binding and CAR functions of candidates still require experimental evaluation.

2.4 Two-Site Mutations and Context-Dependent Search

Combined mutations in antibody affinity depend on sequence context. Adams et al. observed second-order epistasis in an experimental affinity landscape for the 4-4-20 antibody and fluorescein and also reported cases in which individual substitutions were unfavorable but their combination was favorable (5). In model-guided search, multi-objective sequence design for proteins (MosPro) combines gradients of multiple property predictors into an update direction, samples discrete substitutions step by step, and continues the search in the updated sequence context (16); protein heuristic multi-site optimization (ProtHMSO) uses conditional substitution probabilities from a masked protein language model to narrow the multi-site candidate space, then screens by fitness score (17). Thus, conditional updates after mutation and multi-site screening have precedents. AbFUSE implements a mutation-context two-site proposal as an optional extension; it was disabled in the experiments reported here.

3 Method

AbFUSE takes an antibody parent sequence or framework, positive- and negative-target structures, and permitted editing positions. It optimizes a joint objective over a continuous sequence representation, uses a reference positive-target score to penalize deterioration, and can add a static local relation term. If both hard-sequence quality gates pass, discrete candidates are reconstructed and rescored; the optional two-site proposal is disabled in the reported runs. All prediction models remain fixed.

3.1 Sequence Representation and Target-Conditioned Prediction

Let the antibody sequence be x, the number of residues be L, the positive target be A, and the set of negative targets be $\mathcal{B} = \{B_1, \ldots, B_M\}$. Single-domain antibodies and heavy- and light-chain variable domains are each organized by residue position in the same sequence representation. Local

modification retains positions of the parent sequence that are not opened for editing; framework-conditioned design retains the specified framework and optimizes only the open regions.

For each editable position i, let Z_{ia} denote the optimizable parameter for amino acid a, and let $\mathcal{A}_i$ denote the set of amino acids allowed at that position. The amino-acid probability at that position is

$$P_{ia} = \frac{\exp(Z_{ia}/\tau)}{\sum_{b\in\mathcal{A}_i}\exp(Z_{ib}/\tau)}, \qquad a \in \mathcal{A}_i. \tag{1}$$

Here $\tau > 0$ is the temperature, and probabilities outside the set are zero; the set at a fixed position contains only the original amino acid. All targets share the same P. The frozen AlphaFold backend separately predicts complexes of the antibody with each target, and PPAP then calculates binding scores from the sequences and interface structures (7).

3.2 Joint Objective for Positive and Negative Targets

Let q_A and q_{B_j} be predicted binding scores on the pK_D scale, where larger values indicate stronger predicted binding. Negative-target scores are aggregated according to the declared task configuration:

$$\overline{q}_B = \begin{cases} \sum_{j=1}^{M}\omega_j q_{B_j}, & \text{mean mode,} \\ \tau_B \log\sum_{j=1}^{M}\exp(q_{B_j}/\tau_B), & \text{logsumexp mode,} \end{cases} \tag{2}$$

where $\omega_j \geq 0$, $\sum_j \omega_j = 1$, and $\tau_B > 0$. In the reported experiments, negative targets are used only in the STEAP task, where their scores are aggregated by logsumexp with $\tau_B = 0.1$. At the final selectivity weight, the joint loss is

$$\mathcal{L} = -q_A + \lambda_B\overline{q}_B - \lambda_R R_{\mathrm{rel}} + \lambda_K\mathcal{L}_{\mathrm{keep}} + \mathcal{L}_{\mathrm{car}} + \mathcal{L}_{\mathrm{aux}}. \tag{3}$$

The relation term has zero weight when disabled. $\mathcal{L}_{\mathrm{car}}$ is the configured charged-patch objective, either the learned CAR-Toner interval penalty or the separate analytic proxy; it is zero when neither is enabled. $\mathcal{L}_{\mathrm{aux}}$ groups structure-quality, geometry, and sequence-prior terms. In the reported four-stage runs, PPAP is inactive in stages 1–2; the positive-target reference is recorded before the first stage-3 update, when any configured negative-target and relation contributions enter at their full weights.

Affinity gradients are propagated through predicted coordinates to the sequence representation. Each backward pass fixes the residue identities, sequence features, and graph structure corresponding to the current discrete sequence; these inputs are reconstructed after the sequence changes. Therefore, the gradients are used for local updates in the current sequence context, whereas discrete candidates are scored after their complete inputs are recomputed.

3.3 R: Positive-Target Preservation and Compensation

R groups two distinct mechanisms intended to limit deterioration of predicted positive-target binding during negative-target suppression. First, the optimizer records the predicted positive-target binding score q_{ref} from the first stage-3 hard-sequence positive-target PPAP evaluation. With the configured tolerance ε_q, it penalizes a decrease below that score by

$$\mathcal{L}_{\mathrm{keep}}(q_A) = \left[q_{\mathrm{ref}} - \varepsilon_q - q_A\right]_+^2. \tag{4}$$

The reported runs use $\varepsilon_q = 0$. During final candidate screening, when a reference has been established, a candidate is also required to satisfy

$$q_A(x) \geq q_{\mathrm{ref}} - \varepsilon_q. \tag{5}$$

These operations preserve a predicted binding-score lower bound; they do not constrain a fixed set of interface contacts or a geometric pose.

Second, where enabled, a static local relation term compares the positive and negative target neighborhoods at mapped residues. It uses a coarse residue-charge descriptor ($+1$ for K/R, -1 for D/E, and 0 otherwise) and smooth proximity within an 8-Å radius to compute local support values u_j^A and u_k^B for mapped target residues $(j,k) \in \mathcal{M}$. The relation score is

$$R_{\mathrm{rel}} = \frac{1}{|\mathcal{M}|}\sum_{(j,k)\in\mathcal{M}}\frac{|c_j^A - c_k^B|}{2}\left(u_j^A - u_k^B\right), \tag{6}$$

where c_j^A and c_k^B are the coarse target-residue charges. The joint objective rewards a larger R_{rel} when this term is enabled. It is a dimensionless local proxy, not a predicted affinity or a fixed-reference-interface constraint. In the reported experiments it was enabled for the STEAP arms and disabled for the benchmark and case-study A configurations. A separate contact/geometry reference-interface controller exists as an optional mechanism but was disabled in all reported runs; no effect of that controller is claimed here.

3.4 Soft Guidance for CAR-Related Properties

CAR-Toner predicts PCP from the complete binding-domain sequence. Its study notes that effects of PCP optimization on antigen-binding affinity and specificity require further verification (4). In the learned-model arm, AbFUSE adds deviation from a specified PCP interval to the joint loss. The comparison arm instead uses an analytic, reference-relative surface-charge proxy, with its own final-screening requirement; it is not a no-charge-objective control.

CAR-Toner outputs $y(x)$ for the complete binding-domain sequence, corresponding to the base-2 logarithm of PCP. First recover the predicted value with $p(x) = 2^{y(x)}$, then define the target interval $[l_p, u_p]$, where $0 < l_p < u_p$. The normalized deviation loss is

$$\updownarrow_{PCP}(x) = \left(\frac{[l_p - p(x)]_+}{u_p - l_p}\right)^2 + \left(\frac{[p(x) - u_p]_+}{u_p - l_p}\right)^2, \qquad [z]_+ = \max(z, 0). \tag{7}$$

This loss is zero within the interval and increases with deviation outside it. It is evaluated once for the shared antibody sequence and is not a hard candidate-acceptance gate. The analytic proxy instead penalizes an increase in a residue-center positive-charge-patch score above a recorded reference and applies a corresponding screening bound. The proxy is not the CAR-Toner PCP prediction or a measurement of CAR signaling (4).

3.5 Optional Mutation-Context Two-Site Proposal

The implementation can extend the discrete candidate pool by measuring the gradient again after a first substitution, using the gradient change to rank second-site substitutions, and fully rescoring shortlisted double mutants. This module requires a positively eligible parent candidate and an additional model-evaluation budget. It was disabled in all benchmark, case-study A, and STEAP configurations reported here; accordingly, the experiments do not test its contribution. In the four-stage protocol, inverse folding and hard rescoring are entered only after Gates 1 and 2; no reported AbFUSE run reached that stage.

4 Experiments

4.1 Evaluation Protocol

We assess candidate quality on a 37-target positive-target-only benchmark and two single-target case studies (D-IL3 and D-PDL1). The STEAP case evaluates a STEAP1-positive/STEAP2–STEAP3-negative design setting. AbFUSE has two charged-patch conditions: a learned CAR-Toner interval objective and an analytic surface-charge proxy with a reference-relative screening bound. The latter remains an active objective, so this comparison does not isolate the effect of switching CAR-Toner on. The static relation term is enabled in the STEAP conditions and disabled in the other two settings; the contact/geometry controller and the mutation-context two-site module are disabled throughout. The benchmark and case study A use variable domains of heavy-chain-only antibodies (VHHs), whereas STEAP uses a 244-residue single-chain variable fragment (scFv) with a 14-position H3 editing mask.

The prespecified AbFUSE Gate 1 requires binder predicted local distance difference test (pLDDT) ≥ 50 and the campaign's active-interface confidence score (actifpTM) ≥ 0.5. We report the updated 148-slot AbFUSE benchmark ledger separately from the earlier frozen 370-slot ledger. Resource-censored slots remain in the prespecified denominator but do not count as scientifically evaluated

AbFUSE condition	Slots	Evaluated	Gate 1	actifpTM	pLDDT
Learned CAR-Toner	74	70	0	0.127	37.86
Analytic charge proxy	74	70	0	0.124	39.54

Table 1: Updated 37-target benchmark. The last two columns are means over the 70 evaluable slots per condition. Four slots per condition were resource-censored. Gate 1 requires both thresholds stated in the text.

failures. Other methods have different native selection rules, so their design-production counts are not treated as Gate-1 success rates.

For an exploratory assessment of STEAP designs, we submitted 25 designed scFvs against STEAP1, STEAP2, and STEAP3 to AlphaFold Server (AlphaFold 3; AF3), plus the unedited reference scFv against STEAP1 only. The ten AbFUSE sequences submitted for AF3 were exploratory outputs from attempts that did not pass Gate 1: eight from the native protocol and two from later amended search-budget runs; none was an accepted inverse-folding candidate. All 76 jobs returned. Each job used one seed that yielded five samples; the sample with the highest server ranking score was selected by a rule fixed before results were examined. STEAP1 was a 50-residue continuous construct, while STEAP2 and STEAP3 were full-length 490- and 488-residue sequences. The latter representations differ from the cropped or segmented negative targets used during design. We report the binder-chain interface predicted template modelling score (ipTM) and the minimum binder–target chain-pair predicted aligned error (PAE); for each design, the descriptive margin is $\Delta = \min(\mathrm{PAE}_{\mathrm{STEAP2}}, \mathrm{PAE}_{\mathrm{STEAP3}}) - \mathrm{PAE}_{\mathrm{STEAP1}}$. Positive Δ favors STEAP1 under this model score, but does not establish selective binding.

4.2 Benchmark and Native Screening Outcomes

In the updated benchmark, none of the 140 scientifically evaluable AbFUSE slots passed Gate 1 (Table 1). The limiting interface-confidence score was far below its 0.5 threshold in both conditions. Case study A similarly produced no Gate-1-passing candidate across its 50 updated slots. The frozen benchmark and historical case-study design counts use different protocols and outcome definitions and cannot be pooled with these Gate-1 results.

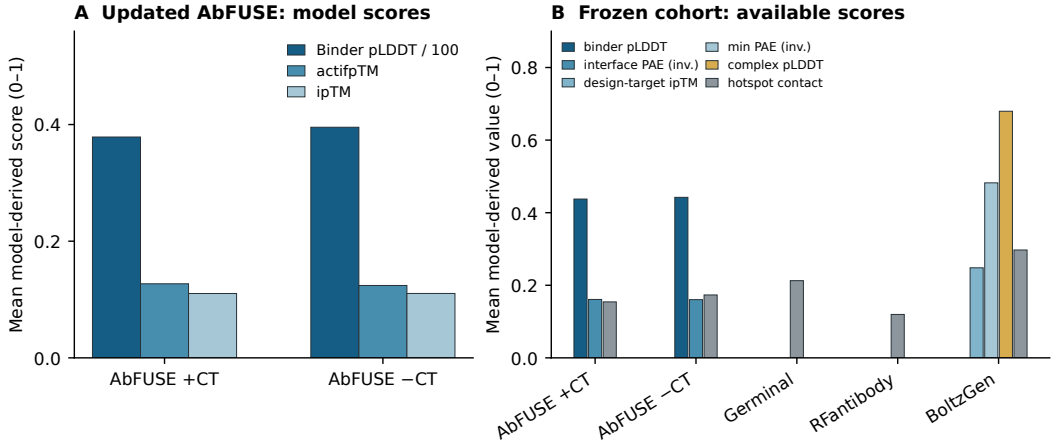

Figure 2: Benchmark model-score summaries. (A) Mean binder pLDDT (divided by 100), actifpTM, and ipTM for the updated AbFUSE arms. (B) Available frozen-cohort metrics on a 0–1 plotting scale: AbFUSE binder pLDDT and inverted interface PAE, BoltzGen design-to-target ipTM, inverted minimum PAE and complex pLDDT, and a common hotspot-contact fraction for all five arms. Metrics have different definitions, and the panels use different run ledgers; bar heights are not direct method rankings. Native pKd, energy, and H-bond scores have incompatible units and are not placed on this axis.

Design source ($n = 5$ each)	S1 ipTM	S1 PAE (Å)	Δ (Å)	PCP
AbFUSE, learned CAR-Toner	0.244	9.80	8.45	39.58
AbFUSE, analytic proxy	0.224	10.41	9.35	29.95
Germinal	0.366	6.49	12.02	41.73
RFantibody	0.278	7.31	8.75	35.24
BoltzGen	0.242	9.33	7.73	34.95

Table 2: Means for the 25 AF3-assessed STEAP designs. S1 denotes STEAP1. The PCP column was scored post hoc on the submitted scFv sequences with the same CAR-Toner checkpoint and is not an experimental measurement. The learned AbFUSE group includes two supplementary search-budget completions; these are excluded from the frozen five-arm native-protocol denominator.

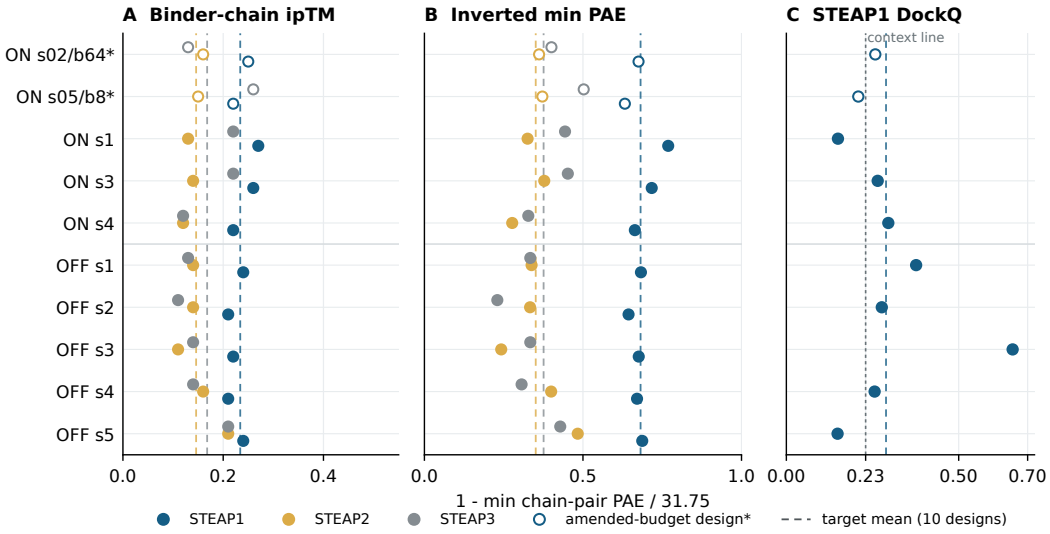

Figure 3: Exploratory AF3 predictions for ten AbFUSE STEAP designs. Each row is one scFv; blue, yellow, and gray indicate STEAP1, STEAP2, and STEAP3. (A) Binder-chain ipTM; (B) inverted minimum chain-pair PAE, scaled as $1 - \mathrm{PAE}/31.75$; (C) STEAP1 DockQ relative to the unedited scFv AF3 reference complex on STEAP1. Colored dashed lines are within-target means over all ten designs, including the two amended-budget sequences (hollow markers). The gray dashed 0.23 line in C is visual context, not an acceptance threshold. Each job used one AF3 seed; different target constructs limit cross-target interpretation, and these scores do not establish selective binding.

4.3 STEAP Predictions and Charged-Patch Scores

All 25 evaluated designs have positive model-derived Δ, but this pattern appears across all five methods (Table 2). The learned AbFUSE group is not the strongest on the reported AF3 interface-confidence proxies; Germinal has the highest mean STEAP1 binder-chain ipTM and Δ in this set. The frozen STEAP1 native protocol yielded no Gate-1-passing AbFUSE attempt in either condition (0/5 each); the two later AbFUSE designs were produced only after increasing the negative-pose search budget and also failed Gate 1. Germinal changed residues outside H3 in four of its five designs, so it is not a mask-matched H3-only comparator.

We additionally applied the CAR-Toner checkpoint after design to all 25 submitted scFv sequences and the unedited reference. The learned-objective AbFUSE group had a mean predicted PCP of 39.58 versus 29.95 for the analytic-proxy group; one of five learned-objective designs and none of the proxy designs lay in the declared 46–56 interval. The reference scFv scored 25.15. These are descriptive post-hoc scores: the groups differ in their objectives and screening rules, two learned-objective designs used amended search budgets, and no controlled comparison isolates the learned PCP term. No measured affinity, binding selectivity, or CAR signaling data are available. The single AF3 seed and unequal target representations further limit interpretation of the predicted STEAP margins.

REFERENCES

[1] Eshhar, Z., Waks, T., Gross, G., and Schindler, D. G. Specific activation and targeting of cytotoxic lymphocytes through chimeric single chains consisting of antibody-binding domains and the gamma or zeta subunits of the immunoglobulin and T-cell receptors. *Proceedings of the National Academy of Sciences of the United States of America*, 90(2), 720–724, 1993. https://doi.org/10.1073/pnas.90.2.720

[2] Porter, D. L., Levine, B. L., Kalos, M., Bagg, A., and June, C. H. Chimeric antigen receptor-modified T cells in chronic lymphoid leukemia. *New England Journal of Medicine*, 365(8), 725–733, 2011. https://doi.org/10.1056/NEJMoa1103849

[3] Chen, J., Qiu, S., Li, W., et al. Tuning charge density of chimeric antigen receptor optimizes tonic signaling and CAR-T cell fitness. *Cell Research*, 33, 341–354, 2023. https://doi.org/10.1038/s41422-023-00789-0

[4] Qiu, S., Chen, J., Wu, T., et al. CAR-Toner: an AI-driven approach for CAR tonic signaling prediction and optimization. *Cell Research*, 34, 386–388, 2024. https://www.nature.com/articles/s41422-024-00936-1

[5] Adams, R. M., Kinney, J. B., Walczak, A. M., and Mora, T. Epistasis in a fitness landscape defined by antibody-antigen binding free energy. *Cell Systems*, 8(1), 86–93.e3, 2019. https://doi.org/10.1016/j.cels.2018.12.004

[6] Jumper, J., Evans, R., Pritzel, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583–589, 2021. https://doi.org/10.1038/s41586-021-03819-2

[7] Qian, J., Yang, L., Duan, Z., Wang, R., and Qi, Y. PPAP: A Protein–protein Affinity Predictor Incorporating Interfacial Contact-Aware Attention. *Journal of Chemical Information and Modeling*, 65(19), 9987–9998, 2025. https://pubs.acs.org/doi/abs/10.1021/acs.jcim.5c01390

[8] Norn, C., Wicky, B. I. M., Juergens, D., et al. Protein sequence design by conformational landscape optimization. *Proceedings of the National Academy of Sciences of the United States of America*, 118(11), e2017228118, 2021. https://doi.org/10.1073/pnas.2017228118

[9] Pacesa, M., Nickel, L., Schellhaas, C., et al. One-shot design of functional protein binders with BindCraft. *Nature*, 646, 483–492, 2025. https://doi.org/10.1038/s41586-025-09429-6

[10] Mille-Fragoso, L. S., Driscoll, C. L., Wang, J. N., et al. Efficient generation of epitope-targeted antibodies with Germinal. *Nature Biotechnology*, 2026. https://www.nature.com/articles/s41587-026-03187-0

[11] Brasas, V., Christensen, C. R., Björnsson, K. H., et al. Specificity-driven protein binder design with Odin-Multi. *bioRxiv*, 2026, preprint. https://www.biorxiv.org/content/10.64898/2026.09.08.749745v1

[12] Xie, Z., and Xu, J. Redesign selective protein binders using contrastive decoding. *bioRxiv*, 2026, preprint. https://www.biorxiv.org/content/10.64898/2026.05.09.722041v1

[13] Gan, Z., Xu, Y., Xu, J., et al. CatESO: Differentiable Enzyme Sequence Optimization Guided by Substrate-Aware k_{cat} Prediction. *Journal of the American Chemical Society*, 2026. https://doi.org/10.1021/jacs.6c14365

[14] Kong, X., Huang, W., and Liu, Y. End-to-End Full-Atom Antibody Design. *Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research*, 202, 17409–17429, 2023. https://proceedings.mlr.press/v202/kong23c.html

[15] Parkinson, J., Hard, R., Ko, Y. S., and Wang, W. RESP2: An Uncertainty Aware Multi-Target Multi-Property Optimization AI Pipeline for Antibody Discovery. *Advanced Science*, 12(38), e04350, 2025. https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202504350

[16] Luo, J., Ding, K., and Luo, Y. Pareto-optimal sampling for multi-objective protein sequence design. *iScience*, 28(3), 112119, 2025. https://doi.org/10.1016/j.isci.2025.112119

[17] Wang, L., Wang, Y., Qiu, C., Xiao, L., Liu, X., and Chen, J. Heuristic multi-site optimization for protein sequence design using Masked Protein Language Models. *PLOS Computational Biology*, 22(6), e1014365, 2026. https://doi.org/10.1371/journal.pcbi.1014365

AI USE STATEMENT

Generative AI tools assisted with research brainstorming and organization, discussion of methodological choices, code implementation and review, translation and drafting/editing of manuscript text, assembly and checking of data summaries, interpretation of supplied results, and preparation of figures and tables. AI tools also assisted with inference-only, post-hoc CAR-Toner scoring of existing STEAP sequences. The authors are responsible for validating and interpreting the experimental records and for the final manuscript, claims, and artifacts.

A ADDITIONAL EXPERIMENTAL DETAILS

A.1 FOUR-STAGE OPTIMIZATION AND CANDIDATE FLOW

Algorithm 1 summarizes the configured four-stage procedure. It shows when the positive-target reference is recorded and when candidate reconstruction would be reached after both gates pass; no reported AbFUSE attempt passed Gate 1. The configured objective terms and screening thresholds are described in Section 3.

A.2 RUN CONFIGURATIONS AND DENOMINATORS

The updated benchmark contains 37 positive-target tasks and two prespecified AbFUSE conditions per task and seed, for 74 slots per condition. The supplied benchmark configurations are positive-target-only designs; their negative-target aggregation and local relation terms are inactive because no negative target is specified for those tasks. Case study A likewise comprises positive-target-only VHH designs for D-IL3 and D-PDL1, with 25 slots per antigen. Thus the benchmark and case study A test candidate production and the quality gate, whereas the STEAP case is the reported positive-versus-negative-target design setting. Their results should not be combined into a single estimate of multi-target selectivity.

All three updated AbFUSE run families use a 140-update, four-stage schedule (75, 45, 5, and 15 updates). The first two stages use soft sequence probabilities; the last two use a hard forward sequence with a straight-through gradient surrogate. The frozen structure model is AlphaFold2 multimer-v3 with zero outer recycles; PPAP and AbMPNN are used in the later affinity and candidate-reconstruction stages, respectively. The benchmark and case study A aggregate negative scores by an equal-weight mean when negative targets exist; STEAP uses logsumexp at temperature 0.1. The reported STEAP design uses a 244-residue scFv with 14 editable H3 positions, STEAP1 as the positive target, and STEAP2/3 as negatives. The static relation term is enabled only in STEAP. The contact/geometry controller and the mutation-context two-site module are disabled in all these runs.

The two STEAP conditions are not a single-factor ablation. The learned condition uses the CAR-Toner interval objective, while the comparison condition uses an active analytic surface-charge proxy and a reference-relative screening requirement. Both configure positive-score preservation and the static relation term. The 46–56 CAR-Toner interval is a model-score design target, not a biological success threshold or a hard acceptance gate.

Algorithm 1 Four-stage AbFUSE optimization and conditional candidate screening

Require: Positive target A, negative targets $\mathcal{B}$, editable sequence positions, stage schedule, screening rules

for each attempt permitted by the trajectory and resource budgets **do**
  Initialize shared sequence logits Z and an unset reference score q_{ref}
  for $s = 1, 2, 3, 4$ **do**
    for each update in stage s **do**
      Set the scheduled temperature and representation: soft for $s \leq 2$, hard straight-through for $s \geq 3$
      if $s = 3$ and this is its first update **then**
        Record q_{ref} from a hard-sequence positive-target complex and PPAP evaluation
      end if
      Evaluate the stage objective with frozen predictors (PPAP is active only for $s \geq 3$) and update Z
    end for
    if $s \geq 3$ **then**
      Freshly predict the hard positive-target complex and apply Gate $(s - 2)$
      if the gate fails **then**
        break the stage loop; this attempt yields no candidates
      end if
    end if
  end for
  if Gate 2 passed **then**
    Sample candidates with AbMPNN from the passing Gate-2 backbone
    for each candidate sequence x **do**
      Recompute hard positive- and negative-target complexes and configured negative poses
      Apply Gate 3 and the configured score, sequence, and property screening rules
    end for
    if any candidate is accepted **then**
      return accepted candidates
    end if
  end if
end for
return no accepted candidate within the budgets

Run family	Slots	Evaluated	Gate 1	Other	Task
Benchmark, learned	74	70	0	4 resource	Positive only
Benchmark, proxy	74	70	0	4 resource	Positive only
Case A, D-IL3	25	25	0	0	Positive only
Case A, D-PDL1	25	25	0	0	Positive only
STEAP, learned	5	3	0	2 technical	Three gate failures
STEAP, proxy	5	5	0	0	Five gate failures

Table 3: Prespecified AbFUSE slots and native Gate-1 outcomes. Resource-censored benchmark slots and technical STEAP failures did not receive an evaluable Gate-1 score and are not treated as measured negative results. For STEAP, the two later search-budget designs are excluded from this native-protocol table.

B STEAP Design-Level Predictions

Table 4 lists the selected AlphaFold Server sample for each of the 25 submitted designs. For every design–target job, the sample with the greatest server ranking score among five samples was selected. One seed (101) was used per job, so the table does not support sampling-variance estimates. P_1, P_2, and P_3 are minimum binder–target chain-pair PAE values for STEAP1, STEAP2, and STEAP3, respectively; $\Delta = \min(P_2, P_3) - P_1$. The ipTM column is the STEAP1 binder-chain ipTM. PCP is an inference-only, post-hoc CAR-Toner prediction from the submitted scFv sequence, not a measured CAR response. The unedited reference scFv has a STEAP1 job and PCP score 25.15 but lacks STEAP2/3 jobs, so no reference Δ is reported.

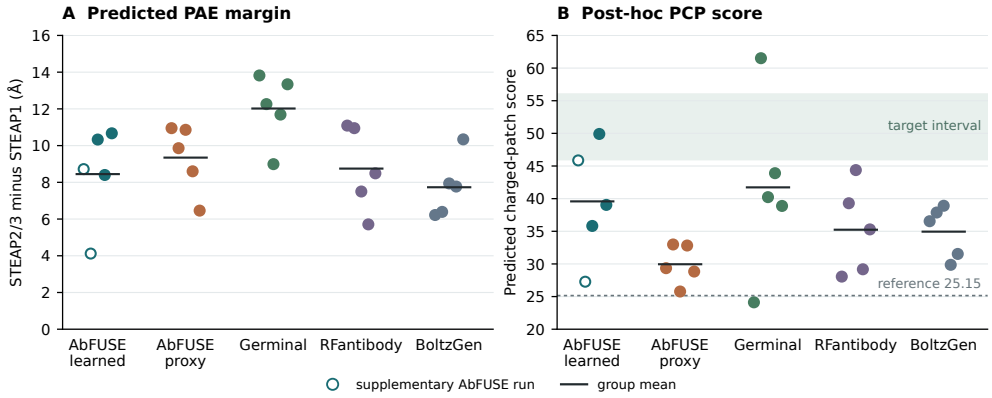

Figure 4: Design-level STEAP diagnostics. (A) Model-derived PAE margin for all 25 designs; higher values indicate a larger predicted difference between STEAP1 and the two negative targets. (B) Post-hoc CAR-Toner PCP scores, with the declared 46–56 interval shaded and the unedited reference indicated by a dashed line. Hollow points are the two supplementary AbFUSE designs; bars mark group means. The groups differ in protocol, and no error bars are inferred from a single AF3 seed.

Source	Design	S1 ipTM	P_1 (Å)	P_2 (Å)	P_3 (Å)	Δ (Å)	PCP
AbFUSE learned	s1	0.27	7.34	21.42	17.67	10.33	35.82
AbFUSE learned	s3	0.26	8.99	19.75	17.39	8.40	49.91
AbFUSE learned	s4	0.22	10.68	22.96	21.35	10.67	39.04
AbFUSE learned	s02/budget64*	0.25	10.31	20.27	19.03	8.72	45.86
AbFUSE learned	s05/budget8*	0.22	11.68	19.94	15.80	4.12	27.28
AbFUSE proxy	s1	0.24	10.07	21.02	21.15	10.95	29.35
AbFUSE proxy	s2	0.21	11.31	21.17	24.45	9.86	32.98
AbFUSE proxy	s3	0.22	10.29	24.06	21.15	10.86	25.78
AbFUSE proxy	s4	0.21	10.46	19.06	22.02	8.60	32.82
AbFUSE proxy	s5	0.24	9.94	16.40	18.15	6.46	28.83
Germinal	s1	0.38	6.53	23.54	20.35	13.82	24.12
Germinal	s2	0.49	4.74	18.57	17.00	12.26	61.51
Germinal	s3	0.25	8.19	17.18	20.24	8.99	40.23
Germinal	s4	0.31	6.76	19.92	18.45	11.69	43.90
Germinal	s5	0.40	6.23	22.05	19.57	13.34	38.88
RFantibody	s1	0.24	10.04	21.13	22.69	11.09	28.06
RFantibody	s2	0.28	7.00	17.95	21.09	10.95	39.30
RFantibody	s3	0.27	8.33	15.83	20.17	7.50	44.38
RFantibody	s4	0.30	5.30	11.01	18.54	5.71	29.18
RFantibody	s5	0.30	5.88	20.51	14.37	8.49	35.29
BoltzGen	s1	0.23	10.30	16.52	18.96	6.22	36.54
BoltzGen	s2	0.26	8.18	14.57	20.63	6.39	37.87
BoltzGen	s3	0.27	7.89	15.83	20.26	7.94	38.91
BoltzGen	s4	0.22	10.47	18.24	20.31	7.77	29.86
BoltzGen	s5	0.23	9.83	22.06	20.17	10.34	31.53

Table 4: Selected model scores for all 25 STEAP designs. *Two AbFUSE learned designs were added after the native five-arm run with amended search budgets; they are not native-protocol successes. Values are model predictions, and the different positive and negative target representations limit their direct comparison.

C Module Status and Evaluation Limits

The scalar positive-score reference penalty is configured in the reported AbFUSE runs. The static local relation score is configured only for STEAP and uses the declared STEAP1–STEAP2 residue correspondence; STEAP3 has no correspondence annotation and remains a separate negative-target scoring branch. The independent contact/geometry reference-interface controller and the two-site mutation-context extension are implemented options but were not run in the reported benchmark, case-study A, or STEAP configurations. Their effects therefore cannot be inferred from these results.

The benchmark and case study A are positive-target-only evaluations. Their zero Gate-1 pass counts identify a candidate-quality bottleneck under the declared gate, but do not quantify cross-reactivity suppression. The STEAP AF3 results are exploratory model assessments: STEAP1 was represented

by a 50-residue construct, while STEAP2/3 used full-length 490-/488-residue sequences; only one AF3 seed was submitted for each job. The parental scFv lacks negative-target AF3 controls, and four of five Germinal designs change residues outside the H3 mask. No measured binding, selectivity, or CAR signaling data are available. Accordingly, these results should be read as predictor outputs rather than evidence of functional antibody performance.