An honest cross-platform comparison of de novo binder design: RFantibody VHHs versus Biohub ESMFold2 minibinders and scFvs against ten challenging cancer antigens

Ashish (@tp53)

Inventcures • inventcures.github.io

Preprint — 4 June 2026. Designs, scores, and code are openly released (§7).

Abstract

We compare two de novo binder design platforms on a shared panel of structurally challenging cancer antigens: **RFantibody** (RFdiffusion → ProteinMPNN → RoseTTAFold2), which we previously ran to design one lead VHH per target, and **Biohub's ESMFold2** binder-design protocol, which designs binders by gradient descent over the binder sequence against an ESMFold2 structure objective, regularized by a down-weighted ESMC-6B language-model term. Running the Biohub protocol on Modal H200 GPUs, we generated minibinders and scFvs for 14 antigens (the 10 RFantibody targets plus 4 additional oncology targets) and scored all designs with ESMFold2 interface confidence (*ipTM*). On every one of the 10 shared targets, Biohub designs reach far higher ESMFold2 *ipTM* than the RFantibody VHH leads (median best *ipTM* 0.85 for both minibinders and scFvs, versus 0.13 for RFantibody VHHs). **We argue this gap is largely an artifact of evaluation circularity**, not a demonstration of superiority: ESMFold2 *ipTM* is precisely the objective the Biohub protocol optimizes, while RFantibody was optimized against RoseTTAFold2 and never "saw" ESMFold2 — the mirror image of our earlier finding that RFantibody leads pass RoseTTAFold2's own filters yet score near zero on ESMFold2. We report the full numbers, the engineering required to run the protocol reproducibly, and — crucially — re-score the best design per platform with **Boltz-2**, an open AlphaFold3-class oracle that neither platform optimized against. Under this neutral judge the gap roughly halves (RFantibody median *ipTM* rises 0.13 → 0.37; Biohub falls to 0.69–0.74), yet Biohub's best design still leads on all 10 shared targets by smaller, more variable margins. We conclude the Biohub advantage is real but modest once home-oracle inflation is removed — and is subject to a best-of-8 versus best-of-1 sampling asymmetry. We release all designs, structures, and code.

1 Introduction

De novo antibody and binder design has progressed from backbone-diffusion pipelines to single-model, all-atom approaches. **RFantibody** [1] couples RFdiffusion backbone generation, ProteinMPNN sequence design, and RoseTTAFold2 (RF2) structure prediction/filtering; in earlier work we applied it to ten "challenging but validated" cancer antigens and selected one best VHH lead per target by RF2 confidence. A newer paradigm, exemplified by Biohub's ESMFold2 family [2], treats design as direct gradient optimization of the binder sequence against a learned structure-prediction model: the **Biohub ESMFold2** binder-design protocol relaxes a binder prompt (a minibinder scaffold or an antibody VH–VL framework) and descends an

ESMFold2 interface objective (predicted inter-/intra-chain contacts), regularized by a *down-weighted* ESMC-6B protein-language-model pseudo-perplexity term. ESMFold2 is the design model; ESMC-6B is an auxiliary sequence prior.

A natural question is whether the newer, single-model approach produces "better" binders than the diffusion pipeline. Answering it cleanly is harder than it looks, because each platform is optimized against its *own* structure oracle. This paper makes that tension the central object of study: we run both platforms on a shared antigen panel, score with a common ESMFold2 metric, report the (large) apparent advantage of the Biohub designs, and then show why that advantage cannot be taken at face value.

2 Methods

2.1 RFantibody designs (baseline)

The RFantibody campaigns and their selection of one lead VHH per target are described in our prior report; here we reuse those leads unchanged. Each lead was generated by the standard three-stage pipeline and ranked by RF2 metrics (interaction-pAE, predicted LDDT, CDR RMSD).

2.2 Biohub ESMFold2 binder design

We deployed Biohub's `binder_design.py` protocol on Modal. For each target we ran two modalities: **minibinders** (single chain, 60–200 aa, `is_antibody=False`) and **scFvs** (the `trastuzumab_framework_vh` VH–VL framework with designed CDRs, `is_antibody=True`). The protocol performs ∼150 gradient steps on the binder sequence: each step folds the complex with an `ESMFold2-Experimental-Fast` inversion model and descends its structure loss (weight 1.0), plus a down-weighted ESMC-6B pseudo-perplexity gradient (weight 0.05 for scFv/antibody, 0.15 for minibinder). The final design is then scored by an ensemble of four ESMFold2-Experimental critic models (the reported *ipTM*). Hardware: one NVIDIA H200 per job; the ∼50–75 GB model footprint mandates small batches (see §5). All artifacts — per-critic *ipTM*, distogram and CDR-distogram *ipTM* proxies, the full loss trajectory, designed sequences, and predicted complex structures (mmCIF) — are persisted to a Modal volume for downstream analysis.

2.3 Targets and antigen sourcing

The panel is the 10 RFantibody oncology targets (B7-H3, CD47, CEACAM5, EGFR, EGFRvIII, EphA2, GPC2, HER2 domain IV, MSLN N-terminal, MSLN C-terminal) plus 4 additional oncology targets (KRAS, TNFSF9/4-1BBL, HDAC8, PHD2/EGLN1). For the RFantibody targets we used the same epitope-truncated antigen the pipeline designed against (Biohub design conditions on the target sequence, so this keeps the antigen matched). **Data-quality note:** two PDB identifiers in our internal target list were incorrect (6W0B is a KcsA potassium channel, not HDAC8; 6R8H is a triosephosphate isomerase, not PHD2); we substituted canonical UniProt sequences (HDAC8 Q9BY41, PHD2 Q9GZT9) and sourced KRAS from 7R0M and TNFSF9 from P41273. The MSLN C-terminal "antigen" is a 15-residue peptide; *ipTM* on such a fragment is unreliable for both platforms and we flag it throughout.

2.4 Scoring

We score interface confidence with ESMFold2 *ipTM* (0–1; higher = more confident interface). RFantibody VHH leads were scored with the hosted `esmfold2-fast-2026-05` model; Biohub designs carry *ipTM* from their ESMFold2-Experimental critics. Both are ESMFold2-family but different variants, so the two columns are *not strictly commensurate* — a caveat we return to. For each Biohub target×modality we report the best single-critic *ipTM* and the best per-design *ipTM* (mean across the four critics).

3 Results

3.1 Biohub designs score much higher ESMFold2 ipTM on every shared target

Figure 1 shows best ESMFold2 *ipTM* per target. The RFantibody VHH leads sit at *ipTM* 0.08–0.50 (median 0.13); Biohub minibinders reach 0.75–0.98 (median 0.85) and scFvs 0.42–0.90 (median 0.85). Biohub's best design exceeds the RFantibody lead on **10/10** shared targets, often by 0.6–0.9 *ipTM* units.

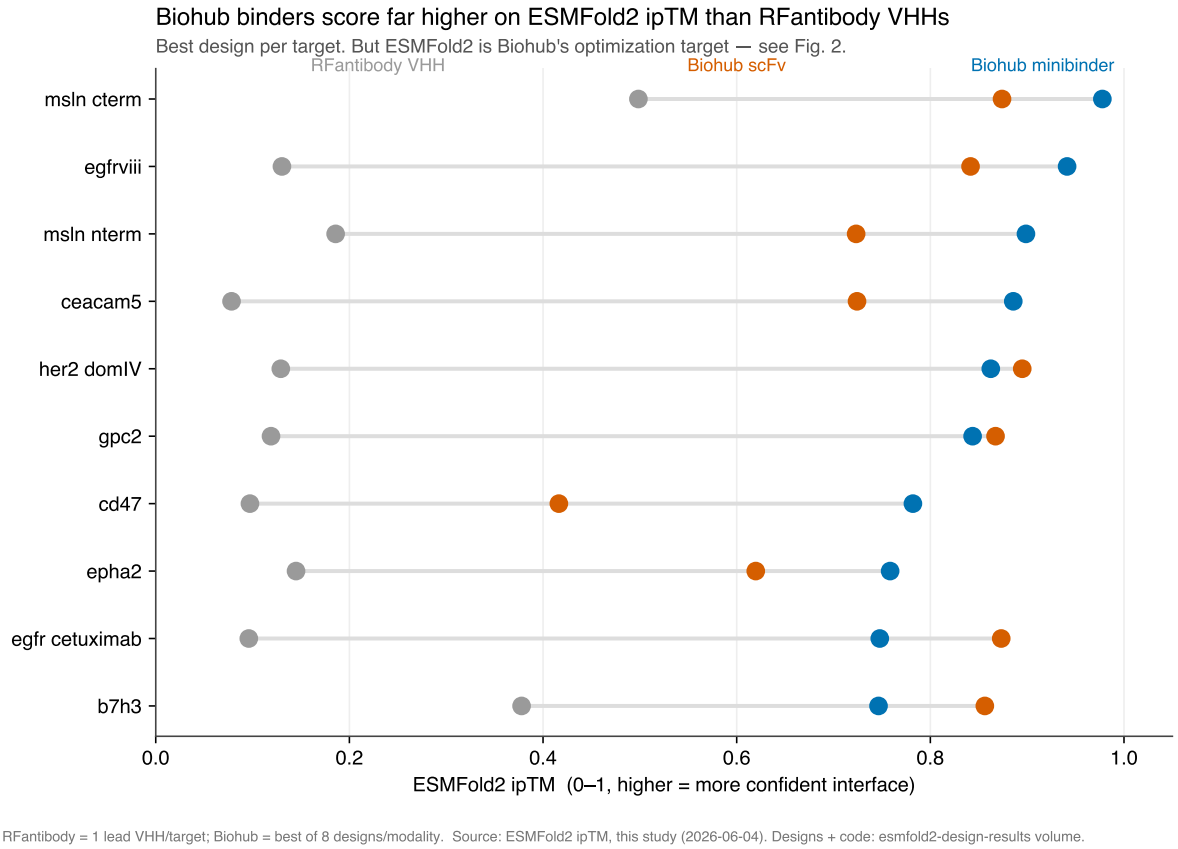

Figure 1. Best ESMFold2 *ipTM* per target for the three design classes. Targets sorted by Biohub minibinder *ipTM*. Points are directly labelled; the grey baseline is RFantibody. See Fig. 2 before interpreting the gap as quality.

3.2 The gap is evaluation circularity, not established superiority

Figure 2 pools the same *ipTM* values. The separation is stark — but ESMFold2 *ipTM* is exactly what the Biohub protocol optimizes, whereas RFantibody was optimized against RoseTTAFold2 and never used ESMFold2. This is the precise mirror of our earlier result: the same RFantibody VHH leads pass RoseTTAFold2's own confidence filters (predicted LDDT ≈ 0.90, interaction-pAE < 3.4 on 8/10 targets) yet collapse to *ipTM* ≈ 0.1 under ESMFold2. **Each platform looks excellent under the oracle that built it.** A high home-oracle score is therefore expected by construction and is not evidence that one platform's binders are more likely to bind in the wet lab.

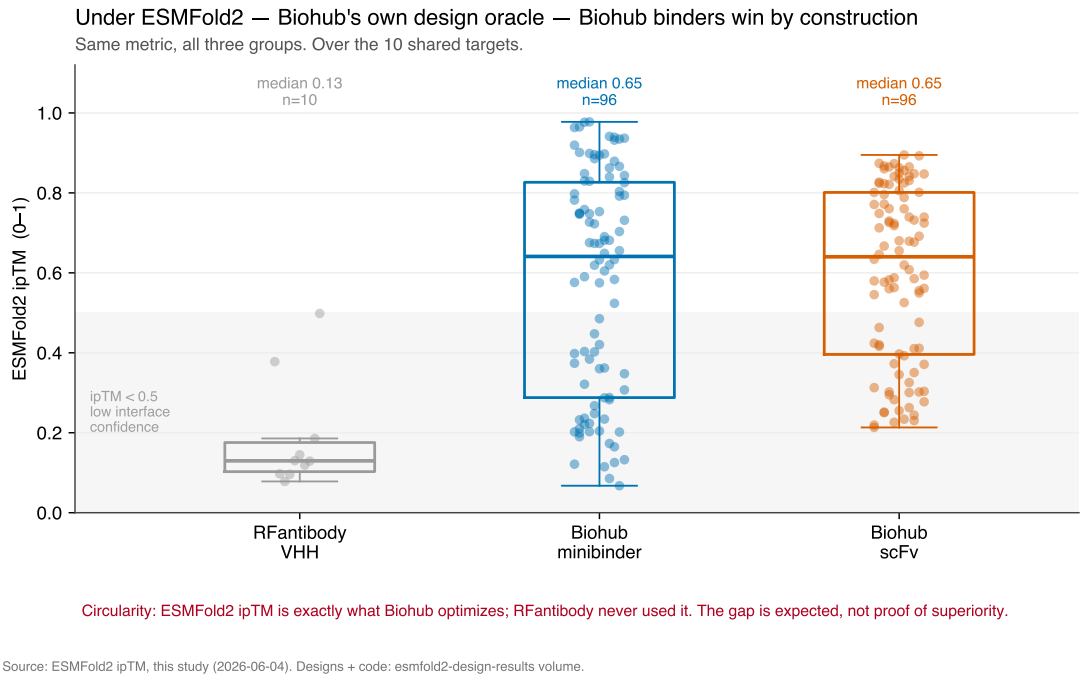

Figure 2. Pooled ESMFold2 *ipTM* for RFantibody VHH leads ($n = 10$, one per target) versus all Biohub minibinder and scFv designs over the same 10 targets. The shaded band marks low interface confidence ($ipTM < 0.5$). The red note states the circularity caveat.

3.3 Minibinders generally outscore scFvs

Across all 14 targets, minibinders reach higher best *ipTM* than scFvs on the majority of targets (Fig. 3). This is consistent with the easier optimization geometry of an unconstrained 60–200 aa scaffold relative to CDR design within a fixed framework — and is again an *ipTM* statement, subject to the same circularity.

3.4 Full numbers

Table 1 gives best and best-per-design *ipTM* for all 14 targets.

4 Neutral-oracle validation (Boltz-2)

Because ESMFold2 *ipTM* is Biohub's own optimization objective, we re-scored the best design per platform per target with **Boltz-2** [3], an open AlphaFold3-class co-folding model that neither

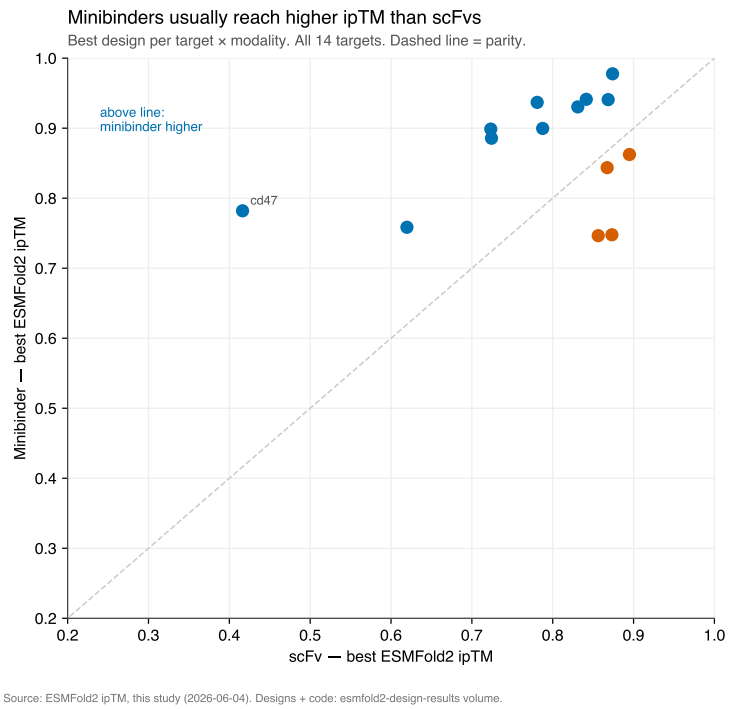

Figure 3. Best ESMFold2 *ipTM*: minibinder versus scFv per target (all 14). Points above the dashed parity line favour minibinders.

platform optimized against (RFantibody used RoseTTAFold2; Biohub used ESMFold2). We folded the 10 RFantibody lead VHHs and the best Biohub minibinder and scFv per target, each as an antigen–binder complex, single-sequence and identically. Genuine DeepMind AlphaFold3 was not used because its server cannot batch ∼200 jobs and its weights are access-gated; Boltz-2 is the runnable, equally independent substitute.

Two things change (Fig. 4, Table 2). **(1)** RFantibody designs recover sharply: median Boltz-2 *ipTM* rises from 0.13 (ESMFold2) to 0.37 — ESMFold2 was unusually harsh on them. **(2)** Biohub designs fall: minibinders 0.85 → 0.74, scFvs 0.85 → 0.69, directly quantifying home-oracle inflation. The Biohub advantage roughly halves under a neutral judge. It does not, however, vanish: the best Biohub design still out-scores the RFantibody lead on **10/10** shared targets under Boltz-2, though by smaller and more variable margins (e.g. on EGFRvIII the RFantibody VHH itself reaches 0.79). This indicates a real signal beyond pure circularity — but it must be read against a **sampling asymmetry**: the Biohub number is the best of 8 designs per modality, the RFantibody number a single lead. A symmetric best-of-N comparison, and ultimately experimental binding, remain the gold standard.

5 Engineering and cost notes

Running the protocol reproducibly required two non-obvious fixes that we document for others. **(1) Memory and context poisoning.** The ESMC-6B + ESMFold2 critic ensemble occupies ∼50–75 GB; at batch size 8 even an H200 (141 GB) ran out of memory, and a CUDA OOM on a *warm, shared* container corrupted its CUDA context, causing cascading cuDNN/CUB failures on subsequent inputs. Batch size 2 (with automatic halving on failure and retry on context-corruption errors) eliminated all failures. **(2) Cost.** On real Modal pricing (H200 $4.54/h, B200 $6.25/h), the dominant cost lever is amortizing the one-time model load and sharing the ∼150-step optimization across a batch — not the GPU tier. A B200 would need >1.38× the

Table 1. ESMFold2 *ipTM* by target. "best" = top single-critic score; "best$_D$" = top per-design score (mean of 4 critics). RFantibody column is the single lead VHH; "—" = no RFantibody design (the 4 extra targets).

	RFab VHH	Biohub minibinder		Biohub scFv	
Target	*ipTM*	best	best$_D$	best	best$_D$
b7h3	0.378	0.746	0.527	0.856	0.704
cd47	0.097	0.782	0.701	0.416	0.311
ceacam5	0.078	0.886	0.803	0.724	0.542
egfr_cetuximab	0.096	0.748	0.628	0.873	0.858
egfrviii	0.130	0.941	0.938	0.842	0.681
epha2	0.145	0.758	0.714	0.620	0.373
gpc2	0.119	0.844	0.798	0.867	0.785
her2_domIV	0.129	0.862	0.688	0.895	0.875
msln_nterm	0.186	0.899	0.897	0.723	0.628
msln_cterm*	0.498	0.978	0.971	0.874	0.720
kras (G12D)	—	0.941	0.927	0.869	0.673
tnfsf9 (4-1BBL)	—	0.930	0.908	0.831	0.730
hdac8	—	0.937	0.935	0.781	0.610
phd2 (egln1)	—	0.900	0.878	0.788	0.610

*15-residue peptide antigen; *ipTM* unreliable.

H200 throughput merely to break even, which this small-batch, latency-bound optimization does not deliver; we found B200 to be roughly cost-neutral. We therefore recommend H200 with the largest stable batch and warm containers.

6 Discussion and limitations

The headline number — Biohub designs beating RFantibody on 10/10 targets by large *ipTM* margins — is real but, on its own, scientifically uninformative about binding. The decisive limitation is **evaluation circularity**: ESMFold2 *ipTM* is the Biohub objective, so high scores are guaranteed by construction, exactly as RF2 confidence was guaranteed-high for RFantibody. The two *ipTM* columns also come from different ESMFold2 variants (hosted `esmfold2-fast` versus the experimental critics), so they are not strictly commensurate. There is **no experimental binding data** for any of these designs; *ipTM* is a confidence proxy, not affinity. The comparison is also format-asymmetric (RFantibody VHH versus minibinder/scFv) and includes one degenerate 15-residue antigen.

Our neutral-oracle pass (§4) directly addresses the circularity: scored by Boltz-2, which neither platform optimized against, the Biohub advantage roughly halves and RFantibody recovers substantially, yet Biohub's best design still leads on all 10 shared targets. The honest reading is therefore intermediate — *not* the crushing ESMFold2 margin, but *not* pure artefact either: a real but modest edge, confounded by a best-of-8 versus best-of-1 sampling asymmetry and the absence of wet-lab data. The remaining gold-standard experiments are a symmetric best-of-N comparison and, ultimately, experimental binding (SPR/BLI); genuine AlphaFold3 scoring (gated weights) would further corroborate the Boltz-2 result.

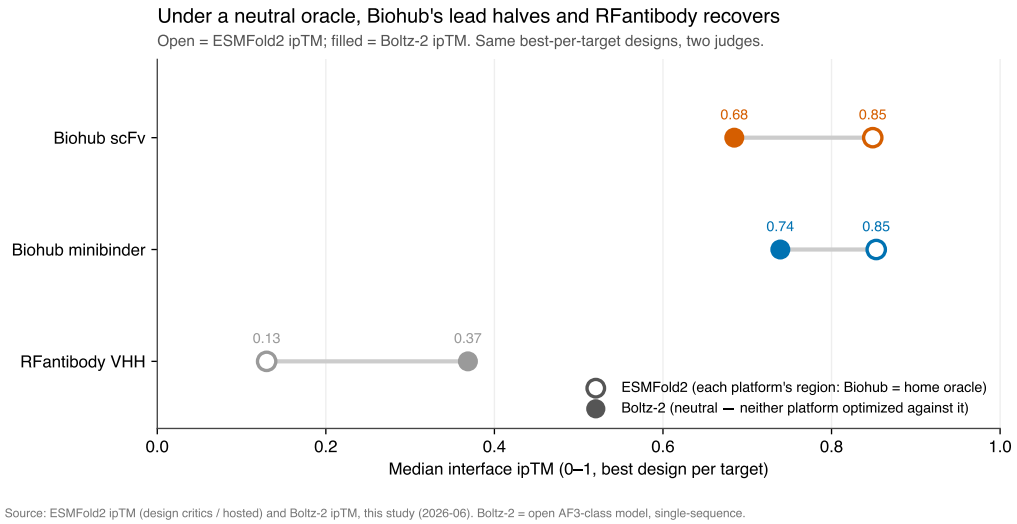

Figure 4. Median best-per-target interface *ipTM* under each platform's design-region oracle (ESMFold2, open markers) versus the neutral Boltz-2 oracle (filled). RFantibody rises; both Biohub modalities fall.

Table 2. Neutral Boltz-2 *ipTM* (single-sequence) for the best design per platform, 10 shared targets. "Biohub best" = max of minibinder and scFv.

Target	RFab VHH	Biohub MB	Biohub scFv	Biohub best
b7h3	0.496	0.793	0.601	0.793
cd47	0.152	0.453	0.507	0.507
ceacam5	0.253	0.320	0.472	0.472
egfr_cetuximab	0.226	0.270	0.782	0.782
egfrviii	0.788	0.894	0.668	0.894
epha2	0.273	0.736	0.399	0.736
gpc2	0.175	0.785	0.861	0.861
her2_domIV	0.464	0.742	0.835	0.835
msln_nterm	0.532	0.907	0.799	0.907
msln_cterm*	0.682	0.904	0.922	0.922
median	0.369	0.739	0.685	—

*15-residue peptide antigen; *ipTM* unreliable.

7 Data and code availability

All Biohub designs (sequences, per-critic metrics, loss trajectories, and predicted mmCIF complexes) are on the Modal volume `esmfold2-design-results`; neutral Boltz-2 confidence scores and structures are on `boltz-neutral-eval`. Driver and figure code: `scripts/biohub_design/` and `scripts/boltz_eval/`. Aggregated tables and this preprint's figures: `results/biohub_design/` and `out/figures/`. We invite readers to identify errors; we will correct them.

References

[1] Bennett, N., Watson, J. et al. Atomically accurate de novo design of antibodies with RFdiffusion. *Nature* (2025/2026).

[2] Candido, S., Hayes, T. et al. Language Modeling Materializes a World Model of Protein Biology. Biohub (2026).

[3] Passaro, S., Corso, G. et al. Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. (2025).