A ONE-SHOT FRAMEWORK FOR DIRECTED EVOLUTION OF ANTIBODIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Improving antibody binding to an antigen without antibody-antigen structure information or antigen-specific data remains a critical challenge in therapeutic protein design. In this work, we propose **AFFINITYENHANCER**, a framework to improve the affinity of an antibody in a one-shot setting. In the *one-shot* setting, we start from a single lead sequence—never fine-tuning on it or using its structure in complex with the antigen or epitope/paratope information—and seek variants that reliably boost affinity. During training, AFFINITYENHANCER utilizes pairs of related sequences with higher versus lower measured binding in a pan-antigen dataset comprising diverse "environments" (antigens) and a shared structure-aware module that learns to transform low-affinity sequences into high-affinity ones, effectively distilling consistent, causal features that drive binding. By incorporating pretrained sequence-structure embeddings and a sequence decoder, our method enables robust generalization to entirely new antibody seeds. Across multiple unseen internal and public seeds, AFFINITYENHANCER identifies key affinity enhancing mutations on the paratope, outperforms existing structure-conditioned and inpainting approaches, achieving substantial (in silico) affinity gains in true, one-shot experiments without ever seeing antigen data.

1 INTRODUCTION

Antibodies are proteins produced by the immune system in response to foreign antigens. In therapeutic settings, antibodies have been developed as drugs against various cancer and autoimmune targets. Antibodies detect harmful antigens (such as bacteria and viruses) by the mechanism of *binding*, attaching to a specific patch on the antigen's surface, called an *epitope*, using six hypervariable loops known as complementarity-determining regions (CDRs). A subset of the residues on these CDRs form the antigen binding surface is known as the *paratope*.

This ability to form highly specific paratopes which are complementary in shape and chemical composition to a extensive repertoire of antigens confers antibodies their unique therapeutic potential, making high-affinity antibodies prime drug candidates. Having the therapeutic potential being driven by the binding mechanism, renders structure information as essential in developing solutions for this tasks. In the typical drug discovery pipeline, a lead antibody with reasonably high affinity and specificity to the antigen of interest, is identified from immunized libraries extracted from animals, followed by optimizing the lead for potency and drug-like properties. Optimizing the potency of the lead routinely involves improving its binding or affinity to the antigen. This is called *affinity maturation*. Experimentally, affinity maturation involves random or directed mutagenesis to generate large diversified libraries (known as diversification or hit-expansion) followed by screening for stronger binding antibodies against the target. Such techniques are common in drug discovery pipelines and have been fairly successful over the last few decades. However, such diversification explores only a miniscule sequence space ($\sim$ order of $\sim 10^6$-10^9) of the entire sequence space (order of 250^{20}; 20 amino acid residues at every position of the variable domain which consists roughly of 250 residues). As a consequence, the resulting sets of designs can be suboptimal and fail to identify sufficient number of antibodies with the desired potency and drug-like properties.

Computational affinity maturation, powered by machine learning models (ML), offers an accelerated alternative to random or directed mutagenesis. However, affinity maturation with ML models becomes

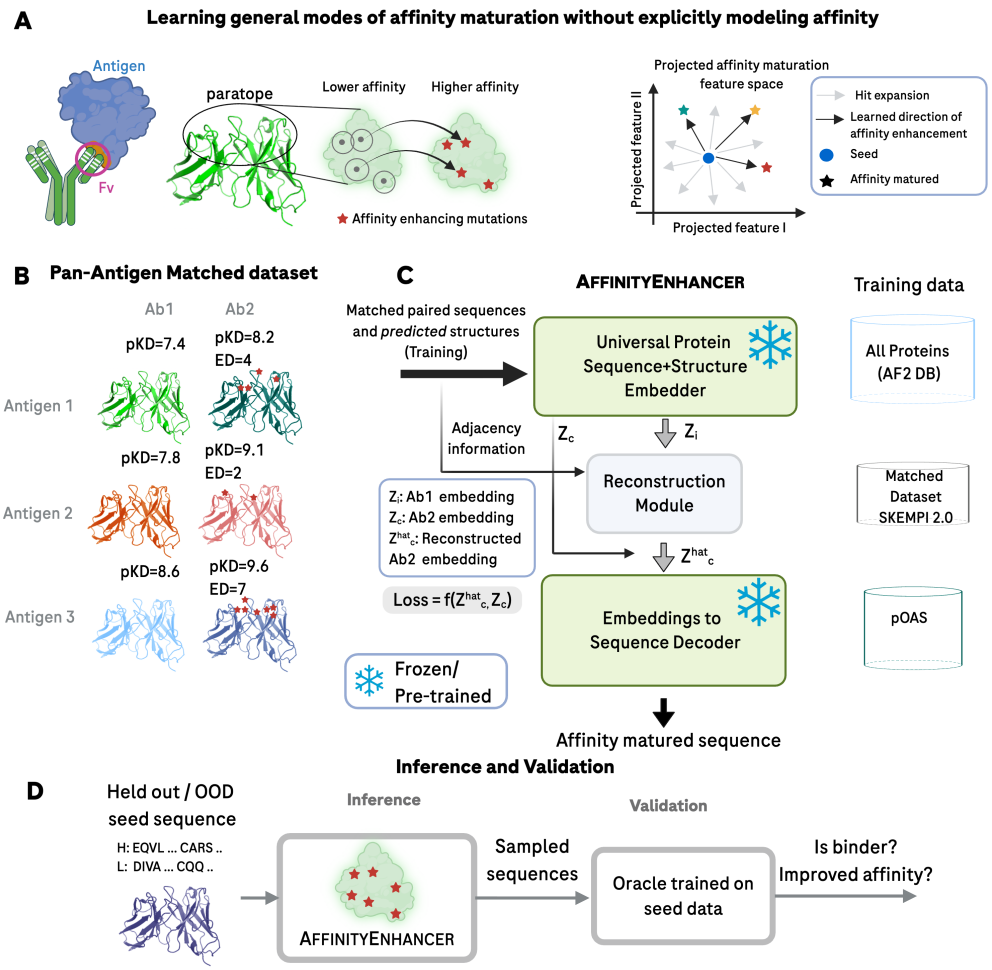

Figure 1: One-shot affinity maturation of antibodies with AFFINITYENHANCER. A) The goal is to implicitly learn modes of affinity maturation by pairing a lower affinity antibody with a higher affinity one. B) Matched datasets are obtained by pairing antibodies against the same target/antigen from the SKEMPI 2.0 database. C) Architecture for AFFINITYENHANCER. D) Inference and validation pipeline for held-out-seed to determine whether sampled sequences are binders or not.

challenging in the one-shot scenario where the lead antibody is far away from the training data, especially in sequence representation. We call this problem *the one-shot affinity maturation*, where the ML model must infer relevant (often, structure-related) modes of affinity enhancement from a single example at inference time. While several ML models have been proposed for both protein and antibody-design, very few are explicitly trained for the objective of improving binding to an antigen in the one-shot scenario. This problem is compounded by the sparsity of antibody-antigen structure and affinity datasets thereby impeding generalization to unseen cases (Hummer et al., 2023).

To bypass the challenges associated with explicitly modeling affinity, Tagasovska et al. (2024) proposed Property Enhancer (PropEn), a property-agnostic model which utilizes data matching to implicitly learn the direction of the gradient for a property of interest with the goal of proposing new optimized designs. It was previously demonstrated that this approach works for a range of tasks, including affinity maturation of antibodies. However, its effectiveness was only demonstrated (i) in sequence-based models and (ii) in cases where a few hundred antibody sequences related to the lead molecule we wish to optimize are available in the training data. In this work, we propose AFFINITYENHANCER, a model that goes beyond the PropEn framework to the one-shot affinity

maturation setup by leveraging structure information and introducing a novel, diversified matching procedure which allows for generalization and transferability. Our main contributions are as follows:

- We propose a one-shot model for affinity maturation *without antigen information* (section 3)
- We leverage matching in heterogeneous datasets to bolster data-sparse regimes (antibody-antigen interactions)
- We provide theoretical analysis supporting OOD transfer (subsection 3.1)
- In empirical results on held-out datasets, we confirm that AFFINITYENHANCER outperforms SOTA structure-conditioned and inverse-folding baselines, producing variants that improve lead-antibody binding (section 5)."

2 BACKGROUND & RELATED WORK

Structure-based design. Most ML models targeted at antibody design, including the design of target-specific antibody libraries rely on structure-conditioned sequence generation, templated on the structure of the lead antibody or, when available, the structure of the antibody-antigen complex (Dreyer et al., 2023; Mahajan et al., 2022; **?**). Such structure-conditioning is necessary in order to restrain the designed sequences to adhere to the shape of the lead antibody. The sequence space can be further controlled when the structure of the antibody-antigen complex is known. For antibody design, in particular, structure-conditioning models such as AbMPNN (Dreyer et al., 2023), AntiFold (Høie et al., 2024), FvHallucinator (Mahajan et al., 2022) and MaskedProtEnT Mahajan et al. (2025) have demonstrated impressive performance on in silico benchmarks. On the other hand, de novo models such as RFDiffusion (Watson et al., 2023), follow a two-step process. First, they design the backbone of the antibody given the context of the antigen, then follow by sequence design with ProteinMPNN conditioned on that backbone in complex with the antigen.

Sequence-based design. Alternatively, sequence-only models have been proposed to generate protein or antibody sequences from a learned distribution or near the seed. Examples of such models include discrete Walk Jump Sampler Frey et al. (2023), latent Walk Jump Sampler, ProGen2(Nijkamp et al., 2022), as well as language-model– and latent-space–guided directed evolution methods such as Hie et al. (2024); Tran & Hy (2024) and Tran et al. (2025). The latter demonstrate that large protein language models or latent generative models can effectively prioritize mutations during iterative directed evolution campaigns, improving protein function given repeated rounds of target-specific screening. However, there are no approaches addressing affinity enhancement in a one-shot setting, and in the absence of the antibody-antigen complex structure. Even de novo models such as RFDiffusion only guarantee binders (not improved binders) given a binding partner or antigen.

Training with matched datasets. We adopt a *matching-based supervision scheme* in which training pairs are formed by selecting, for each anchor the nearest neighbor such that (i) it lies within an input-space radius (ii) achieves a strictly higher measured affinity. This construction follows the spirit of PropEn which demonstrated that matching, implicitly recovers the ascent directions for a property of interest. Here we extend the matching to the one-shot antibody setting by including structure-aware embeddings and explicit environment control. In other words, to the PropEn requirements for matching, we add: (iii) pairing antibodies targeting the same antigen, i.e. *same environment*. Unlike PropEn, which uses sequence representation only, the AFFINITYENHANCERs matching operates in a geometry induced by pretrained encoders and a residual graph transformer to map low-affinity embeddings to higher-affinity counterparts.

Conceptually, this pairing induces *pairwise preferences* ($x' \succ x$), connecting our approach to preference learning (Zhang & Ranganath, 2025) methods such as Direct Preference Optimization (DPO) (Rafailov et al., 2023) from the LLM literature, where models are updated toward preference winners under KL regularization. Preference Learning has recently inspired a new direction in protein design. For backbone generation, Huguet et al. (2024) introduce Reinforced Fine-Tuning (ReFT): a supervised fine-tuning pass on a dataset filtered by auxiliary rewards to create a *preferential* subset, effectively supervised fine tuning on matched positives. In antibody co-design, Zhou et al. (2024) learn over *paired* samples by defining residue-level energy preferences and optimize a conditional diffusion model with a direct preference objective showing gains via energy decomposition and gradient-surgery to resolve conflicts. For peptide/protein binder design, (Mistani & Mysore, 2024) explicitly formulate multi-objective alignment with DPO on curated chosen/rejected receptor-binder

pairs, demonstrating that preference learning on matched datasets steers a protein LM toward binders satisfying specificity and developability (e.g., pI) constraints.

Despite the common points, two major differences should be noted. First in preference learning the sampled data consist of pairs going from lower to higher property, without any limitation on the closeness of the datapoints or their measured values. Second, preference learning focuses on taking an existing generator that inputs receptors and outputs binders and improving that generator so the outputted binders have a higher score given a receptor. In contrast, AFFINITYENHANCER seeks to produce an improved binder given an existing (lead) binder.

3 METHOD - AFFINITY ENHANCER

We formalize AFFINITYENHANCER as learning from matched improvements under fixed environments. In what follows, we state the data-generative model from which training pairs are drawn. Then, we derive the constraints that make the signal dominantly causal.

Problem setup & method summary. Let $\mathcal{X}$ denote the space of antibody sequences and let $\mathcal{Y} \subset \mathbb{R}$ denote measured binding affinities. We assume access to E *environments* indexed by $e = 1, \ldots, E$, where each environment corresponds to a distinct lead antibody (we use the terms leads or "seeds" interchangeably). In environment e we observe a small subset of sequences (order of 10) with measured affinities $\{(x_j^e, y_j^e)\}$. Our goal is, for a held-out environment e^* (the "one-shot" seed corresponding to an antigen not seen in the training set), to propose a set of new designs that reliably improve on the lead affinity $y_{\text{lead}}^{e^*}$, despite never fine-tuning on e^* or using its antigen structure. To do so, we propose AFFINITYENHANCER, summarized in Figure 1:

1. **Form matched pairs.** $\mathcal{M} = (x_i, x_i' | e = e')$ in every environment e by finding, for each low-affinity sequence x_i, the nearest neighbors x_i whose measured affinity is $y_i' > y_i$, under a capped distance threshold δ_x.
2. **Extract embeddings.** For each antibody in the matched pairs, extract sequence-structure embeddings form a foundational model $\psi : \mathcal{X} \to \mathbb{R}^{L \times d}$.
3. **Learn a worse embedding→ better embedding map.** Given matched embeddings, learn a *Graph Transformer* G_θ acting per residue and used in residual form
$$f(z) \;:=\; z + G_\theta\big(z; A, P\big), \qquad z = \psi(x) \in \mathbb{R}^{L \times d},$$
where A is a residue–residue adjacency (from predicted structure) and P positional/edge features.
4. **Embeddings→sequence decoder.** Train a light-weight decoder $\rho : \mathbb{R}^{L \times d} \to \mathcal{X}$ that maps per-residue embeddings to amino-acid distributions.
5. **Sampling for an unseen lead.** At test time, compute $z_{\text{lead}} = \psi(x_{\text{lead}}^{e^*})$ and apply the residual map
$$\tilde{z} \;=\; f(z_{\text{lead}}) \;=\; z_{\text{lead}} + G_\theta\big(z_{\text{lead}}; A, P\big), \qquad \tilde{x} \;=\; \rho(\tilde{z}).$$

3.1 FROM MATCHED DATA TO CAUSAL SIGNALS

Data-generative process. We posit that x factors into latent components:

$$x = f(s, c) \quad y = h(c, e)$$

where c collects the **causal factors** that determine affinity and s collects **spurious factors** (such as batch effects, library or lead/antigen idiosyncrasies, etc.) that influence x but not y once e are fixed. c causally affects y for fixed e. In an idealized world, we would sample independently

$$c \sim q(c), \quad s \sim q(s), \quad e \sim q(e), \quad x = f(s, c), \quad y = h(c, e).$$

In practice, *selection* (Pearl, 2009) induces dependencies: only some (c, s) are assayed, and not every x is tested in every e. The observable joint is therefore summarized as

$$(c, s) \sim p(c, s), x = f(s, c), \quad e \sim p(e \mid x), y = h(c, e),$$

In particular, y depends on the target i.e. environment e, hence s and e may spuriously correlate with y through selection rather than causation.

Matched pair selection as targeted conditional. For each anchor x assayed in environment e with outcome y, we seek a nearby variant x' that improves the outcome, *in the same environment*:

$$p(x'|x, d(x, x') < \varepsilon, y' - y > \Delta y, e' = e) \tag{1}$$

with distance d on $\mathcal{X}$, a small neighborhood radius $\varepsilon > 0$, and an improvement margin $\Delta y > 0$. Conditioning on $e' = e$ removes environment-driven gains; only changes in x can explain improvements. This conditional represents the data-matching rule that defines our train set. For simplicity we include a deterministic analysis free of measurement noise.

We impose two standard smoothness assumptions which align with biophysical/representational assumptions as well.

Assumption 1 (Property smoothness)**.** *For fixed e, the property function is K_y-Lipschitz in the causal latent:*

$$|h(c_1, e) - h(c_2, e)| \leq K_y d(c_1, c_2).$$

Assumption 2 (Responsive observation/bi-Lipschitz renderer)**.** *There exists K_x such that for all $(c, s), (c', s')$,*

$$\frac{1}{K_x} d([c, s], [c', s']) \leq d(f(s, c), f(s', c')) \leq K_x d([c, s], [c', s']),$$

and the latent metric decomposes additively,

$$d([c, s], [c', s']) = d(c, c') + d(s, s').$$

Intuitively, small moves in x, imply small moves in the underlying factors; no large cancellation can hide a big change in c by counter-moving s.

Theorem 1 (Improvement Bounds)**.** *Consider a* matched *pair (x, x') measured in the same environment with*

$$d(x, x') < \varepsilon \qquad \textit{and} \qquad y' - y = h(c', e) - h(c, e) \; > \; \Delta_y \; > \; 0. \tag{2}$$

Then:

*1. (**Minimum causal movement**)*

$$d(c', c) \; > \; \Delta_y / K_y. \tag{3}$$

*2. (**Spurious-movement cap**) If, in addition, $K_x \varepsilon - \Delta_y / K_y \geq 0$, then*

$$d(s', s) \; < \; K_x \, \varepsilon \; - \; \Delta_y / K_y. \tag{4}$$

Proof. From equation 2 and A1,

$$\Delta_y \; < \; h(c', e) - h(c, e) \; \leq \; K_y \, d(c', c) \quad \Rightarrow \quad d(c', c) \; > \; \Delta_y / K_y,$$

which proves equation 3. Next, by equation 2 and A2,

$$d(c', c) + d(s', s) \; \leq \; K_x \, d\big(f(s, c), f(s', c')\big) \; \leq \; K_x \, \varepsilon.$$

Subtracting the lower bound on $d(c', c)$ from the left-hand side yields

$$d(s', s) \; < \; K_x \, \varepsilon - d(c', c) \; \leq \; K_x \, \varepsilon - \Delta_y / K_y,$$

establishing equation 4 whenever the right-hand side is nonnegative. □

The matching rule is feasible only if $K_x \varepsilon - \frac{\Delta y}{K_y} \geq 0$; otherwise no pair can simultaneously be close in x and improve y. From equation 3 and equation 4, each pair enforces a minimal step along causal directions and leaves a strictly bounded budget for spurious drift. Hence, the supervision from matched improvements is dominated by *causal variation*.

Training AFFINITYENHANCER Let $z = \psi(x)$ be a sequence-structure embeddings (frozen). The embedding-to-embedding module learns a residual map $f_\theta(z) = z + G_\theta(z; A, P)$, trained to reconstruct matched targets in embedding space, by minimizing

$$\mathcal{L}(\theta) = \frac{1}{|M|}\Sigma_{(x,x')\in\mathcal{M}}||\psi(x') - f_\theta(\psi(x))||_2^2.$$

At test time, for a held-out seed x_{lead} in unseen environment e^* we compute $z_{lead} = \psi(x_{lead})$, apply the residual map $\tilde{z} = f(z_{\text{lead}})$, and decode $x^* = \rho(\tilde{z})$.

Why this objective isolates causal signals? By equation 1 and equation 2, each training pair constrains the model with a guaranteed minimum shift in the causal coordinates and a tight upper bound on spurious motion. Averaged over many environments, spurious directions fluctuate and cancel, while causal directions align across pairs; minimizing $\mathcal{L}$ therefore compels G_θ to model the shared environment-invariant components that consistently explain affinity gains.

Given the selection rule equation 1 and the assumptions, every matched pair obeys

$$d(c', c) > \Delta y/K_y \quad and \quad d(s', s) < k_x\varepsilon - \Delta y/K_y,$$

so the training signal is *necessarily* a causal movement plus a bounded spurious residue. AFFINITYENHANCER exploits this by learning a residual embedding-space operator that reconstructs matched targets and, at inference steps in the same causal direction on held-out seeds. This "invariance-by-matching" view will underlie all experiments that follow.

4 AFFINITYENHANCER IMPLEMENTATION

Our theoretical formulation proposed above lends a direct implementation in our AFFINITYENHANCER which consists of three main modules (Figure 1A). The structure and sequence embedder (Embedder), the reconstruction module and the embeddings to sequence decoder (Decoder) module. The Embedder embeds the antibody sequence and structure to a semantically meaningful embedded space. To this end, we utilize GearNet Zhang et al. (2023), a representation learning model trained on 600k sequences and structures from the AlphaFold2 database. To map the embeddings to antibody sequence, we trained a sequence decoder which maps GearNet (frozen) embeddings to antibody sequences on the paired Observed Antibody Space (pOAS), (Olsen et al., 2022). Once the sequence decoder is trained, it is also frozen. The reconstruction module, a Graph Transformer (GT), learns to reconstruct the embedding of the lower affinity antibody to the embedding of the higher affinity antibody. The reconstruction module is trained on the matched datasets prepared from SKEMPI 2.0. These modules allow us to embed sequences to a general embedding space that is trained on a massively large database of protein and antibody sequence and residue environments. Utilizing these pretrained modules allows us to leverage learned representations from all proteins and antibodies and generalize to blind or unseen test seeds.

5 EXPERIMENTS

The main challenge we address is whether it is possible to propose sequences of affinity enhanced designs starting from a single lead antibody sequence without *any* context or structure related to the antigen. Our validation pipeline is included in Figure 1B. We train AFFINITYENHANCER on a matched dataset that excludes any sequences in the vicinity of held-out seeds. Additionally, we utilize a predicitve model, Coretx, (Gruver et al., 2023) (Appendix E) trained and validated on labeled expression and high-quality affinity data in vicinity of the held-out seeds. We then propose designs with AFFINITYENHANCER and predict the binding and affinity for the proposed designs with the oracle.

Metrics. We evaluate sampled designs by reporting edit distances from the seed sequence, the number of designs that are predicted to be binders, and number of improved binders over the seed. Additionally we include the binding and improved rates as well as the average performance across seeds to ease summarizing the performance per baseline.

Baselines. We compare AFFINITYENHANCER to three baselines – PropEn, trained on the same matched dataset as AFFINITYENHANCER, AntiFold, an antibody-specific, structure-conditioned

inverse folding model and IgCraft (Greenig et al., 2025), an antibody-specific generative inpainting model.

Ablations. We systematically explore the effect of each component in AFFINITYENHANCER, dataset matching, embeddings, adjacency information and model architecture (local sequential kernels - convolutional neural networks, versus adjacency-informed graph transformers).

5.1 RESULTS

We demonstrate the application of AFFINITYENHANCER on four held-out seeds – 3 internal seeds and one public (Trastuzumab) antibody, all of them with edit distance between 64 and 87 (60-70% sequence similarity [1]). The edit distances of the held-out seeds to the train set are reported in Table 7. Additionally, in Table 8 we report whether any samples in the trainset have matching germlines to the test seeds.

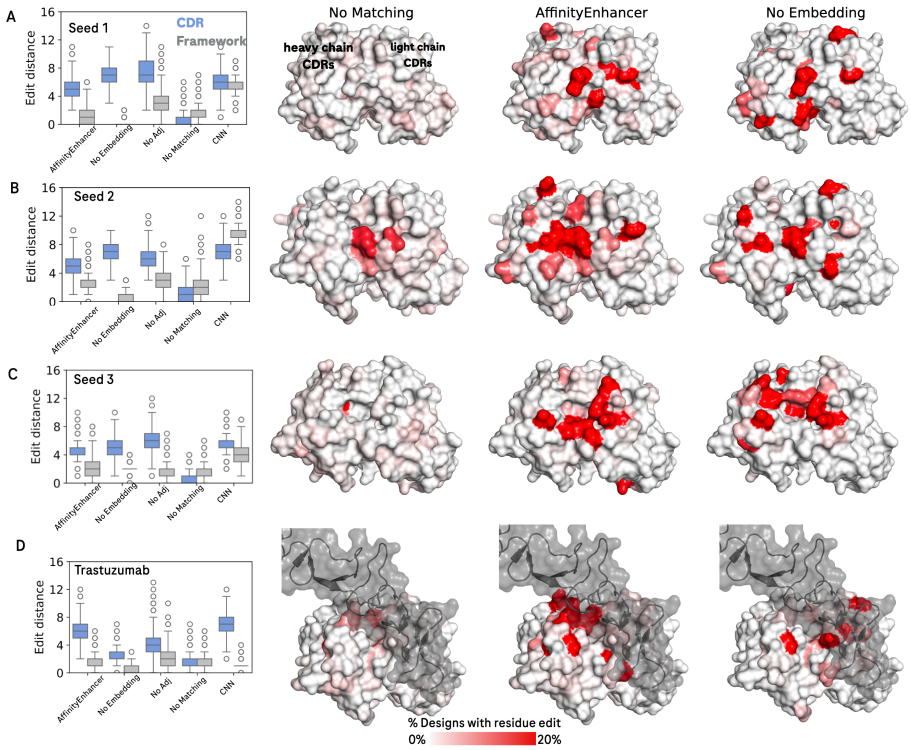

Figure 2: AFFINITYENHANCER identifies distinct and structurally important positions for each antibody. Each residue on the surface representation of the seed antibody is colored on a spectrum ranging from positions modified by the model in 0.0 (white) percent of designs to 20 percent (red) of designs. For Trastuzumab, we also show the antigen in gray. All antibody structure models were obtained with ESMFold. For Trastuzumab where the crystal structure of the antibody-antigen complex structure is available, we aligned the ESMFold structure to the crystal structure to map the position of the antigen.

AFFINITYENHANCER targets edits that retain and improve binding We asked how each model prior localizes affinity-enhancing edits across the antibody and, when known, at the antibody–antigen interface. In Figure 2A–D and Figure 5 we compare edit distance (leftmost panels) with the positions of edited residues on the binding surface (top-view CDRs). The model without matching ("No matching") serves as the baseline: it proposes few, nearly uniform edits across CDRs and frameworks, with no clear positional preferences (aside from Seed 2). In contrast, models trained with the matching intervention show distinct spatial patterns. The CNN variant makes more edits overall, spanning both

[1] For reference, antibody design experts consider edit distance of 8 to be a different molecule.

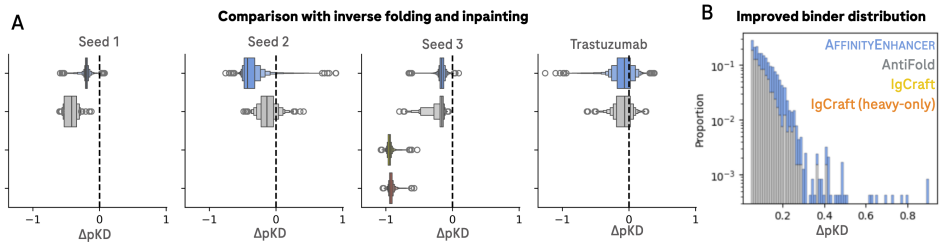

Figure 3: One-shot guided sampling with AFFINITYENHANCER. A) Comparison of AFFINITYENHANCER with the antibody-specific structure-conditioned inverse folding model, AntiFold and inpainting model IgCraft. Distribution of predicted pKD (negative log10 of dissociation constant KD for unique designs with edit distance between [5,12] for 3 internal seeds and the Trastuzumab antibody. We report difference of the predicted pKD from the pKD of the seed. IgCraft designs were sampled for all-CDRs (IgCraft) or heavy chain CDRs only (IgCraft (heavy-only)). B) Distribution of affinity improvement for AFFINITYENHANCER, AntiFold and IgCraft for designs with improved affinities (Δ pKD> 0.05).

Table 1: **Ablation study for AFFINITYENHANCER.** We sample 5,000 sequences for Trastuzumab and for three internal seeds per model. ED = minimum edit distance to the parent; "ED window" counts designs with ED $\in [5, 12]$. "Binders" and "Improved binders" are wet-lab positives and affinity-improved positives, respectively. AFFINITYENHANCER uses GearNet embeddings, pOAS decoder, an adjacency-informed Graph Transformer, and data matching. PropEn is the sequence-only baseline from Tagasovska et al. (2024). For each model, sampling settings were chosen to maximize the number of designs sampled in the ED $\in [5, 12]$ range.

Model	Set	ED	ED window	Binders	Improved	Binder rate	Improved rate
AFFINITYENHANCER	Trastuzumab	7.9 ± 1.8	4,815	**3,970**	1,575	**79.4 %**	31.50 %
	Seed 1	6.5 ± 1.6	4,382	**1,105**	2	22.1 %	0.04 %
	Seed 2	7.4 ± 1.8	4,672	**3,612**	113	72.2 %	2.26 %
	Seed 3	6.5 ± 1.7	4,352	**1,334**	3	26.7 %	0.06 %
	Mean over seeds	7.08	4,555	2,505	423	**50.10 %**	8.46 %
				Seeds improved (Trastuzumab + Seeds 1–3)			**4 /4**
PropEn (– *Structure*)	Trastuzumab	28.6	0	0	0	0.0 %	0.00 %
	Seed 1	68.0	0	0	0	0.0 %	0.00 %
	Seed 2	30.9	0	0	0	0.0 %	0.00 %
	Seed 3	68.4	0	0	0	0.0 %	0.00 %
	Mean over seeds	55.8	0	0	0	0.0 %	0.00 %
				Seeds improved (Trastuzumab + Seeds 1–3)			**0 /4**
AFFINITYENHANCER (– *Matching*)	Trastuzumab	2.8	447	392	162	7.8 %	3.24 %
	Seed 1	2.6	253	45	0	0.9 %	0.00 %
	Seed 2	3.4	954	838	**696**	16.8 %	**13.92 %**
	Seed 3	2.3	98	47	0	0.9 %	0.00 %
	Mean over seeds	2.78	438	331	215	6.61 %	4.29 %
				Seeds improved (Trastuzumab + Seeds 1–3)			**2 /4**
AFFINITYENHANCER (– *Embedding*)	Trastuzumab	3.1	161	39	14	0.8 %	0.28 %
	Seed 1	7.1	2,366	171	**134**	3.4 %	**2.68 %**
	Seed 2	7.6	3,486	**3,457**	112	69.1 %	2.24 %
	Seed 3	7.3	1,992	1,737	4	34.7 %	0.08 %
	Mean over seeds	6.27	2,001	1,351	66	27.02 %	1.32 %
				Seeds improved (Trastuzumab + Seeds 1–3)			**4 /4**
AFFINITYENHANCER (– *Graph Transformer*)	Trastuzumab	8.0	4,570	2,027	124	40.5 %	2.48 %
	Seed 1	11.2	3,812	1,085	0	21.7 %	0.00 %
	Seed 2	17.0	13	13	0	0.3 %	0.00 %
	Seed 3	9.4	4,719	89	2	1.8 %	0.04 %
	Mean over seeds	11.40	3,279	804	32	16.07 %	0.63 %
				Seeds improved (Trastuzumab + Seeds 1–3)			**2 /4**
AFFINITYENHANCER (– *Adjacency Matrix*)	Trastuzumab	6.8	4,196	3,297	**1,951**	65.9 %	**39.02 %**
	Seed 1	10.7	3,939	179	8	3.6 %	0.16 %
	Seed 2	9.1	4,769	2,873	38	57.5 %	0.76 %
	Seed 3	7.2	4,423	659	0	13.2 %	0.00 %
	Mean over seeds	8.45	4,332	1,752	499	35.04 %	**9.98 %**
				Seeds improved (Trastuzumab + Seeds 1–3)			**3 /4**

CDRs and frameworks; Graph-Transformer (GT) variants concentrate edits in CDRs; and the "No Embedding" ablation makes the fewest framework edits. Across seeds, matched models repeatedly target protruding CDR motifs. For Trastuzumab, where the interface is known, many edited positions fall in direct contact with the antigen (Figure 2D).

Outperforming inverse folding and inpainting baselines. Across all seeds, AFFINITYENHANCER shifts the predicted-affinity distribution decisively upward relative to AntiFold (Figure 3A). Whereas AntiFold—by conforming to the seed antibody's structure—mostly proposes variants that retain binding with similar or lower affinity, AFFINITYENHANCER consistently produces affinity-improving designs for nearly every seed. The inpainting sequence based model IgCraft fails to propose CDR sequences (all CDrs or heavy-only CDRs) which retain or improve binding given the context of the framework residues. This further strengthens our claim that models which learn antibody sequence distributions are insufficient to generate CDR sequences that retain binding. The magnitude of these gains also exceeds those from both AntiFold and the inpainting model IgCraft (Figure 3B).

Ablations: Which components of AFFINITYENHANCER matter and why.

Sequence-Only Baseline We first compare AFFINITYENHANCER to PropEn (sequence-only) across Seed 1–3 and Trastuzumab (*Table 1*). PropEn proposes designs more than 25 edits from the seed in every case, i.e., it fails to generate variants in the seed's neighborhood; none of its designs are predicted binders.

Across all seeds, AFFINITYENHANCER generates designs close to the seed (Tables 1), with 26–78% predicted binders and non-zero counts of improved binders for each seed. Edit distance is controllable via sampling (iterations/temperature), enabling small-to-moderate edits at low settings and larger edits at higher settings (Tables 2–5, Figure 4).

(– Matching) Autoencoder Without Guidance Removing the matching intervention reduces the model to an embedding-space autoencoder. This yields low-diversity proposals clustered near the seed and few binders or improved binders (notable exception: Seed 2). Matching is therefore critical for shifting probability mass toward functional, higher-affinity regions.

(– Embedding) Generalization Without the Embedder Without GearNet embeddings and the pOAS decoder, the model still produces some improved binders across all seeds and, for Seeds 1 and 3, the most improved binders among ablations. This suggests that structural priors plus matching capture useful causal signal even without the embedder. However, sequence diversity and binder counts—especially for Trastuzumab—lag the full model. Furthermore, edit distances are less controllable and limited to a single iteration (Tables 2–5, Figure 4).

(CNN) Weaker Structural Prior, Weaker Binders Replacing the Graph Transformer with a CNN (PropEn-style) increases edit distances and weakens edit-distance control (Tables 2–5, Figure 4). Binder and improved-binder yields drop substantially, indicating that the GT's relational bias is important for localized, functional edits.

(– Adjacency) Losing Contacts, Losing Control Using a fully connected Graph Transformer (no adjacency) similarly inflates edits and reduces controllability with sampling knobs (Tables 2–5, Figure 4). This highlights the role of explicit adjacency in guiding compact, physics-aware modifications.

Comparison to experimental data and biological insights: What is AFFINITYENHANCER (AE) able to learn and where is it still lacking.

- *AE identifies positions at the rim of the antibody-antigen interface without the knowledge of the structure of the complex:* For three seeds (Seed 1, Trastuzumab and Seed 4 - additional internal seed with known complex structure and an edit distance of 73 from the trainset Table 7), we were able to find experimentally solved crystal structures for the antibody-antigen complex. We mapped the most edited positions by AE to the antibody-antigen interface (Figure 6). Strikingly, the majority of the highest edit incidence, lies along the rim of the interface, while the core exhibits very low edit rates (Figure 6). This is a biologically meaningful pattern since more often than not, affinity enhancement, especially starting with high affinity seeds, involves augmenting the existing core-binding interface with additional affinity enhancing mutations at the rim. This is a difficult task since the antibody-antigen interface is rarely known and it's prediction is still an open problem Goudeau & Georges (2023); Polonsky et al. (2024); Svensson & others (2025). Thus, in the absence of known antibody-antigen complex structures, our implicit matching framework enables AE to infer biologically relevant positions for affinity enhancement.
- *Top 25 percentile AE mutations capture highest affinity enhancing positions for two out of three seeds* For Seeds 1, 5 and 6 (additional seeds with high edit distance from the trainset

Table 7), we were able to find high quality experimental data of single mutations to the seed. In Figure 7, we show the distribution of the maximum experimentally measured affinity improvement per position as a function of whether this position was preferentially edited by AE. For Seeds 5 and 6, the top 25 percentile of positions edited by AE includes the position that yields the maximum affinity improvement in experiments Figure 7A. For Seed 1, AE is able to identify two positions which were found to improve affinity by 0.7 pK_D in experiments but it misses two positions which lead to the largest affinity gains Figure 7A. While, for each seed, the positions with the largest affinity gains from experiments are identified by AE in the top 50 percentile of edit positions Figure 7B, it is unable to propose these positions with the highest incidence Figure 7C. This indicates an area of improvement for the current model. We speculate that greater granularity in the model, including explicit structure reconstruction and including epitope context may mitigate this issue. However, given the error in prediction in both the precise structure of the CDR loops and the antibody-antigen interface with current SOTA models and the difficulty of acquiring experimentally resolved structures of antibody-antigen complexes, the inaccuracy in capturing affinity enhancing mutations will remain challenging.

- *AE identifies semantically meaningful amino acid substitutions that enhance affinity:* For Seed 1, a noticeable number of experimentally identified affinity enhancing single-point mutations were substitutions to a negatively charged residues (D or E). We found this amino acid substitution to be prominent in the designs as well. In Figure 8A, we compare the electrostatic surfaces of the Seed, the top design with a D mutation in CDR H3 and one of the top mutations from experiments with a D substitution in CDR H3 ($\Delta \text{p}K_D$=1.3). The electrostatic surface comparison reveals that both the top experimentally identified substitution and the designed variant result in more negatively charged electrostatic surface.
- *AE fails to identify a key affinity enhancing amino acid substitution for one seed:* For Seed 5, while the most potent single-point mutations in experiments are hydrophobic (L, Y, W etc.), atleast a handful of top experimental mutations are substitutions to a positively charged residues (K). We found this amino acid substitution to be absent from designs. In Figure 8B, we compare the electrostatic surfaces of the Seed, the top design and one of the top mutations from experiments with a K substitution in CDR H3 ($\Delta \text{p}K_D$=0.65).

6 Conclusion

In this work, we tackle the one-shot task of affinity maturing a lead antibody for blind or unseen seeds. AffinityEnhancer combines dataset matching with pretrained sequence–structure representations, an antibody-specific decoder, and lightweight structural priors to propose targeted edits directly from the lead sequence. Empirically, we show it recovers binding-relevant features from sequence alone and generates affinity-enhancing mutations. Across held-out evaluations, it outperforms sequence-only PropEn, a structure-conditioned inverse-folding baseline, and a sequence-inpainting model, sampling variants with consistently higher affinity. Unlike reconstruction-driven approaches, AffinityEnhancer is designed to discover *causal*, affinity-improving mutations—yielding practical gains for directed evolution.

Beyond accuracy, AffinityEnhancer offers practical advantages for directed evolution: it generalizes to new seeds in a one-shot regime, provides controllable edit distances for risk-aware exploration, and remains data-efficient by leveraging pretrained biomolecular priors. These properties make it a useful drop-in companion for antibody lead optimization when structural complexes or large labeled datasets are unavailable. Looking ahead, AffinityEnhancer creates a clear path for further gains. Incorporating epitope or antigen context could disambiguate multiple plausible routes to improvement, while expanding labeled affinity resources will broaden coverage of binding modes. We view these as opportunities to extend a framework that already delivers strong, sequence-only affinity maturation with minimal assumptions and maximal practical impact.

7 Reproducibility Statement

We will release (i) the exact SKEMPI-derived matched pairs (IDs and thresholds), (ii) code to recompute matches from raw SKEMPI/pOAS, (iv) all hyperparameters, (v) pre-trained weights for and G_θ, and (vi) scripts to reproduce all figures/tables from a single make entrypoint. We report complete sampling settings and will upload an anonymous artifact with code and models at submission time.

References

Frédéric A Dreyer, Daniel Cutting, Constantin Schneider, Henry Kenlay, and Charlotte M Deane. Inverse folding for antibody sequence design using deep learning. *arXiv preprint arXiv:2310.19513*, 2023.

Nathan C. Frey, Daniel Berenberg, Karina Zadorozhny, Joseph Kleinhenz, Julien Lafrance-Vanasse, Isidro Hotzel, Yan Wu, Stephen Ra, Richard Bonneau, Kyunghyun Cho, Andreas Loukas, Vladimir Gligorijevic, and Saeed Saremi. Protein discovery with discrete walk-jump sampling, 2023.

Nathan C Frey, Taylor Joren, Aya Abdelsalam Ismail, Allen Goodman, Richard Bonneau, Kyunghyun Cho, and Vladimir Gligorijević. Cramming protein language model training in 24 gpu hours. *bioRxiv*, pp. 2024–05, 2024.

Nathan C Frey, Isidro Hötzel, Samuel D Stanton, Ryan Kelly, Robert G Alberstein, Emily Makowski, Karolis Martinkus, Daniel Berenberg, Jack Bevers III, Tyler Bryson, et al. Lab-in-the-loop therapeutic antibody design with deep learning. *bioRxiv*, pp. 2025–02, 2025.

Laurent E Goudeau and Guy Georges. Challenges in antibody structure prediction. *mAbs*, 15(1): 2175319, 2023. doi: 10.1080/19420862.2023.2175319.

Matthew Greenig, Haowen Zhao, Vladimir Radenkovic, Aubin Ramon, and Pietro Sormanni. Igcraft: A versatile sequence generation framework for antibody discovery and engineering. *arXiv preprint arXiv:2503.19821*, 2025.

Nate Gruver, Samuel Stanton, Nathan Frey, Tim GJ Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew G Wilson. Protein design with guided discrete diffusion. *Advances in neural information processing systems*, 36:12489–12517, 2023.

Brian L Hie, Varun R Shanker, Duo Xu, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, Wesley Wu, John E Pak, and Peter S Kim. Efficient evolution of human antibodies from general protein language models. *Nature biotechnology*, 42(2):275–283, 2024.

Magnus Haraldson Høie, Alissa Hummer, Tobias H Olsen, Broncio Aguilar-Sanjuan, Morten Nielsen, and Charlotte M Deane. AntiFold: Improved antibody structure-based design using inverse folding, 2024. URL https://arxiv.org/abs/2405.03370.

Guillaume Huguet, James Vuckovic, Kilian Fatras, Eric Thibodeau-Laufer, Pablo Lemos, Riashat Islam, Chenghao Liu, Jarrid Rector-Brooks, Tara Akhound-Sadegh, Michael Bronstein, et al. Sequence-augmented se (3)-flow matching for conditional protein generation. *Advances in neural information processing systems*, 37:33007–33036, 2024.

Alissa M. Hummer, Constantin Schneider, Lewis Chinery, and Charlotte M. Deane. Investigating the Volume and Diversity of Data Needed for Generalizable Antibody-Antigen G Prediction. *bioRxiv*, pp. 2023.05.17.541222, 2023. URL https://www.biorxiv.org/content/10.1101/2023.05.17.541222v1%0Ahttps://www.biorxiv.org/content/10.1101/2023.05.17.541222v1.abstract.

Sai Pooja Mahajan, Jeffrey A. Ruffolo, Rahel Frick, and Jeffrey J. Gray. Hallucinating structure-conditioned antibody libraries for target-specific binders. *Frontiers in Immunology*, 13, October 2022. doi: 10.3389/fimmu.2022.999034. URL https://doi.org/10.3389/fimmu.2022.999034.

Sai Pooja Mahajan, Fátima A Dávila-Hernández, Jeffrey A Ruffolo, and Jeffrey J Gray. How well do contextual protein encodings learn structure, function, and evolutionary context? *Cell Systems*, 16(3), 3 2025. ISSN 2405-4712. doi: 10.1016/j.cels.2025.101201. URL https://doi.org/10.1016/j.cels.2025.101201.

Pouria Mistani and Venkatesh Mysore. Preference optimization of protein language models as a multi-objective binder design paradigm. *arXiv preprint arXiv:2403.04187*, 2024.

Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. Progen2: Exploring the boundaries of protein language models, 2022.

Tobias H Olsen, Fergus Boyles, and Charlotte M Deane. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31 (1):141–146, 2022.

Judea Pearl. Causal inference in statistics: An overview. 2009.

Joshua R Polonsky, Tom Kelleher, Bernat Abanades, and Charlotte M Deane. Ai-augmented physics-based docking for antibody-antigen complex prediction. *Bioinformatics*, 41(4):btaf129, 2024. doi: 10.1093/bioinformatics/btae129.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Rickard Svensson and others. Improved prediction of antibody and their complexes with clustered generative modelling ensembles. *Bioinformatics Advances*, 5(1):vbaf161, 2025. doi: 10.1093/bioadv/vbaf161.

Nataša Tagasovska, Vladimir Gligorijevic, Kyunghyun Cho, and Andreas Loukas. Implicitly guided design with propen: Match your data to follow the gradient. *Advances in Neural Information Processing Systems*, 37:35973–36001, 2024.

Thanh VT Tran and Truong Son Hy. Protein design by directed evolution guided by large language models. *IEEE Transactions on Evolutionary Computation*, 2024.

Thanh VT Tran, Nhat Khang Ngo, Viet Thanh Duy Nguyen, and Truong-Son Hy. Latentde: latent-based directed evolution for protein sequence design. *Machine Learning: Science and Technology*, 6(1):015070, 2025.

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, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Lily H Zhang and Rajesh Ranganath. Preference learning made easy: Everything should be understood through win rate. In *Forty-second International Conference on Machine Learning*, 2025.

Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. *arXiv*, 2023. doi: 10.48550/arxiv.2203.06125. URL https://arxiv.org/abs/2203.06125.

Xiangxin Zhou, Dongyu Xue, Ruizhe Chen, Zaixiang Zheng, Liang Wang, and Quanquan Gu. Antigen-specific antibody design via direct energy-based preference optimization. *Advances in Neural Information Processing Systems*, 37:120861–120891, 2024.

A Appendix

B Probabilistic bounds for Theorem 1

We now allow *only* affinity measurements to be noisy, with noise *independent* of the environment. The sequence/embedding path is noise-free. Specifically,

$$z = \psi(x), \qquad y_{\text{obs}} = h(c, e) + \xi_y,$$

where ξ_y is zero-mean sub-Gaussian with proxy σ_y^2, i.i.d. across samples, and independent of (e, c, s). (The sub-Gaussian assumption yields tight, environment-agnostic high-probability margins; weaker moment assumptions are possible with looser bounds.)

Observed matching rule. We form matches using *observed* improvements in the same environment and exact embedding proximity:

$$d\big(\psi(x), \psi(x')\big) < \varepsilon, \qquad y'_{\text{obs}} - y_{\text{obs}} > \Delta_y, \qquad e = e'. \tag{5}$$

High-probability causal movement. Since $y'_{\text{obs}} - y_{\text{obs}} = \big(h(c', e) - h(c, e)\big) + (\xi'_y - \xi_y)$, sub-Gaussianity implies that for any $\delta \in (0, 1)$ there exists

$$\Gamma_y(\delta) = 2\sigma_y\sqrt{2\log(1/\delta)} \quad \text{s.t.} \quad \mathbb{P}(h(c', e) - h(c, e) > \Delta_y - \Gamma_y(\delta)) \geq 1 - \delta.$$

By A1, with the same probability,

$$d(c', c) \;>\; \frac{\Delta_y - \Gamma_y(\delta)}{K_y}. \tag{6}$$

High-probability spurious cap (no x-noise). Assume A2 and denote $z = \psi(x)$. From equation 5, $d(c', c) + d(s', s) \leq K_x\,\varepsilon$. Combining with equation 6 yields, with probability $\geq 1 - \delta$,

$$d(s', s) \;<\; K_x\,\varepsilon \;-\; \frac{\Delta_y - \Gamma_y(\delta)}{K_y}. \tag{7}$$

Feasibility and interpretation. Non-vacuous guarantees require $\Delta_y > \Gamma_y(\delta)$ and $K_x\varepsilon - (\Delta_y - \Gamma_y(\delta))/K_y \geq 0$. Because ξ_y is independent of the environment, the margin $\Gamma_y(\delta)$ is *uniform across all* e. Equations equation 6–equation 7 are the noise-robust analogues of the deterministic bounds equation 3–equation 4 when y is noisy.

C Model and Training

Dataset: The matched dataset was prepared with an edit distance threshold of 5 and a pKD threshold of 1.5. **Inputs and embeddings**: All structures are predicted with ESMFold. Per-residue GearNet embeddings concatenated over all 6 layers (dimension=3072) are obtained for the full fv (heavy and light chains), followed by padding to a fixed heavy and light chain length of 151 and 150 respectively. **Sequence decoder**: The sequence decoder is a 2-layer multi-layer perceptron with a hidden dimension of 32 and ReLU activation. The decoder is trained with the GearNet embeddings (frozen) on the paired Observed Antibody Space (pOAS) Table2. **Model**: AffinityEnhancer has 4.2M parameters 3. The GraphTransformer was adapted from lucidrains implementation on Github (https://github.com/lucidrains/graph-transformer-pytorch). It has 4 blocks. Each block consists of normalization layer, an attention layer, a gated residual connection with 4 attention heads and a hidden dimension of 256. The model was trained for 200 epochs. Training times as a function of dataset size are reported in Table 4. Seeds 1, 2 and 3 were trained on trainset of size 2200 whereas Trastuzumab was trained on trainset of size 1300 (after removing all sequences in the vicinity of the seed). At inference, it takes 3-3.5 mins to generate 5000 samples (batched inference with batch size of 64) on a single A100 or a G5 GPU.

Table 2: Parameter count for the "GearNet_MLP" autoencoder, composed of a frozen GearNet embedder and an MLP decoder (embeddings to sequence). This model required 3 hours and 30 minutes of training time on a single A100 GPU.

	Name	Type	Params	Mode
0	encoder	GearNet	20.1 M	eval
1	decoder	MLP	99.0 K	train

Table 3: Model Architecture and Training Details

	Name	Type	Params	Mode
0	autoencoder	Gearnet_MLP	20.2 M	eval
1	model	GraphTransformer	4.2 M	train

Table 4: Typical training configuration on A100 GPU and wall-clock time as a function of matched dataset size.

Matched dataset size	Training time (hours)	# of GPUs
1300	~ 2.5	1
2200	~ 5	1
5100	~ 14	1
7640	~ 48	2

D AFFINITY ORACLE

For our in silico validation, we use Cortex (Gruver et al., 2023), a multi-task fine-tuning framework that uses pre-trained Language models for Biological Sequence Transformation and Evolutionary Representation (LBSTER) (Frey et al., 2024) to simultaneously model multiple properties of interest (binding affinity, expression). Cortex has been trained on diverse set of targets, including the leads and their surrounding data included in our manuscript. This oracle has been recently suggested in an extensive lab-in-the-loop study for affinity maturation of antibodes (Frey et al., 2025).

Table 5: Overall performance for CORTEX (in-distribution affinity prediction).

Model	**Mean Binder Accuracy (%)**	**Standard Error (%)**	**Spearman ρ (pKD)**
Cortex	82.9	0.4	0.90

Table 6: CORTEX Per-target accuracies.

Seed	Accuracy (%)
Seed 1	72.4
Seed 2	62.0
Seed 3	70.0
Trastuzumab	78.4

E ANTIFOLD AND IGCRAFT

For AntiFold, for each seed, we sampled 5000 sequences at temperatures 0.2 and 0.5. For IgCraft, we sampled sequences with default parameters and for additional setting (lower sampling temperature of 0.05 and number of steps set to 10).

F ADDITIONAL RESULTS FROM EXPERIMENTS

G POSITIONING OF AFFINITYENHANCER WITH RESPECT TO SOTA METHODS.

H COMPARISON OF AFFINITYENHANCER WITH ORACLE-GUIDED LATENT MODELS

A common class of protein optimization algorithms relies on generative models guided by oracles/predictors in latent space, with directed evolution approaches such as Tran & Hy (2024) and

Table 7: Edit Distance and Sequence identity (SI) to trainset for test set seeds. Seeds 1-3 and Trastuzumab are used for affinity enhancement experiments.Internal Seeds 4-6 were used for comparison with experimental data and biological insights. For reference, a typical heavy chain is 115 residues whereas a typical light chain is 106 residues in length. A sequence identity of 90-95% (10-30 residues) is commonly used to demarcate out-of-distribution samples for antibody sequences in prior works.

Seed	SI	heavy	light	full	L1	L2	L3	H1	H2	H3
Seed1	63	42	37	84	4	2	5	5	6	8
Seed2	69	28	35	69	3	3	3	2	4	6
Seed3	71	30	18	64	4	2	4	4	5	4
Trastuzumab	72	28	33	64	6	3	5	3	5	6
Seed 4 (structure comparison)	68	44	19	73	3	3	2	6	5	5
Seed 5 (experimental comparison)	66	32	35	75	3	1	4	3	4	6
Seed 6 (experimental comparison)	78	30	22	52	10	3	4	4	5	7

Table 8: Overlap between train set and test set germlines. Seeds 1-3 and Trastuzumab are used for affinity enhancement experiments. Additional internal Seeds 4-6 are used for comparison with experimental data and biological insights.

	Matching gene in train set			
Seed	# heavy V-gene	# heavy J-gene	# light V-gene	# light J-gene
Seed 1	0	193	0	0
Seed 2	0	0	0	1785
Seed 3	0	85	0	179
Trastuzumab	0	72	0	813
Seed 4 (Structure comparison)	371	85	0	279
Seed 5 (experimental comparison)	0	98	0	279
Seed 6 (experimental comparison)	941	85	0	179

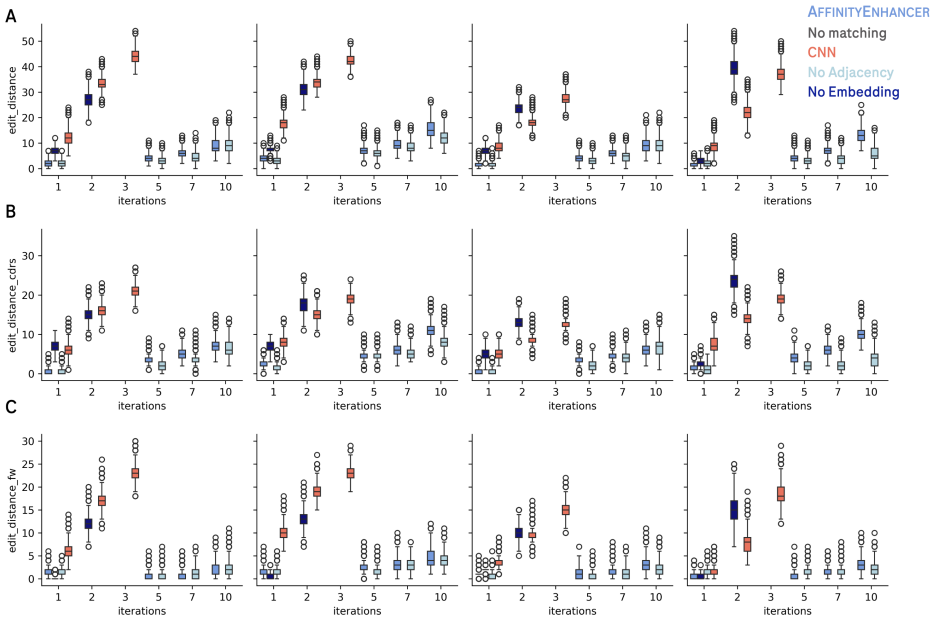

Figure 4: Edit distance distribution as a function of number of iterations at sampling for different model ablations A) Full length antibody B) CDRs only C) Framework regions only.

Table 9: Selecting sampling parameters for Trastuzumab with the maximum number of designs within an edit distance of [5,12] from the seed sequence. We sampled 5000 designs for 3 internal seeds and Trastuzumab. AFFINITYENHANCER is the base model with a GearNet embedder and pOAS sequence decoder, an adjacency-informed Graph Transformer and with data matching. PropEn is the sequence-only model from Tagasovska et al. (2024)

model	iterations	temperature	edit_distance	N edit_distance [5,12]
AFFINITYENHANCER	1	0.7	1.1 ± 0.3	0
	1	1.0	2.0 ± 0.9	20
	5	0.7	3.4 ± 0.9	118
	5	1.0	4.8 ± 1.5	2544
	7	0.7	5.7 ± 1.1	2427
	7	1.0	7.9 ± 1.8	4815
	10	0.7	11.3 ± 1.3	3543
	10	1.0	14.6 ± 2.2	805
AFFINITYENHANCER (No Matching)	1	0.7	1.3 ± 0.5	0
	1	1.0	2.8 ± 1.3	393
	5	0.7	1.4 ± 0.6	0
	5	1.0	2.7 ± 1.3	361
	7	0.7	1.4 ± 0.6	0
	7	1.0	2.8 ± 1.3	407
	10	0.7	1.4 ± 0.6	0
	10	1.0	2.8 ± 1.3	447
AFFINITYENHANCER (No Embed)	1	0.7	2.5 ± 0.9	2
	1	1.0	3.1 ± 1.1	161
	2	0.7	37.2 ± 2.9	0
	2	1.0	41.0 ± 3.5	0
AFFINITYENHANCER (CNN)	1	0.5	7.3 ± 1.3	2319
	1	0.7	8.0 ± 1.5	4570
	1	1.0	10.8 ± 2.1	3988
	2	0.5	20.0 ± 1.7	0
	2	0.7	21.4 ± 2.1	0
	2	1.0	25.4 ± 2.7	0
	3	0.5	35.3 ± 1.9	0
	3	0.7	36.7 ± 2.3	0
	3	1.0	40.1 ± 2.8	0
AFFINITYENHANCER (No Adj)	1	0.7	1.2 ± 0.4	0
	1	1.0	2.5 ± 1.2	266
	5	0.7	1.6 ± 0.7	0
	5	1.0	3.5 ± 1.6	1049
	7	0.7	2.1 ± 0.8	6
	7	1.0	4.5 ± 1.9	2269
	10	0.7	3.2 ± 1.0	238
	10	1.0	6.8 ± 2.2	4196

Tran et al. (2025) being representative examples. To compare AffinityEnhancer to this framework, we focus on Tran & Hy (2024) and use the authors' implementation as shared in the public repository. To adapt the method for binding affinity, we first train the ESM2-based decoder on the SKEMPI v2 dataset (after removing sequences close to Seed 1 and Trastuzumab for the one-shot setup). We use the default parameters dec hidden dim = 1280, batch size=256, lr=5e-5 , and num epochs=50. The validation MAE for the decoder is 0.451 for Seed 1 and 1.167 for Trastuzumab. We then sample new sequences with n steps=10, population=5000, num proposes per var=4 , population ratio per mask=0.6. We choose a lower number of steps than the default (60), since this yields designs closer to the seed, for which we have greater confidence in the oracle predictions.

We were unable to generate a sufficiently large number of designs for MLDETran & Hy (2024) with reasonable edit distances from the seed. For example, default settings yielded edit distances of >30 edits in the seed (Seed 1). We selected a lower number of iterations to obtain MLDE designs in the vicinity of the seed. To match this setting, we also generated AffinityEnhancer designs with lower temperatures and iterations. We also compared predicted affinity values between MLDE and AE and found later to give better affinities. AE also primarily identified edits in the CDRs versus MLDE

Table 10: Selecting sampling parameters for Seed 1 with the maximum number of designs within an edit distance of [5,12] from the seed sequence. We sampled 5000 designs for 3 internal seeds and Trastuzumab. AFFINITYENHANCER is the base model with a GearNet embedder and pOAS sequence decoder, an adjacency-informed Graph Transformer and with data matching. PropEn is the sequence-only model from Tagasovska et al. (2024)

model	iterations	temperature	edit_distance	N edit_distance [5,12]
AFFINITYENHANCER	1	0.7	1.1 ± 0.3	0
	1	1.0	2.2 ± 1.0	83
	5	0.7	3.3 ± 0.7	33
	5	1.0	4.6 ± 1.3	1985
	7	0.7	4.8 ± 0.9	1148
	7	1.0	6.5 ± 1.6	4382
	10	0.7	6.9 ± 1.2	4178
	10	1.0	10.4 ± 2.1	4182
AFFINITYENHANCER (No Matching)	1	0.7	1.2 ± 0.4	0
	1	1.0	2.3 ± 1.1	123
	5	0.7	1.1 ± 0.4	0
	5	1.0	2.4 ± 1.1	154
	7	0.7	1.2 ± 0.5	1
	7	1.0	2.4 ± 1.2	210
	10	0.7	1.2 ± 0.5	0
	10	1.0	2.6 ± 1.2	253
AFFINITYENHANCER (No Embed)	1	0.7	6.7 ± 1.1	703
	1	1.0	7.1 ± 1.2	2366
	2	0.7	26.2 ± 2.1	0
	2	1.0	28.4 ± 2.6	0
AFFINITYENHANCER (CNN)	1	0.5	10.1 ± 1.5	3645
	1	0.7	11.2 ± 1.8	3812
	1	1.0	14.7 ± 2.3	871
	2	0.5	32.6 ± 1.7	0
	2	0.7	32.9 ± 2.0	0
	2	1.0	34.8 ± 2.5	0
	3	0.5	42.8 ± 1.6	0
	3	0.7	43.6 ± 1.9	0
	3	1.0	45.9 ± 2.4	0
AFFINITYENHANCER (No Adj)	1	0.7	1.1 ± 0.4	0
	1	1.0	2.2 ± 1.1	109
	5	0.7	2.0 ± 0.7	0
	5	1.0	3.4 ± 1.3	754
	7	0.7	3.1 ± 0.9	76
	7	1.0	5.2 ± 1.8	2960
	10	0.7	6.7 ± 1.4	3848
	10	1.0	10.7 ± 2.3	3939

which made edits in both CDRs and frameworks regions to the same extent. In all settings tested, AffinityEnhancer outperforms MLDE. (Table 14).

Table 11: Selecting sampling parameters for Seed 2 with the maximum number of designs within an edit distance of [5,12] from the seed sequence. We sampled 5000 designs for 3 internal seeds and Trastuzumab. AFFINITYENHANCER is the base model with a GearNet embedder and pOAS sequence decoder, an adjacency-informed Graph Transformer and with data matching. PropEn is the sequence-only model from Tagasovska et al. (2024)

model	iterations	temperature	edit_distance	N edit_distance [5,12]
AFFINITYENHANCER	1	0.7	2.7 ± 0.7	3
	1	1.0	3.9 ± 1.3	1194
	5	0.7	5.4 ± 1.0	1293
	5	1.0	7.4 ± 1.8	4672
	7	0.7	7.5 ± 1.2	3515
	7	1.0	10.6 ± 2.1	4130
	10	0.7	13.3 ± 1.7	1619
	10	1.0	17.6 ± 2.4	48
AFFINITYENHANCER (No Matching)	1	0.7	1.4 ± 0.6	0
	1	1.0	2.8 ± 1.3	443
	5	0.7	1.5 ± 0.6	0
	5	1.0	3.0 ± 1.4	557
	7	0.7	1.6 ± 0.6	0
	7	1.0	3.1 ± 1.4	672
	10	0.7	1.8 ± 0.7	1
	10	1.0	3.4 ± 1.5	954
AFFINITYENHANCER (No Embed)	1	0.7	7.3 ± 1.2	1470
	1	1.0	7.6 ± 1.3	3486
	2	0.7	30.1 ± 2.1	0
	2	1.0	32.1 ± 2.6	0
AFFINITYENHANCER (CNN)	1	0.5	16.3 ± 1.5	10
	1	0.7	17.0 ± 1.7	13
	1	1.0	19.5 ± 2.2	2
	2	0.5	32.3 ± 1.3	0
	2	0.7	33.4 ± 1.7	0
	2	1.0	35.6 ± 2.1	0
	3	0.5	41.6 ± 1.6	0
	3	0.7	42.0 ± 1.8	0
	3	1.0	43.1 ± 1.9	0
AFFINITYENHANCER (No Adj)	1	0.7	1.8 ± 0.5	0
	1	1.0	3.0 ± 1.2	440
	5	0.7	5.0 ± 0.9	840
	5	1.0	6.6 ± 1.6	4353
	7	0.7	7.0 ± 1.1	2089
	7	1.0	9.1 ± 1.8	4769
	10	0.7	10.1 ± 1.3	3918
	10	1.0	13.6 ± 2.2	1547

Table 12: Selecting sampling parameters for Seed 3 with the maximum number of designs within an edit distance of [5,12] from the seed sequence. We sampled 5000 designs for 3 internal seeds and Trastuzumab. AFFINITYENHANCER is the base model with a GearNet embedder and pOAS sequence decoder, an adjacency-informed Graph Transformer and with data matching. PropEn is the sequence-only model from Tagasovska et al. (2024)

model	iterations	temperature	edit_distance	N edit_distance [5,12]
AFFINITYENHANCER	1	0.7	1.1 ± 0.3	0
	1	1.0	1.9 ± 0.9	19
	5	0.7	3.3 ± 0.9	78
	5	1.0	4.6 ± 1.4	2106
	7	0.7	4.7 ± 1.0	1021
	7	1.0	6.5 ± 1.7	4352
	10	0.7	7.2 ± 1.4	4066
	10	1.0	10.6 ± 2.2	4062
AFFINITYENHANCER (No Matching)	1	0.7	1.1 ± 0.3	0
	1	1.0	1.9 ± 0.9	39
	5	0.7	1.2 ± 0.4	0
	5	1.0	2.1 ± 0.9	37
	7	0.7	1.2 ± 0.4	0
	7	1.0	2.1 ± 1.0	54
	10	0.7	1.3 ± 0.5	0
	10	1.0	2.3 ± 1.0	98
AFFINITYENHANCER (No Embed)	1	0.7	6.9 ± 1.1	434
	1	1.0	7.3 ± 1.3	1992
	2	0.7	23.7 ± 2.1	0
AFFINITYENHANCER (CNN)	1	0.5	7.1 ± 1.0	834
	1	0.7	7.6 ± 1.2	2933
	1	1.0	9.4 ± 1.8	4719
	2	0.5	16.7 ± 1.4	3
	2	0.7	17.4 ± 1.7	6
	2	1.0	19.7 ± 2.1	1
	3	0.5	26.3 ± 1.5	0
	3	0.7	26.9 ± 1.8	0
	3	1.0	29.0 ± 2.3	0
AFFINITYENHANCER (No Adj)	1	0.7	1.1 ± 0.4	0
	1	1.0	2.0 ± 0.9	31
	5	0.7	1.9 ± 0.8	0
	5	1.0	3.3 ± 1.4	747
	7	0.7	3.2 ± 1.1	199
	7	1.0	5.4 ± 1.9	3274
	10	0.7	7.2 ± 1.7	4423
	10	1.0	11.3 ± 2.5	3466

Table 13: Positioning of AFFINITYENHANCER with respect to SOTA methods.

	IID optimization	OOD optimization	single-shot	improved binders with CDR edits
AFFINITYENHANCER (ours)	✓	✓	✓	✓
Property Enhancer (Tagasovska et al. (2024))	✓	✗	✗	✓
AntiFold ((Høie et al., 2024))	✓	✓	✗	✗
Walk-Jump, diffusion (Frey et al. (2023))	✓	✗	✓	✗
EffEVO, LM-based (Hie et al. (2024))	✓	✓	✗	✗
IgCraft (Greenig et al. (2025))	✓	✓	✓	✗
	✓	✗	✓	✗
Directed Evolution (Tran et al. (2025); Tran & Hy (2024))	✓	✗	✓	✗

Table 14: Comparison of AffinityEnhancer (AE) for Seeds 1 and Trastuzumab with MLDETran & Hy (2024)

Method	seed	ED	ED window	Binders	Improved	Binder rate	Improved rate
MLDE (low ED)	Seed 1	5.9 ± 0.8	98/128	32/98	0	34.7%	0%
AE (low ED)	Seed 1	5.2 ± 0.46	283/497	103/283	0	36.4%	0.0%
MLDE	Seed 1	16.2 ± 1.0	5/5000	0/5	0	0%	0%
MLDE	Trastuzumab	13.6 ± 1.1	816/5000	0/816	0	0%	0%
AE	Seed 1	6.5 ± 1.6	4382/5000	1,105	2	22.1%	0.04%
AE	Trastuzumab	7.9 ± 1.8	4815/5000	3970	1575	79.4%	31.5%

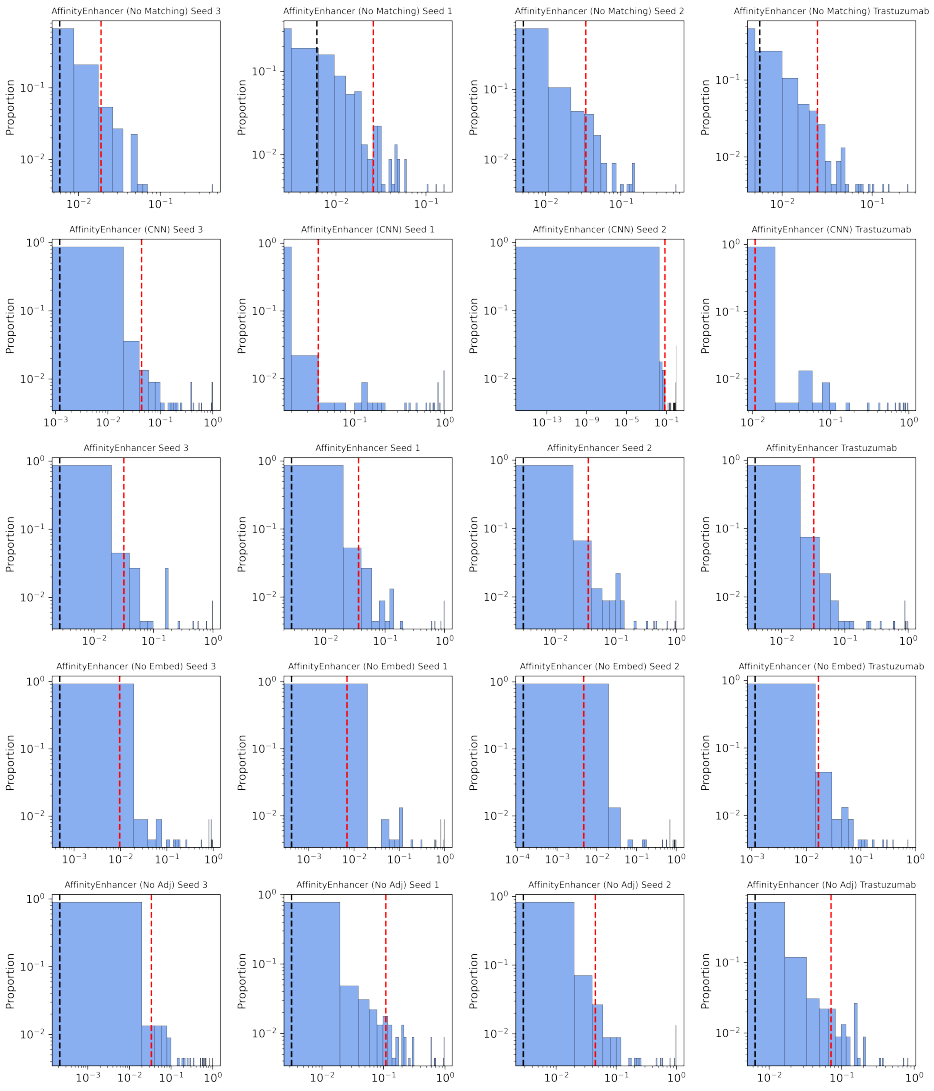

Figure 5: Distribution of fraction of designs with edits per-residue for each model and seed. Black and red dashed lines mark the 50th and 90th percentile respectively.

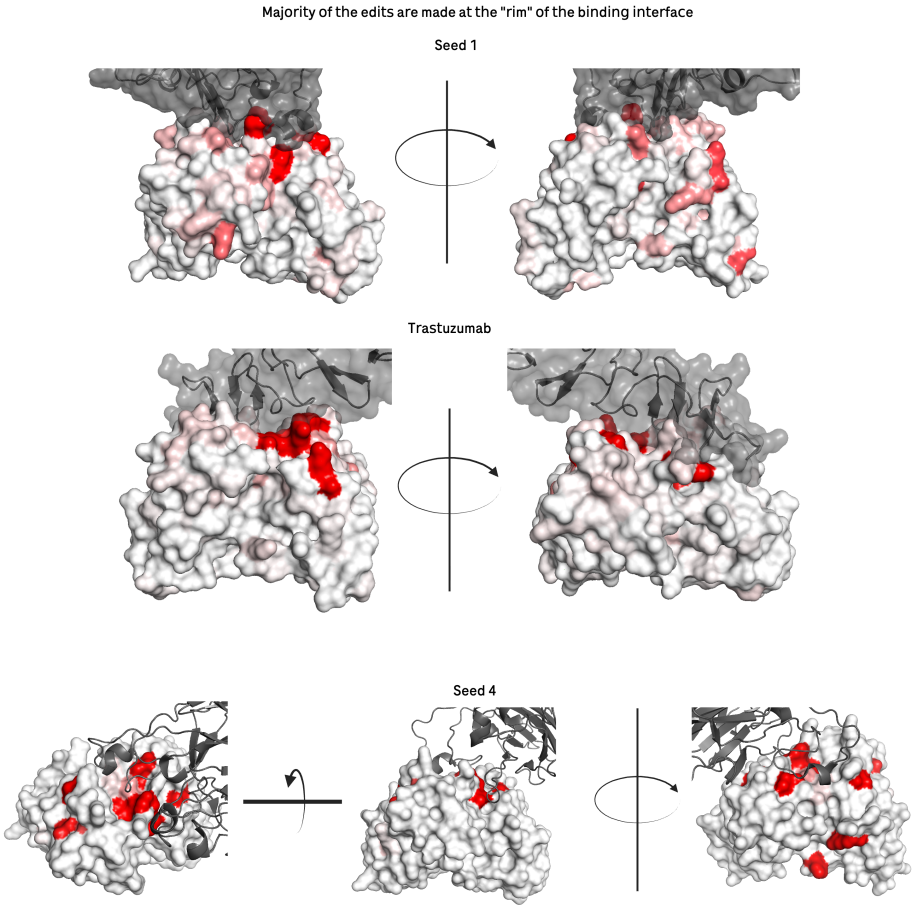

Figure 6: AffinityEnhancer identifies positions relevant to the antibody-antigen interface. (Top) Seed 1 in complex with its antigen. (Middle) Trastuzumab in complex with it's antigen HER2. (Bottom) Internal Seed 4 in complex with its antigen. Most edited positions by the AFFINITYENHANCER are colored red. Proposed affinity-enhancing positions are concentrated in the rim as opposed to the core of the binding surface.

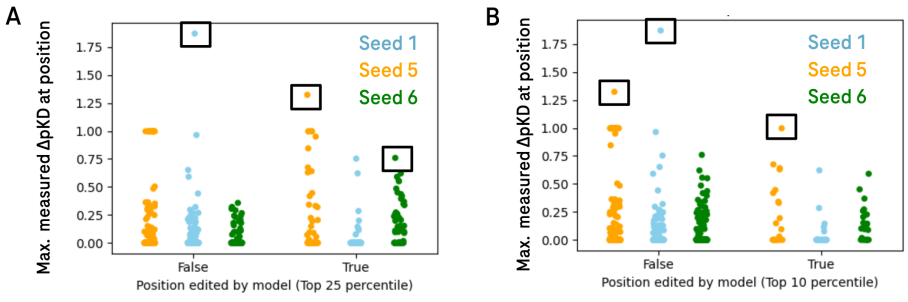

Figure 7: Comparison of AFFINITYENHANCER edits to experiments. Distribution of maximum measured improvement in pKD over seed for a position edited by AFFINITYENHANCER. A) "True" refers to a position in the top 25 percentile of the edited positions for that seed. B) "True" refers to a position in the top 10 percentile of the edited positions for that seed. Positions with the highest experimentally measured improvements in affinity are highlighted with a black box.

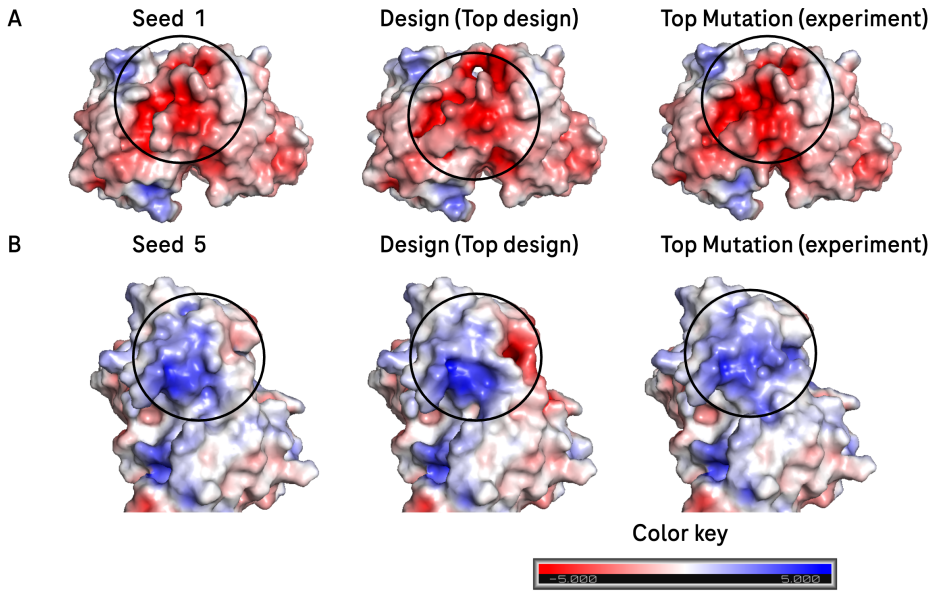

Figure 8: Comparison of amino acid substitutions with experiments. Comparison of the electrostatic surface of the Seed, top design and a top experimental single-point mutant for A) Seed 1 and B) Seed 5. Regions with the amino acid substitution in question is circled.