EVOTILT-MPNN: DISTRIBUTIONAL TILTING FOR VARIATION-AWARE BINDER DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Designed protein binders can fail as their targets evolve: a single interface point mutation can abolish binding, making escape a major failure mode in therapeutic and diagnostic applications. However, current binder-design methods condition on a single target, making their designs vulnerable to mutations. To address this, we introduce a structure-based binder-design approach in which each target-interface residue is specified by a distribution π over likely mutations. During autoregressive design, we sample each binder residue from an inverse-folding model's conditional distribution, tilted toward residues that reduce the KL divergence between π and the model's predicted distribution at the target interface. Memoization computes this tilt at a small additional cost. Tilting produces a consistent distributional response that transfers to held-out inverse-folding models across twenty-one targets. In silico co-folding on fifteen of these shows a small, heterogeneous, tail-concentrated improvement in worst-tail confidence across sampled target mutations, relative to an untilted ProteinMPNN baseline. Our method provides an efficient way to incorporate plausible target variation into sequence design. Because it accepts any per-residue distribution, π can be drawn from deep mutational scanning, phylogenetic frequencies, or escape predictors, making the approach a general route to variation-aware binder design.

1 INTRODUCTION

We introduce a binder-design method that conditions on how a target may vary rather than on a single fixed target sequence. Targets are not fixed: proteins are under continual evolutionary pressure, and their sequences change over time. Viral, bacterial, and tumor-associated proteins accumulate resistance mutations under immune and therapeutic selection. A single substitution at the interface can abolish recognition: the SARS-CoV-2 spike substitution E484K abolishes neutralization by bamlanivimab (Starr et al., 2021), and the EGFR ectodomain substitution S492R prevents cetuximab binding in colorectal cancer (Montagut et al., 2012). However, not all binders are equally vulnerable. For example, naturally elicited antibodies against the HIV-1 fusion peptide tolerate substitutions across the epitope, whereas vaccine-elicited antibodies to the same site lose binding to single mutations (Reveiz et al., 2024). Tolerance to substitutions is a property of the binder which can be optimized for: we call a binder *robust* when it retains binding across the substitutions its target is likely to undergo at the positions it engages.

Current structure-based binder-design methods condition on a single target and do not optimize for robustness. When robustness is sought at all, it is obtained afterwards by screening designs against explicit variants. Multistate design methods inherit that requirement: every variant must be enumerated in advance, realized as a structure, and scored (Löffler et al., 2017; Cheng et al., 2025). Instead of requiring variants to be enumerated explicitly, we let the designer supply a per-position distribution π_j over the residues that each interface position of the target might plausibly adopt, obtained from empirical lineage information, deep mutational scanning data, or from a protein language model.

A related line of work uses inverse-folding models as a specificity filter: designs are retained when the target-side posterior recovers the on-target target sequence more confidently than related off-target sequences (Liu et al., 2025). We use the same target-side posterior at the design stage, shaping it toward a distribution π over plausible target variants rather than filtering against a discrete off-

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.

target set. More concretely, we bias the inverse-folding model's per-residue sampling distribution during binder decoding: candidate residues that pull the model's target-side predictions toward π are favored. Each stepwise bias adds little cost over ordinary decoding.

We evaluate on 21 target families spanning an established *in silico* binder-design benchmark and a broader set of design targets (Stark et al., 2025). The resulting decoder is fast and shifts interface predictions toward the biological prior. When designs are challenged by co-folding against variant panels, we observe a heterogeneous improvement in worst-tail confidence at mild tilt. We frame this as a first step toward variation-aware binder design: a precise, cheap objective that can be composed with other stages of the design pipeline.

Contributions. We formulate robust binder design as an exponential tilt of an inverse-folding model toward a per-position prior π over tolerated target substitutions, without materializing any specific variant target structure. We give EvoTilt-MPNN, an algorithm that computes the tilt by caching leave-one-out decoding states at close to the cost of ordinary decoding. Finally, we characterize the result across 21 target families, with held-out inverse-folding scorers and single-mutant variant panels drawn from π alongside uniform and anti-π control arms.

2 Related work

Structure generation and inverse folding. De novo binder design is usually staged: a generative model proposes a binder backbone against a single target (Watson et al., 2023; Krishna et al., 2024; Didi et al., 2026; Stark et al., 2025), an inverse-folding model samples a sequence from that backbone (Dauparas et al., 2022), and subsequent structure-prediction filters score the resulting complex against the same single state (Abramson et al., 2024; Passaro et al., 2025). ProteinMPNN and related inverse-folding models such as ESM-IF1 (Hsu et al., 2022) and LigandMPNN (Dauparas et al., 2025) all expose a per-position conditional distribution over residue identity, not merely a sequence; most pipelines consume only its argmax or samples, whereas our method acts on the distribution itself.

Selectivity uses the same quantity, pointed the other way. Where designs must discriminate between related targets, the field already scores the target side and retains designs whose on-target score beats their off-target scores. Liu et al. (2025) run ProteinMPNN to predict the target peptide's sequence in the context of a designed binder, retaining designs whose on-target peptide is recovered more confidently than off-target proteome peptides; Sehgal et al. (2026) use AlphaFold3 predicted aligned error to distinguish on-target from off-target DNA at single-base resolution. Both are negative design implemented as post-hoc filters. The quantity in Liu et al. (2025) is the target-side posterior we tilt, applied at design time rather than as a filter; the structure-prediction confidence used by Sehgal et al. (2026) is the same kind of signal we rely on for co-folding evaluation.

Designing against more than one target state. Multistate design established the formalism: a single sequence is scored against a set of positive and negative structural states and optimized against their combination (Löffler et al., 2017). Its defining computational property is that the state set is enumerated and each state is realized as a structure. Learned methods inherit this requirement in different places, and almost all address structural variation of the target or the designed protein rather than sequence-level variation at a fixed interface: ProDualNet (Cheng et al., 2025) and DynamicMPNN (Abrudan et al., 2025) extend inverse folding to multiple supplied backbones, while ProDiT (Jing et al., 2025), SwitchCraft (Jing et al., 2026), and Chamaileon (Cao et al., 2026) take co-design or gradient-based routes over multiple explicit target contexts. Concurrent work on antiviral peptide design (Liang et al., 2026) enumerates variants from viral phylogenies and designs against that discrete variant set. Our formulation instead treats variation at the target's sequence directly: tolerated substitutions at fixed interface geometry are supplied as a per-position marginal, and no target variant is materialized.

3 Method

An inverse-folding model conditioned on a complex exposes a per-position posterior over target residues, which shifts as the binder changes. We tilt the design distribution toward binder sequences

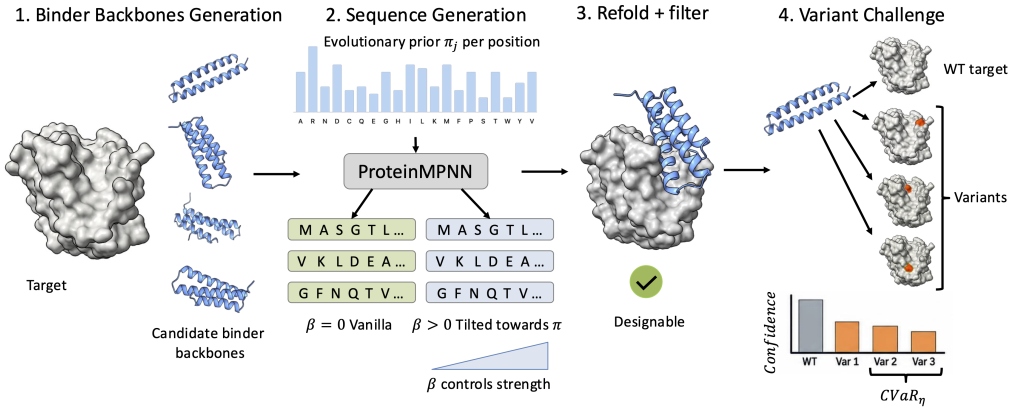


Figure 1: **Design and evaluation pipeline.** Binder backbones are generated conditioned on the reference target, then inverse-folded with either unmodified ($\beta = 0$) or tilted ($\beta > 0$) toward a per-position prior π over tolerated target substitutions. Surviving designs are challenged by co-folding against panels of target variants sampled from π or uniform distributions.

whose target-side posterior at the interface agrees with a prior over tolerated substitutions, without enumerating or building any variant. The full pipeline is shown in Figure 1; this section discusses only its sequence-design stage.

3.1 Setup

Let the target sequence be s^T and the binder sequence s^B, with s_v the residue at position v. The target has structure X_T (held fixed), and X_B the binder backbone structure. Write $\mathcal{A}$ for the amino-acid alphabet and n for the number of residue positions in the complex. The *interface* $\mathcal{I}$ is the set of target residues with a binder atom within 8 Å of their backbone, with $Q = |\mathcal{I}|$, computed for each binder candidate rather than specified in advance.

Throughout, $P_\theta(\cdot \mid X_B, s^B, s^T_{-j}, X_T)$ is the distribution an inverse-folding model places over the residue at target position j when that position is masked and the rest of the complex, binder backbone included, is given.

3.2 Robust design as distribution matching

A binder is robust to target position j when it accommodates plausible substitutions at that location. Given a categorical prior π_j over $\mathcal{A}$ describing that tolerated variation, we measure agreement by the Kullback–Leibler divergence, averaged over the interface:

$$D(s^B) \;=\; \frac{1}{Q}\sum_{j\in\mathcal{I}} \mathrm{KL}\big(\pi_j \,\big\|\, P_\theta(\cdot \mid X_B, s^B, s^T_{-j}, X_T)\big)\,. \tag{1}$$

We used this formulation with a mode-covering divergence so a binder is penalized for failing to accommodate a substitution which π_j deems plausible but not for tolerating ones it does not anticipate. We also divide by Q for two reasons. Numerically, it keeps β comparable across targets with different interface sizes. Conceptually, engaging a position is a commitment to handle its variation, and a design should not be penalized for making a larger interface.

We then incorporate this score into a design procedure. Let $p_0(s^B \mid X_B, s^T)$ be the distribution over binder sequences induced by an unmodified inverse-folding model on a fixed backbone. Rather than search for the s^B that minimizes D, we tilt p_0 by it:

$$p_\beta(s^B \mid X_B, s^T) \;\propto\; p_0(s^B \mid X_B, s^T)\,\exp\big(-\beta\, D(s^B)\big), \qquad \beta \geq 0. \tag{2}$$

At $\beta = 0$ this is exactly the unmodified design distribution (our vanilla ProteinMPNN baseline).

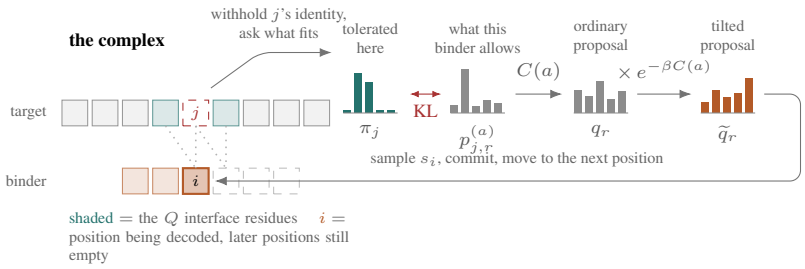


Figure 2: **One tilted decoding step.** At binder position i, each candidate residue a is scored by $C(a)$, the mean KL divergence between the prior π_j and the model's prediction $p^{(a)}_{j,r}$ at each interface position j, with j's identity withheld. The ordinary proposal q_r is reweighted by $e^{-\beta C(a)}$ to give the tilted proposal $\tilde{q}_r$, from which s_i is sampled.

Equation 2 cannot be sampled directly, since D depends on all of s^B at once, so we act one position at a time inside the decoder, with a greedy sequential approximation to p_β. The core operation is a withheld-position query: for each interface position $j \in \mathcal{I}$, we hide j's identity and read out the model's prediction at j given the decoded binder so far. Concretely, at step r decoding position i, let q_r be the model's ordinary proposal and, for candidate a and interface position j, let $p^{(a)}_{j,r}$ be its prediction at j with j itself withheld. Define

$$C_r(a) = \frac{1}{Q} \sum_{j \in \mathcal{I}} \mathrm{KL}\Big(\pi_j \,\Big\|\, p^{(a)}_{j,r}\Big), \qquad \widetilde{q}_r(a) = \frac{q_r(a)\, e^{-\beta C_r(a)}}{\sum_{b \in \mathcal{A}} q_r(b)\, e^{-\beta C_r(b)}}, \tag{3}$$

and sample $s_i \sim \widetilde{q}_r$: each candidate is reweighted by how far it pushes the target's withheld predictions away from π. Figure 2 shows one such step in context.

3.3 Exact computation by cached leave-one-out decoding

Evaluating equation 3 at every design step requires A candidate residues, each scored by Q leave-one-out predictions over the interface, for a total of $A \cdot Q$ decoder queries per position. Each query withholds a distinct interface residue j from the decoding order, which invalidates the running cache that makes ordinary decoding cheap: with j's identity removed at every layer, the cached states of j's neighbours no longer reflect the correct conditioning. Reusing the running cache and recomputing only node j is not a valid shortcut, and the error it introduces is not small relative to the effects we report. Without a fix, each prediction demands a fresh pass over all n positions, giving $\Theta(d\,A\,Q\,n\,L\,U)$ against $\Theta(n\,L\,U)$ for ordinary decoding (here L is the number of decoder layers, U the cost of one layer update, and d the number of designable binder positions).

We instead maintain $Q + 1$ *worlds* alongside the decoder. World $\perp$ is the ordinary decoder and supplies q_r; world j is the same decode with the identity of position j withheld everywhere while its geometry is retained. Each step appends the newly decoded position to all worlds, and the readout for position j takes its neighbours from world j, so every state entering it is j-free. This reproduces leave-one-at-a-time recomputation at cost $\Theta([\,n(Q+1) + d\,A\,Q\,]\,L\,U)$. A further exact reduction keeps only the interface queries whose neighbour list contains the position being decoded; this is the form used in all experiments. Appendix C provides additional information for this optimization.

4 Experimental setup

Pipeline. Binder backbones ($n = 10,000$ per target) are generated with BoltzGen (Stark et al., 2025) conditioned on the target. Each fixed backbone is inverse-folded with ProteinMPNN under both the unmodified decoder ($\beta = 0$) and the tilted decoder of Section 3 ($\beta > 0$). Every design is then re-folded with Boltz-2 from its sequence alone, templating the target backbone with no multiple sequence alignment. Refolded structures passing a self-consistency gate are ranked by predicted confidence and the top $K = 40$ (per target, β) are carried forward. Those selected designs are the ones put to the robustness challenge, described below.

Table 1: **Distributional response relative to the untilted baseline across 21 target families.** Each entry is the equal-family mean paired change $\Delta M = M(\beta) - M(0)$ with its between-family SEM; the parenthetical value counts families with $\Delta M < 0$. Negative KL and JS indicate closer agreement with the prior and negative log-likelihood a designability cost; entropy is descriptive.

β	ΔKL	ΔJS	Δentropy	Δlog-likelihood
5	-0.020 ± 0.004 (21)	-0.004 ± 0.001 (21)	0.002 ± 0.001 (11)	-0.002 ± 0.002 (10)
10	-0.039 ± 0.006 (21)	-0.007 ± 0.001 (21)	0.006 ± 0.003 (10)	-0.007 ± 0.003 (10)
15	-0.057 ± 0.008 (21)	-0.010 ± 0.002 (21)	0.012 ± 0.005 (10)	-0.013 ± 0.005 (10)
20	-0.074 ± 0.010 (21)	-0.013 ± 0.002 (21)	0.018 ± 0.007 (9)	-0.022 ± 0.008 (12)
100	-0.255 ± 0.019 (21)	-0.041 ± 0.004 (21)	0.104 ± 0.030 (6)	-0.191 ± 0.037 (20)

Targets and the prior. We use 21 target families in nested cohorts. All 21 are designed and wild-type-refolded; 17 complete the cross-variant co-folding challenge at $\beta \in \{0, 10, 20\}$, and seven of those also support the finer $\beta \in \{0, 5, 10, 15, 20, 100\}$ sweep. The seven fine-grid runs are drawn from an established de-novo binder benchmark for which *in silico* design is demonstrated. The primary analyses use the complete-on-interface subsets defined below: 15 families on the common grid and six on the fine grid. All experiments use one construction of π per target, a masked-position ESM-2 marginal truncated to the top 60% of probability mass to enforce sparsity and conserved-like positions. The truncated marginal distributions are re-normalized (Appendix A), and held fixed across all experiments.

Robustness challenge. Each target is challenged against a fixed, design-agnostic variant panel, built from π and target surface geometry (relative SASA ≥ 0.25). Substitutions are drawn under three arms, π, uniform, and anti-π, of which only π is ever used by the decoder, so the arms are held out from design. The panel consists of 40 single-mutant variants per arm, giving 121 co-folds per design. Full parameters, thresholds and per-family provenance are in Appendix A. Co-folding uses Boltz-2 with a backbone-only target template and no MSA. The common grid covers $\beta \in \{0, 10, 20\}$ on 17 targets, and the fine grid covers $\beta \in \{0, 5, 10, 15, 20, 100\}$ on seven. We record 13 confidence readouts, define retention as the variant-minus-wild-type change, and scope readouts as `on` (the mutated site lies on that design's own interface) or `off`. We define the primary evolutionary robustness metric as $\mathrm{CVaR}_{0.10}$ of `interface_iptm`, scope `on`, arm π, where CVaR_η denotes the mean of the worst η fraction of variant scores per design.

5 Results

5.1 Distributional tilting shifts interface predictions toward plausible target substitutions

We first test whether increasing the tilt changes the distributional quantities that the method acts on. Table 1 reports paired changes in interface KL divergence, interface Jensen–Shannon divergence, posterior entropy, and binder log-likelihood relative to the untilted $\beta = 0$ condition. The 21 family-level changes are pooled with equal weight.

Tilting produced a consistent dose response in both measures of agreement with the substitution prior. Interface KL and Jensen-Shannon divergence both decreased monotonically with β, with KL falling from -0.039 ± 0.006 nats at $\beta = 10$ to -0.255 ± 0.019 at $\beta = 100$ (a 15.2% reduction relative to the untilted mean). All 21 families had lower KL and JS than baseline at every positive tilt strength. The KL reduction could arise trivially, if the model simply grew more confident in the wild-type residue at each interface position, or substantively, if it reallocated mass among non-WT residues toward those favored by the prior. We decomposed the same-backbone KL reduction at prior-variable interface sites into these two components. To test whether the mutation-preference shift is prior-specific, we compared its reduction under prior-weighted (G_π) and uniform-weighted (G_{unif}) aggregation over non-WT residues (Appendix D). At $\beta = 10$, mutation-specific preference accounted for 75.0% of the KL reduction while mean WT posterior probability was essentially unchanged, and the conditional-mutant reduction was larger under the prior ($G_\pi = 0.0307$) than under uniform weighting ($G_{\text{unif}} = 0.0088$), with the contrast positive in all 21 families. Mild tilting therefore changes which substitutions the model favors rather than merely increasing WT confidence.

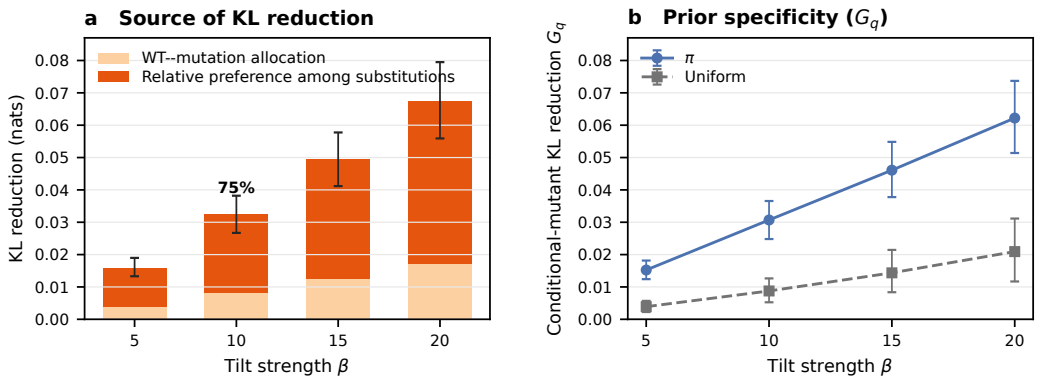


Figure 3: **KL reductions primarily reflect prior-specific redistribution among plausible substitutions. a,** Decomposition of the KL reduction into WT-mutation allocation and relative preference among substitutions, at each tilt strength. Bars are equal-family means across 21 families; the label gives the substitution-preference share at $\beta = 10$. **b,** Conditional-mutant KL reduction G_q under prior or uniform weighting over non-WT residues. Whiskers are 95% family-bootstrap intervals throughout; only prior-variable interface sites contribute.

Interface entropy also rose on average, though it fell below baseline in 6 of 21 families. The sequence-model likelihood cost was negligible at mild tilt ($\Delta = -0.007 \pm 0.003$ nats per residue at $\beta = 10$) but pronounced at $\beta = 100$ ($\Delta = -0.191 \pm 0.037$, lower than baseline in 20 of 21 families). This progressive tightening against the prior at a growing likelihood cost motivates separate structural evaluation at the mild-tilt regime ($\beta \leq 20$).

5.2 The distributional shift transfers to held-out inverse-folding models

The distributional shifts reported so far could be an artifact of optimizing and evaluating with the same model. We therefore re-scored the same designs with two held-out ProteinMPNN checkpoints and with ESM-IF1. We define the transferred fraction ϕ for each (family, held-out model) pair as the held-out KL decrease divided by the optimized-model KL decrease, then average across families with equal weight per family. To prevent unstable ratios, families whose optimized-model KL decrease is smaller than 0.02 nats are excluded from ϕ at that tilt strength.

The KL reduction transferred strongly between ProteinMPNN checkpoints and partially across architectures (Figure 4). At $\beta = 10$, the two held-out ProteinMPNN checkpoints retained $\phi = 0.75$ and 0.73, while ESM-IF1 retained 0.47; the family-level correlation between the optimized and ESM-IF1 KL decreases was strong (Pearson $r = 0.84$), and ESM-IF1 KL decreased in every family. The same pattern held at $\beta \geq 15$, with all 21 families entering the estimate ($\phi \in [0.71, 0.74]$ for ProteinMPNN checkpoints, $[0.44, 0.47]$ for ESM-IF1). The distributional response is thus not confined to the model weights used to construct the tilt, though cross-architecture transfer is incomplete.

5.3 Mild tilting largely maintains wild-type structural confidence

We next ask whether the distributional gain comes at the cost of the wild-type target. Mild tilting preserved wild-type confidence. Equal-family means moved by at most a few thousandths across all readouts: interface iPTM was 0.478 (vanilla), 0.476 ($\beta = 10$), and 0.475 ($\beta = 20$); binder pLDDT was 0.647, 0.644, and 0.641 (Table 2). The stricter downstream eligibility rate showed the same pattern, from 3.8% to 3.7% to 3.4%. Mild tilting therefore preserved average wild-type structural confidence; whether tolerance improves against target variation is addressed next (Section 5.4).

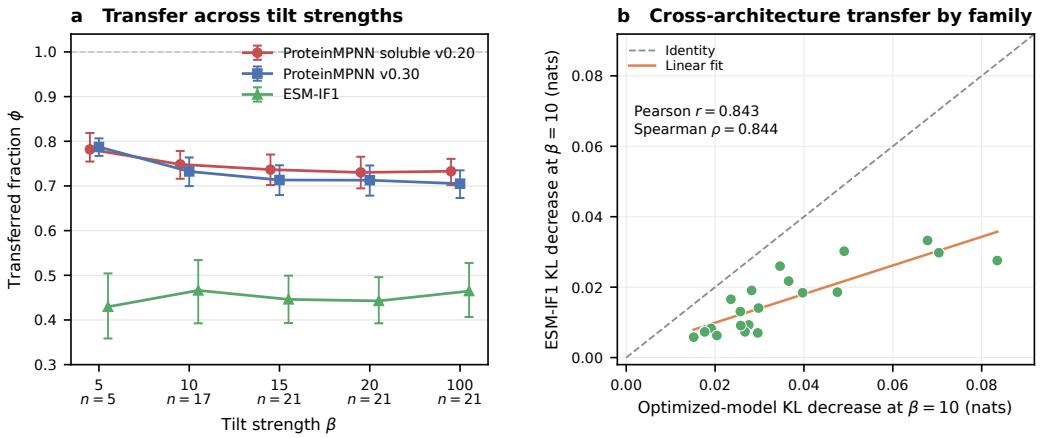


Figure 4: **a,** Equal-family mean transferred fraction ϕ across positive tilt strengths for two held-out ProteinMPNN checkpoints and ESM-IF1. Whiskers are 95% family-bootstrap confidence intervals; eligible family counts shown below the axis. **b,** Family-level optimized-model and ESM-IF1 KL decreases at $\beta = 10$. Each point is one family. The dashed line is identity; the solid line is the OLS fit.

Table 2: **Wild-type refolding confidence across 21 target families.** Entries are absolute, equally weighted means of the 21 family-specific means, with the number of families better than $\beta = 0$ in parentheses. RMSDs are in Å. Cohort, selection and eligibility criteria are as in Appendix A.

	Absolute equal-family mean (number of families better than $\beta = 0$)						
β	interface iPTM↑	binder pLDDT↑	complex pLDDT↑	global $\text{iPTM}_{\text{best5}}$↑	tgt-aln RMSD↓	binder RMSD↓	downstream eligible (%)↑
0	0.478 (–)	0.647 (–)	0.804 (–)	0.506 (–)	24.84 (–)	6.28 (–)	3.8 (–)
10	0.476 (8/21)	0.644 (6/21)	0.803 (6/21)	0.505 (9/21)	24.84 (13/21)	6.34 (6/21)	3.7 (12/21)
20	0.475 (9/21)	0.641 (2/21)	0.802 (3/21)	0.503 (9/21)	24.87 (13/21)	6.41 (2/21)	3.4 (13/21)

5.4 Co-folding against target variants tests evolutionary robustness

Section 5.3 showed WT confidence essentially unchanged at mild tilt; we now test whether the same tilt shifts confidence on plausible substitutions at the interface. The WT references below are the selected top-40 designs re-evaluated under the five-sample co-folding protocol and are therefore not directly comparable to the all-design refolding means in Table 2. At $\beta = 10$, the improvement was concentrated in the worst tail. Interface iPTM at the worst decile rose from 0.6300 to 0.6476 ($\Delta = +0.0176$, positive in 9 of 15 families), while the corresponding WT and mean-mutant changes were smaller. Complex PDE showed the same tail-focused pattern, with worst-decile PDE falling from 0.7384 to 0.7132 ($\Delta = -0.0252$, improved in 10 of 15 families). Thus, the largest gains appeared in the difficult mutant tail rather than in the WT reference.

The common-grid response attenuated at stronger tilt. At $\beta = 20$, WT interface iPTM fell from 0.8283 to 0.8196, worst-decile iPTM declined from its $\beta = 10$ peak of 0.6476 to 0.6387 but remained above the 0.6300 baseline, and worst-decile complex PDE drifted from 0.7132 back to 0.7264, still below the 0.7384 baseline. Across the broader readouts, 12 of 13 worst-decile readouts moved favorably at $\beta = 10$, but most moved back toward baseline at $\beta = 20$ (Appendix B). These results support a mild, tail-specific regime rather than monotonic improvement with tilt strength.

The finer six-strength sweep on the six-family complete-on-interface subset resolved this non-monotonic response more clearly. Worst-decile interface iPTM increased from 0.6586 at $\beta = 0$ to 0.6853 at $\beta = 5$ and 0.6859 at $\beta = 10$, then fell below baseline at $\beta \geq 15$. Complex PDE showed the complementary pattern, decreasing from 0.6841 to 0.6490 and 0.6450 before rising above baseline at $\beta \geq 15$ (Figure 5).

The on-interface prior-versus-uniform contrast pointed in the favorable direction, while the off-interface estimate was near zero; this specificity analysis is exploratory and reported in Figure 6.

Table 3: **Mild tilting improves the worst tail while maintaining WT confidence.** Equal-family means across the 15-family complete-on-interface cohort; all 40 selected designs contribute for every family and tilt strength. Parentheses: number of families improved over $\beta = 0$; bold: best value in each row. $\mathrm{CVaR}_{0.10}$ is the mean of the worst decile per design. Higher interface iPTM and lower complex PDE are favorable.

Readout	Quantity	$\beta = 0$	$\beta = 10$	$\beta = 20$
Interface iPTM ↑	WT reference	0.8283 (–)	**0.8302** (8/15)	0.8196 (6/15)
	Mean mutant ($\eta = 1$)	0.7925 (–)	**0.7969** (8/15)	0.7951 (7/15)
	Worst-decile $\mathrm{CVaR}_{0.10}$	0.6300 (–)	**0.6476** (9/15)	0.6387 (9/15)
Complex PDE ↓	WT reference	0.5244 (–)	**0.5172** (10/15)	0.5315 (5/15)
	Mean mutant ($\eta = 1$)	0.5554 (–)	**0.5458** (10/15)	0.5530 (9/15)
	Worst-decile $\mathrm{CVaR}_{0.10}$	0.7384 (–)	**0.7132** (10/15)	0.7264 (9/15)

5.5 Memoization computes the tilt at practical cost

Caching makes exact tilting practical. We timed complete sampler calls for vanilla ProteinMPNN, the original uncached exact tilt, and the two cached forms of Section 3.3 on a panel of 14 BoltzGen backbones drawn from seven target families (Table 7). The production local cached implementation was $147.0\times$ faster than the original exact computation and required only $1.40\times$ the wall-clock time of vanilla ProteinMPNN; local readout was $4.63\times$ faster than the full cached memoization program. Appendix C reports the interface geometry that determines the local speedup and the operation-count analysis behind these ratios.

6 Discussion

Across the readouts, the tilt behaves as designed at the distributional level and produces a small effect at the structural level. Interface KL falls monotonically with β in all 21 families, and the reduction transfers to held-out inverse-folding checkpoints; at mild tilt ($\beta = 10$), wild-type confidence is essentially unchanged and worst-decile substitution confidence improves on 12 of 13 readouts, with the primary endpoint (worst-decile interface iPTM) improving in 9 of 15 families. The effect attenuates at $\beta = 20$ on the 15-family common grid and reverses from $\beta \geq 15$ in the six-family finer sweep. The non-monotonic response likely reflects an expected trade-off: interface-adjacent binder residues must accommodate target variation while still satisfying local packing and global folding constraints. As β increases, the tilt increasingly overrides ProteinMPNN's structural preferences at these positions. Consistent with this interpretation, binder log-likelihood declines monotonically and wild-type refolding confidence deteriorates, eventually outweighing the tolerance gained at mild tilt.

The gap between the clean distributional response and the small structural response constrains what a stronger version of the method should aim for. Interface KL decreases by 15% at $\beta = 100$, while 75% of the mild-tilt reduction reflects mutation-specific redistribution rather than increased wild-type confidence; the distributional prior-versus-uniform contrast is positive in all 21 families (Section 5.1). The structural response, by contrast, is small. One explanation is limited assay sensitivity: single substitutions often weakly perturb predicted structures, and only a small, design-dependent subset of the fixed panel intersects each binder's modeled interface, leaving some worst-tail estimates supported by few mutations. The favorable on-interface specificity contrast and near-zero off-interface contrast are directionally encouraging, but this secondary analysis is exploratory and its family-level confidence interval includes zero. Larger interface-targeted panels or multi-substitution challenges may provide greater dynamic range. Second, interface divergence weights all Q positions alike, whereas the positions a particular design depends on for binding are a subset that π knows nothing about. Whether π is well-calibrated to binding sensitivity, and whether a measured prior (for example, from deep mutational scanning) rather than a language-model-inferred one behaves differently, are the most likely routes to a larger effect and are untested here. A related direction lies in the decoder itself: the current greedy per-position sampler approximates the joint tilt p_β (Equation 2) by a product of per-step tilts, and a beam search over the tilted proposal would coordinate choices across positions and close part of that gap.

The modest effect size may partly reflect where the objective enters the pipeline. EvoTilt can change the binder sequence, but the backbone geometry, epitope, interface footprint, and binding orientation have already been fixed by a generator conditioned on the wild-type target. Sequence design therefore has limited freedom to accommodate target variation while preserving local packing, folding, and wild-type compatibility. Some robust solutions may require a different interface geometry altogether. Incorporating the distributional compatibility score as a reward or conditioning signal during backbone generation would allow sequence and structure to adapt jointly and could produce stronger effects.

7 LIMITATIONS

Several limitations bound our claims. All evidence is computational: co-fold confidence is not measured affinity, and Boltz-2 evaluation is not fully independent of BoltzGen backbone generation. The held-out inverse-folding models establish transfer of the distributional surrogate, not independent functional validation; orthogonal physics-based evaluation and ultimately experimental binding measurements are needed. Moreover, all experiments use one protein-language-model-derived prior. Its marginal distributions omit substitution co-occurrence and epistasis, while the design-agnostic single-mutant panels may not probe the positions most important to every binder. The modest and heterogeneous structural effects are estimated across only 15 target families with complete on-interface design coverage, limiting power for family-level inference. Additional independent families would improve statistical power, while larger variant panels would stabilize worst-tail estimates. Multi-mutant challenges could provide greater dynamic range, although they constitute a distinct and more stringent robustness test.

Despite these limitations, the variant panels and co-folding evaluation protocol constitute a reusable benchmark for the community to assess variation-aware binder-design methods. The present targets are established *in silico* design systems rather than clinically relevant cases. Future work should extend this benchmark to evolving therapeutic and diagnostic targets using empirical priors derived from deep mutational scanning, phylogenetic surveillance, or measured escape, together with physics-based and experimental validation.

8 CONCLUSION

We formulated variation-aware binder design as distributional matching rather than optimization over enumerated target states. Plausible target substitutions are supplied as per-position marginals, and the autoregressive binder proposal is tilted toward sequences whose target-side posterior better matches those distributions. No target variant is instantiated during design, and cached leave-one-out decoding computes each stepwise tilt exactly at only $1.40\times$ the runtime of vanilla ProteinMPNN. Across 21 target families, tilting produced a consistent distributional response that transferred to held-out inverse-folding models. In the 15-family complete-on-interface co-folding analysis, mild tilting yielded a modest, directionally favorable change in worst-tail confidence that attenuated or reversed at stronger tilt. Together, the method and benchmark provide a practical foundation for variation-aware binder design, while the limited structural effect suggests that sequence optimization on a fixed backbone leaves important degrees of freedom inaccessible. Incorporating the same objective during backbone generation, where the interface geometry and footprint are chosen, is a promising next step.

AI USE STATEMENT

We used large-language-model assistants (OpenAI ChatGPT/Codex and Anthropic Claude) to draft and edit prose, locate candidate literature, check bibliographic metadata, and assist in writing, reviewing, and debugging research and analysis code. AI was used to help with code implementation and refactoring. These tools were not treated as evidentiary sources. The authors ran the scripts, made the final experimental and interpretive decisions, recomputed quantitative claims from stored outputs, and checked literature claims against primary sources.

ETHICS STATEMENT

This study is entirely computational and uses public protein data and published model checkpoints. It involves no human participants, animals, clinical records, or personally identifying data. Variation-aware binder design has potential value for therapeutics and diagnostics, but protein-design capabilities can be dual use. Although this work does not modify pathogens or experimentally enhance pathogenicity, applications to biological targets should be conducted under appropriate institutional biosafety, biosecurity, and legal oversight.

REPRODUCIBILITY STATEMENT

The appendices specify the model inputs, target cohorts, selection rules, variant-panel construction, evaluation metrics, and runtime protocol. The source code will accompany the paper release.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630:493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Alex Abrudan, Sebastian Pujalte Ojeda, Chaitanya K. Joshi, Matthew Greenig, Felipe Engelberger, Alena Khmelinskaia, Jens Meiler, Michele Vendruscolo, and Tuomas P. J. Knowles. Multi-state protein design with DynamicMPNN. *arXiv preprint arXiv:2507.21938*, 2025.

Hengyuan Cao, Shizhuo Cheng, Mingxuan Liu, Weicheng Huang, Yunhong Lu, Chenxi Cai, Yan Zhang, and Min Zhang. Chamaileon: Cross-context binder design with contextualized modeling and mixed sampling. *arXiv preprint arXiv:2607.23518*, 2026.

Liu Cheng, Ting Wei, Xiaochen Cui, Hai-Feng Chen, and Zhangsheng Yu. ProDualNet: dual-target protein sequence design method based on protein language model and structure model. *Briefings in Bioinformatics*, 26(4):bbaf391, 2025. doi: 10.1093/bib/bbaf391.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187.

Justas Dauparas, Gyu Rie Lee, Robert Pecoraro, Linna An, Ivan Anishchenko, Cameron Glasscock, and David Baker. Atomic context-conditioned protein sequence design using LigandMPNN. *Nature Methods*, 22(4):717–723, 2025. doi: 10.1038/s41592-025-02626-1.

Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, et al. Scaling atomistic protein binder design with generative pretraining and test-time compute. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qmCpJtFZra.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8946–8970, 2022. URL https://proceedings.mlr.press/v162/hsu22a.html.

Bowen Jing, Anna Sappington, Mihir Bafna, Ravi Shah, Adrina Tang, Rohith Krishna, Adam Klivans, Daniel J. Diaz, and Bonnie Berger. Generating functional and multistate proteins with a multimodal diffusion transformer. *bioRxiv*, 2025. doi: 10.1101/2025.09.03.672144.

Bowen Jing, Mihir Bafna, Anisha Parsan, Heyuan Michael Ni, David Kwabi-Addo, Bryan Bryson, Adam Klivans, and Bonnie Berger. SwitchCraft: A programmatic framework for designing state-switching proteins. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.31236.

Rohith Krishna, Jue Wang, Woody Ahern, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, 384(6693):eadl2528, 2024. doi: 10.1126/science.adl2528.

Kimberly Liang, Navya Nori, Sophia Vincoff, Elizabeth H. Mahood, Yinuo Zhang, and Pranam Chatterjee. PROPHET: Phylogenetically robust antiviral peptide design against heterogeneous evolutionary trajectories. In *The 2026 Workshop on Generative and Agentic AI for Biology (GenBio 2026) at ICML*, 2026. URL https://openreview.net/forum?id=zQJ0Ll2IR5.

Bingxu Liu, Nathan F. Greenwood, Julia E. Bonzanini, et al. Design of high-specificity binders for peptide-MHC-I complexes. *Science*, 389(6758):386–391, 2025. doi: 10.1126/science.adv0185.

Patrick Löffler, Samuel Schmitz, Enrico Hupfeld, Reinhard Sterner, and Rainer Merkl. Rosetta:MSF: a modular framework for multi-state computational protein design. *PLOS Computational Biology*, 13(6):e1005600, 2017. doi: 10.1371/journal.pcbi.1005600.

Clara Montagut, Alba Dalmases, Beatriz Bellosillo, et al. Identification of a mutation in the extracellular domain of the epidermal growth factor receptor conferring cetuximab resistance in colorectal cancer. *Nature Medicine*, 18(2):221–223, 2012. doi: 10.1038/nm.2609.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707.

Mateo Reveiz, Kai Xu, Myungjin Lee, Shuishu Wang, Adam S. Olia, Darcy R. Harris, Kevin Liu, Tracy Liu, Andrew J. Schaub, Tyler Stephens, Yiran Wang, Baoshan Zhang, Rick Huang, Yaroslav Tsybovsky, Peter D. Kwong, and Reda Rawi. Vaccine-elicited and naturally elicited antibodies differ in their recognition of the HIV-1 fusion peptide. *Frontiers in Immunology*, 15: 1484029, 2024. doi: 10.3389/fimmu.2024.1484029.

Enisha Sehgal, Yuliya Politanska, Raktim Mitra, Paul T. Kim, Nayim González Rodríguez, Tushar Warrier, et al. Generative design of sequence specific DNA binding proteins. *bioRxiv*, 2026. doi: 10.64898/2026.04.27.720408.

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy John O'Donnell, et al. BoltzGen: Toward universal binder design. *bioRxiv*, 2025. doi: 10.1101/2025.11.20.689494.

Tyler N. Starr, Allison J. Greaney, Adam S. Dingens, and Jesse D. Bloom. Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. *Cell Reports Medicine*, 2(4):100255, 2021. doi: 10.1016/j.xcrm.2021.100255. URL https://pubmed.ncbi.nlm.nih.gov/33842902/.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620:1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

A EXPERIMENTAL DETAILS

A.1 TARGETS

We evaluate on 21 target families spanning 83–288 residues, drawn from three groups. Seven come from an established *in silico* binder benchmark: `bhrf1`, `il7ra`, `insulin`, `mdm2`, `pdgfr_d23`, `pdl1`, `trka`. The benchmark is selected for designability, so we added ten further families: `a1m`, `calm`, `egfr_d3`, `gm2a`, `hnmt`, `idi2`, `mzb1`, `orm2`, `rfk`, `tie2`. Together, these 17 families completed the cross-variant co-folding runs. The primary on-interface aggregation contains the 15 families with complete 40-design coverage at $\beta = 0/10/20$; `insulin` and `rfk` are excluded only from that aggregation. Four additional families (`cd5`, `trem2`, `mgat1`, `mbp`) are designed and wild-type-refolded but not co-folded.

A.2 PRIOR CONSTRUCTION

We query ESM-2 (`esm2_t33_650M_UR50D`) with each target position masked and the surrounding window as context, then truncate the predicted marginal to the residues carrying the top 60% of probability mass and renormalize.

A.3 BACKBONE GENERATION AND TILTED SEQUENCE DESIGN

Binder backbones are generated with BoltzGen conditioned on the target, $10{,}000$ per family, with binder lengths sampled from per-target ranges (spanning 40–140 residues). Backbones are held fixed across every condition so that variation across β isolates the sequence-design stage. For each backbone we sample four sequences at each $\beta \in \{0, 5, 10, 15, 20, 100\}$, giving $240{,}000$ designs per family.

A.4 DESIGNABILITY FILTER AND SELECTION

Every designed sequence is re-folded with Boltz-2, which templates the target backbone and predicts the complex from the binder sequence alone. A design passes the self-consistency gate if the re-folded binder matches its design backbone within 2.0Å RMSD and the target-aligned complex matches within 2.5,Å. Survivors are ranked by a weighted combination of ipTM and pTM, and the top $K = 40$ per (target, β) are carried forward (Table 4).

Table 4: **Wild-type refolding metrics by target family.** Each cell reports the absolute family mean in the order $\beta = 0/10/20$. RMSDs are in Å; downstream co-folding eligibility is reported as a percentage.

Target	interface iPTM↑	binder pLDDT↑	complex pLDDT↑	global iPTM$_{\text{best5}}$↑	tgt-aln RMSD↓	binder RMSD↓	downstream eligible (%)↑
mdm2	0.558 / 0.563 / **0.565**	**0.675** / **0.675** / 0.673	0.785 / **0.786** / 0.784	0.610 / **0.616** / **0.616**	12.06 / **11.88** / 11.93	5.62 / **5.54** / 5.65	12.3 / 12.6 / **12.7**
insulin	0.441 / 0.447 / **0.449**	**0.706** / 0.700 / 0.698	**0.842** / 0.839 / 0.839	0.455 / 0.461 / **0.463**	31.13 / 30.78 / **30.76**	**4.60** / 4.75 / 4.75	4.0 / 5.2 / **5.6**
bhrf1	**0.615** / 0.598 / 0.575	**0.752** / 0.744 / 0.729	**0.864** / 0.861 / 0.854	**0.734** / 0.716 / 0.692	**17.88** / 18.59 / 19.56	**7.11** / 7.27 / 7.63	**19.3** / 16.9 / 13.4
trka	**0.547** / 0.545 / 0.540	**0.697** / 0.691 / 0.686	**0.816** / 0.813 / 0.810	**0.560** / 0.559 / 0.555	**17.58** / 17.81 / 17.95	**5.49** / 5.72 / 5.84	**6.5** / 6.0 / 5.6
pdgfr_d23	**0.550** / 0.540 / 0.531	**0.690** / 0.684 / 0.679	**0.803** / 0.800 / 0.798	**0.550** / 0.541 / 0.531	**16.44** / 17.10 / 17.92	**3.88** / 4.01 / 4.13	**4.5** / 3.6 / 2.8
pdl1	0.627 / **0.630** / 0.628	**0.786** / **0.786** / 0.785	**0.855** / **0.855** / **0.855**	0.636 / **0.638** / 0.636	13.08 / 13.02 / **12.92**	**2.85** / **2.85** / 2.88	6.1 / 6.6 / **6.8**
il7ra	**0.432** / 0.427 / 0.425	**0.660** / 0.655 / 0.650	**0.829** / 0.827 / 0.826	**0.435** / 0.430 / 0.429	**36.06** / 36.29 / 36.46	**5.27** / 5.44 / 5.59	**1.1** / 1.0 / 0.9
cd5	**0.502** / 0.499 / 0.495	**0.609** / 0.604 / 0.596	**0.758** / 0.756 / 0.753	**0.513** / 0.511 / 0.507	32.46 / **32.41** / 32.42	**6.72** / 6.85 / 7.01	**0.1** / **0.1** / **0.1**
trem2	**0.513** / 0.508 / 0.512	**0.672** / 0.671 / 0.667	**0.817** / 0.816 / 0.815	**0.519** / 0.514 / 0.518	32.56 / 32.51 / **32.24**	**5.60** / 5.61 / 5.73	0.2 / **0.3** / **0.3**
calm	0.538 / 0.539 / **0.540**	0.543 / 0.546 / **0.548**	0.683 / 0.685 / **0.687**	0.565 / 0.568 / **0.569**	21.77 / 21.66 / **21.23**	10.84 / 10.76 / **10.70**	1.3 / 1.4 / **1.6**
rfk	0.362 / **0.366** / **0.366**	**0.571** / **0.571** / 0.570	0.708 / **0.709** / 0.708	0.395 / **0.400** / 0.399	18.02 / 17.87 / **17.55**	7.58 / **7.57** / 7.59	2.1 / **2.5** / 2.4
mzb1	**0.538** / 0.534 / **0.538**	**0.698** / 0.696 / 0.696	**0.787** / 0.786 / 0.786	0.541 / 0.538 / **0.542**	32.27 / 32.15 / **32.10**	**5.17** / 5.25 / 5.23	1.2 / 1.2 / **1.4**
gm2a	0.445 / 0.445 / **0.446**	0.526 / 0.526 / **0.527**	**0.709** / **0.709** / **0.709**	0.509 / **0.510** / **0.510**	23.35 / 23.23 / **22.77**	**9.28** / 9.35 / 9.29	2.0 / 2.3 / **2.8**
a1m	0.540 / 0.543 / **0.551**	**0.661** / 0.658 / 0.657	**0.800** / **0.800** / 0.799	0.610 / 0.613 / **0.621**	20.31 / 19.82 / **19.26**	**5.75** / 5.82 / 5.85	1.3 / **1.4** / 1.3
orm2	**0.497** / 0.494 / 0.493	**0.668** / 0.666 / 0.663	**0.805** / 0.804 / 0.803	**0.570** / 0.569 / 0.567	**20.93** / 21.14 / 21.06	**5.62** / 5.63 / 5.70	**2.6** / 2.2 / 2.1
tie2	**0.491** / 0.489 / 0.485	**0.664** / 0.661 / 0.653	**0.810** / 0.809 / 0.806	**0.512** / **0.512** / 0.509	**18.92** / 19.18 / 19.09	**5.14** / 5.26 / 5.45	**6.9** / 6.2 / 5.3
egfr_d3	**0.490** / **0.490** / 0.489	**0.711** / **0.711** / 0.708	**0.842** / **0.842** / 0.841	**0.494** / **0.494** / 0.493	19.77 / 19.75 / **19.56**	4.08 / **4.06** / 4.15	2.4 / 2.3 / **2.5**
idi2	0.365 / **0.366** / 0.361	**0.622** / 0.621 / 0.617	**0.839** / **0.839** / 0.837	**0.373** / **0.373** / 0.368	**28.33** / 28.51 / 29.13	**6.41** / 6.44 / 6.51	3.7 / **3.9** / 3.1
hnmt	**0.331** / 0.328 / 0.326	**0.564** / **0.564** / 0.561	**0.811** / **0.811** / 0.810	**0.357** / 0.354 / 0.353	**35.07** / 35.35 / 35.92	**8.71** / 8.75 / 8.82	**1.4** / 1.0 / 0.9
mgat1	**0.323** / 0.320 / 0.322	**0.580** / 0.579 / 0.578	**0.859** / 0.858 / 0.858	0.341 / 0.340 / **0.342**	39.73 / **39.17** / 39.41	7.80 / **7.78** / 7.80	**0.2** / **0.2** / **0.2**
mbp	0.332 / 0.334 / **0.336**	**0.523** / **0.523** / 0.522	**0.858** / 0.857 / 0.857	0.340 / 0.342 / **0.344**	33.91 / 33.35 / **33.04**	8.41 / **8.36** / 8.41	0.3 / **0.4** / **0.4**

A.5 VARIANT PANELS

Sites are surface-exposed residues (FreeSASA relative SASA ≥ 0.25) ranked by escape mass under π. At each site, each arm produces an ordered candidate list: π-weighted (Efraimidis-Spirakis),

uniform over the 19 non-wild-type residues, or ordered by ascending mass under the untruncated ESM prior (anti-π). Sites are visited in escape-mass order via round-robin: round 0 takes each site's top candidate, round 1 the next, and so on, stopping at 40 total (site, substitution) pairs per target and arm. This fixed target-level panel gives 121 co-folds per design (one wild-type plus 120 variants across three arms). Anti-π weighting is close to flat over 20 residues, so this arm is mildly rather than strongly adversarial.

A.6 Co-folding and readouts

Each (design, variant) pair is co-folded with Boltz-2 using a backbone-only target template, no MSA, and five diffusion samples, scored on the best of five. Co-folding covers $\beta \in \{0, 10, 20\}$ on all 17 co-folded targets and extends to $\beta \in \{0, 5, 10, 15, 20, 100\}$ on seven. A design contributes to an `on` statistic only if at least one panel substitution hits its modeled interface. To avoid unequal design coverage, the primary common-grid aggregation requires all 40 selected designs to contribute at each tilt strength. Fifteen families meet this outcome-independent coverage rule; `insulin` ($7/7/14$ contributing designs at $\beta = 0/10/20$) and `rfk` ($31/27/29$) are excluded only from co-fold aggregation and remain in every upstream analysis. Applying the same common-grid rule to the fine sweep leaves six families. All six contribute 40 designs through $\beta = 20$; at $\beta = 100$, `il7ra` and `pdgfr_d23` contribute 30 each.

A.7 Held-out scorers

We re-score the same designs under three held-out models: two sibling ProteinMPNN checkpoints (`v_48_020.pt` from `soluble_model_weights` and `v_48_030.pt` from `vanilla_model_weights`) and ESM-IF1 (`esm_if1_gvp4_t16_142M_UR50`). This is a read-only recompute from stored designs, with no re-design and no re-folding, on a frozen subsample of 1,000 backbones per family; decode order is shared across models and tilt strengths.

B CROSS-VARIANT CHALLENGE CONFIDENCE READOUT SET

Table 5 reports the complete confidence readouts for the same π-arm, on-interface, worst-decile analysis used for the primary endpoint.

Table 5: **Worst-decile absolute co-fold confidence across all stored readouts.** Values are equal-family means over the 15-family complete-on-interface common grid; all 40 selected designs contribute in every family and condition. Parentheses give the number of families moving in the favorable direction relative to $\beta = 0$ (out of 15).

Readout	$\beta = 0$	$\beta = 10$	$\beta = 20$
Interface iPTM ↑	0.6300	0.6476 (9/15)	0.6387 (9/15)
Global iPTM ↑	0.6451	0.6656 (11/15)	0.6549 (10/15)
Confidence score ↑	0.8177	0.8259 (9/15)	0.8204 (7/15)
Complex ipLDDT ↑	0.8097	0.8154 (9/15)	0.8097 (8/15)
Geometric-interface pLDDT ↑	0.8338	0.8352 (9/15)	0.8331 (6/15)
Complex pLDDT ↑	0.8559	0.8613 (10/15)	0.8571 (9/15)
Binder pLDDT ↑	0.7969	0.8073 (10/15)	0.7975 (7/15)
Mutated-residue pLDDT ↑	0.6757	0.6807 (10/15)	0.6722 (6/15)
Binder pTM ↑	0.8877	0.8938 (8/15)	0.8874 (8/15)
pTM ↑	0.8056	0.8119 (11/15)	0.8083 (9/15)
Target pTM ↑	0.9025	0.9022 (7/15)	0.9019 (9/15)
Target pLDDT ↑	0.8842	0.8861 (8/15)	0.8848 (7/15)
Complex PDE ↓	0.7384	0.7132 (10/15)	0.7264 (9/15)

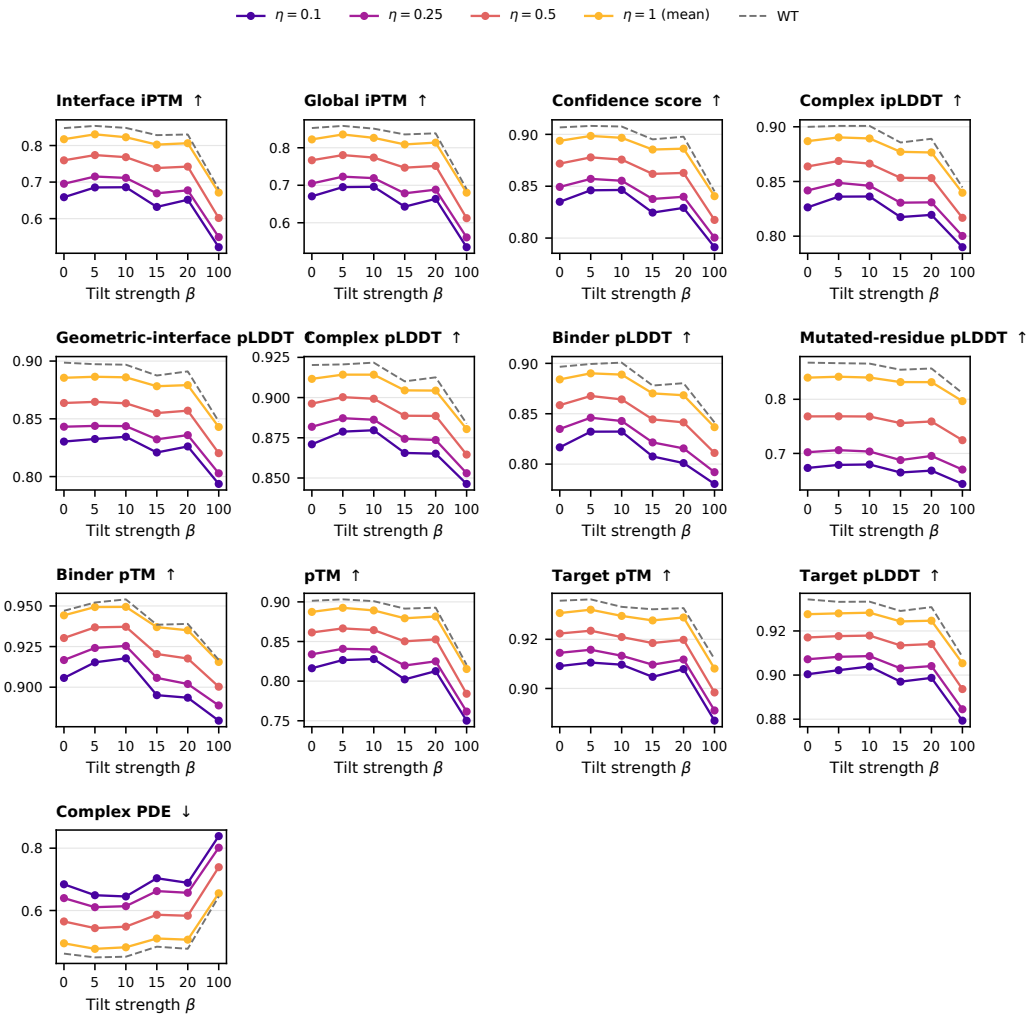


Figure 5: **Complete six-strength co-fold confidence readouts in the six-family complete-on-interface subset.** Each panel is an absolute equal-family mean for π-drawn on-interface substitutions. Colored curves show CVaR_η across four worst-tail fractions; the dashed gray line shows the matched WT reference. All six families contribute 40 designs through $\beta = 20$; at $\beta = 100$, `il7ra` and `pdgfr_d23` contribute 30 designs each.

Table 6: **Primary co-fold confidence readouts by target family.** The 15 rows are the complete-on-interface analysis cohort. Each cell reports the absolute value in the order $\beta = 0/10/20$ for π-drawn single substitutions that hit the design's modeled interface. Mean-mutant and worst-decile CVaR quantities follow Table 3; n_{design} gives the contributing design count at each tilt strength.

Target	n_{design}	iPTM WT↑	iPTM mean↑	iPTM $\mathrm{CVaR}_{0.10}$↑	PDE WT↓	PDE mean↓	PDE $\mathrm{CVaR}_{0.10}$↓
a1m	40 / 40 / 40	0.7321 / **0.7403** / 0.7045	**0.7129** / 0.7098 / 0.6879	**0.5830** / 0.5727 / 0.5499	0.5401 / **0.5313** / 0.5509	0.5577 / 0.5572 / **0.5546**	**0.6701** / 0.6786 / 0.6738
bhrf1	40 / 40 / 40	0.8665 / **0.8859** / 0.8352	0.8781 / **0.9104** / 0.8553	0.7442 / **0.8223** / 0.7069	0.4420 / **0.3995** / 0.4675	0.4275 / **0.3841** / 0.4531	0.5873 / **0.4838** / 0.6292
calm	40 / 40 / 40	0.7264 / 0.7513 / **0.7655**	0.7184 / 0.7497 / **0.7669**	0.6130 / 0.6421 / **0.6662**	0.7244 / 0.7042 / **0.6668**	0.7431 / 0.7066 / **0.6634**	0.9943 / 0.9374 / **0.9001**
egfr_d3	40 / 40 / 40	0.8226 / **0.8393** / 0.7969	0.7948 / 0.8008 / **0.8027**	0.6655 / **0.6793** / 0.6731	0.4658 / **0.4651** / 0.4899	0.4886 / 0.4902 / **0.4885**	0.5895 / **0.5847** / 0.6009
gm2a	40 / 40 / 40	0.7734 / 0.7744 / **0.7808**	0.7613 / 0.7569 / **0.7785**	0.5769 / 0.5675 / **0.6466**	0.7175 / **0.6606** / 0.6817	0.7372 / **0.6712** / 0.6847	1.0106 / 0.9108 / **0.8695**
hnmt	40 / 40 / 40	**0.8068** / 0.7480 / 0.7912	**0.7235** / 0.6756 / 0.7152	**0.3534** / 0.3465 / 0.3505	**0.5988** / 0.6580 / 0.6146	**0.6592** / 0.6995 / 0.6692	**0.9079** / 0.9214 / 0.9293
idi2	40 / 40 / 40	0.8617 / **0.9332** / 0.8773	0.6969 / **0.7758** / 0.7646	0.4500 / **0.5618** / 0.5021	0.4278 / **0.3763** / 0.4158	0.5207 / **0.4668** / 0.4855	0.6626 / **0.5809** / 0.6327
il7ra	40 / 40 / 40	0.6733 / 0.6669 / **0.6805**	0.6261 / 0.6300 / **0.6461**	0.4549 / **0.4881** / 0.4829	**0.5972** / 0.6160 / 0.6019	0.6373 / 0.6410 / **0.6262**	0.7735 / 0.7578 / **0.7505**
mdm2	40 / 40 / 40	**0.9335** / 0.9240 / 0.9070	0.9087 / **0.9173** / 0.8951	0.7591 / **0.8018** / 0.7660	**0.3638** / 0.3659 / 0.3891	0.3930 / **0.3771** / 0.4023	0.6038 / **0.5457** / 0.5700
mzb1	40 / 40 / 40	0.8456 / 0.8450 / **0.8864**	0.8375 / 0.8446 / **0.8893**	0.7497 / 0.7687 / **0.8347**	0.5879 / 0.5671 / **0.5198**	0.5924 / 0.5779 / **0.5191**	0.7449 / 0.7226 / **0.6197**
orm2	40 / 40 / 40	**0.8587** / 0.8464 / 0.8358	**0.8299** / 0.8267 / 0.8234	0.6930 / 0.6858 / **0.7029**	0.5155 / **0.5130** / 0.5257	0.5369 / 0.5357 / **0.5323**	0.6662 / 0.6891 / **0.6419**
pdgfr_d23	40 / 40 / 40	**0.8556** / 0.8257 / 0.8186	**0.8230** / 0.8128 / 0.7976	**0.6996** / 0.6856 / 0.6610	**0.5269** / 0.5886 / 0.5779	**0.5760** / 0.5949 / 0.6148	**0.7507** / 0.7644 / 0.8282
pdl1	40 / 40 / 40	0.8872 / 0.8896 / **0.8906**	0.8709 / 0.8742 / **0.8784**	0.7814 / 0.7851 / **0.7923**	0.4035 / **0.3743** / 0.3827	0.4171 / **0.3918** / 0.3929	0.5097 / 0.4930 / **0.4927**
tie2	40 / 40 / 40	**0.9106** / 0.8855 / 0.8740	**0.9089** / 0.8778 / 0.8601	**0.8143** / 0.7750 / 0.7436	**0.5184** / 0.5719 / 0.6442	**0.5271** / 0.5899 / 0.6589	**0.7257** / 0.8026 / 0.8961
trka	40 / 40 / 40	0.8701 / **0.8975** / 0.8496	**0.7971** / 0.7916 / 0.7656	0.5127 / **0.5324** / 0.5012	0.4366 / **0.3665** / 0.4440	0.5177 / **0.5031** / 0.5493	0.8798 / **0.8255** / 0.8619

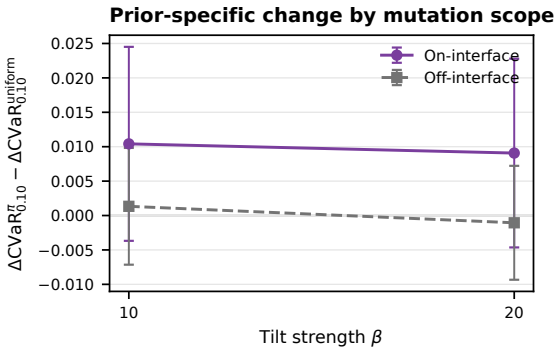


Figure 6: **Prior-specific changes differ by mutation scope.** Points show the equal-family mean difference-in-differences in worst-decile interface iPTM: $[\mathrm{CVaR}_{\pi}(\beta) - \mathrm{CVaR}_{\pi}(0)] - [\mathrm{CVaR}_{\mathrm{uniform}}(\beta) - \mathrm{CVaR}_{\mathrm{uniform}}(0)]$, computed separately for substitutions on and off each design's modeled interface. Positive values indicate that tilting improves the π arm more than the uniform arm. Whiskers are 95% Student-t intervals across the 15-family complete-on-interface cohort; all 40 selected designs contribute in each condition.

Table 7: **Runtime of exact tilted decoding.** Values are median seconds per sampled design [interquartile range] over 42 calls per method. The panel contains two distinct BoltzGen backbones from each of seven target families, three seeds per backbone, one sampled sequence per call ($B = 1$), and one timed repeat. Complete sampler calls include structural encoding but exclude CIF parsing, feature construction, and decode-order generation. All methods for a backbone were run on the same one of eight NVIDIA RTX A6000 GPUs, with method order position-balanced across backbone–seed pairs. Ratios are ratios of pooled medians.

Method	Seconds [IQR]	Runtime / vanilla	Speedup vs. uncached
Vanilla ProteinMPNN	0.553 [0.519, 0.697]	1.00×	—
Original exact (uncached)	114.014 [52.455, 122.648]	206.34×	1.00×
Full cached DP	3.593 [2.883, 4.065]	6.50×	31.74×
Local cached DP	0.776 [0.686, 0.915]	1.40×	147.01×

C CACHED LEAVE-ONE-OUT TILTED DECODING

Complexity and terminology. Let $n = |V|$, $d = |\mathcal{D}|$, $A = |\mathcal{A}|$, and let U be the cost of one decoder-layer update and L the number of layers. Ordinary, untilted ProteinMPNN decoding costs $\Theta(nLU)$. The original exact implementation recomputes the complete decoder separately for every design position, amino-acid candidate, and target query, at cost $\Theta(dAQnLU)$. The cached algorithm instead costs:

$$\Theta([n(Q+1) + dAQ]LU)\,, \tag{4}$$

where the first term maintains the $Q + 1$ worlds and the second performs the candidate readouts.

The implementation further uses the locality of ProteinMPNN's neighbor graph. At a design position i, only target queries j whose neighbor list contains i can vary across candidate identities; every other query adds the same constant to all candidate logits and therefore cancels in the proposal softmax. If $\bar{Q}_{\mathrm{loc}}$ is the mean number of such candidate-dependent queries, the production local-readout algorithm costs

$$\Theta\big([n(Q+1) + dA\bar{Q}_{\mathrm{loc}}]LU\big)\,. \tag{5}$$

The accompanying GPU benchmark measures all four paths on a deterministic panel of two distinct generated backbones from each of seven target families.

Empirical runtime. We measured complete sampler calls, including structural encoding but excluding CIF parsing, feature construction, and decode-order generation, on eight identical NVIDIA RTX A6000 GPUs. Each of the 14 backbones was evaluated with three random seeds and one sampled design per call ($B = 1$), giving 42 calls per method. All four methods for a given backbone were assigned to the same GPU, and method order was position-balanced across backbone–seed pairs. All 42 comparisons among the three exact tilted implementations produced identical binder sequences; the maximum absolute difference in proposal probability was 1.61×10^{-6}. Table 7 reports the resulting medians and interquartile ranges.

The production local method was therefore $147\times$ faster than the original exact implementation while requiring only $1.40\times$ the wall-clock time of vanilla ProteinMPNN; locality provided a further $4.63\times$ speedup over the full cached method. Exact reconstruction of the neighbor graphs gave a median interface size of $Q = 30$ and a median per-backbone $\bar{Q}_{\mathrm{loc}} = 4.44$ (range 2.36–8.89), so only about 15% of interface queries were candidate-dependent at a typical binder position. The resulting operation counts predict a median $4.06\times$ reduction from the full to the local cached method, close to the observed $4.63\times$ wall-clock gain. The larger operation-count improvements—$200\times$ from uncached to full caching and $867\times$ from uncached to local caching at the median backbone—exceed the respective $31.7\times$ and $147\times$ wall-clock gains, as expected when nominal operations are executed concurrently in GPU batches.

D KL DECOMPOSITION AND CONDITIONAL-MUTANT CONTROLS

Let w_j be the WT residue at interface site j, let π_j be the supplied target distribution, and let $P_{\beta,j}$ be the inverse-folding posterior under designs sampled at tilt strength β. We focus on sites where π_j assigns nonzero probability to at least one substitution.

To separate preferences among substitutions from the overall WT probability, we renormalize the posterior over the 19 non-WT amino acids:

$$P^{\mathrm{mut}}_{\beta,j}(a) = \frac{P_{\beta,j}(a)}{1 - P_{\beta,j}(w_j)}, \qquad a \neq w_j. \tag{6}$$

We then measure how much tilting moves this distribution toward a reference distribution q_j:

$$G_{q,j}(\beta) = \sum_{a \neq w_j} q_j(a) \log \frac{P^{\mathrm{mut}}_{\beta,j}(a)}{P^{\mathrm{mut}}_{0,j}(a)}. \tag{7}$$

Equivalently, $G_{q,j}(\beta) = \mathrm{KL}(q_j \| P^{\mathrm{mut}}_{0,j}) - \mathrm{KL}(q_j \| P^{\mathrm{mut}}_{\beta,j})$, so positive $G_{q,j}$ means that the tilted posterior is closer to q_j than the untilted posterior. We consider three reference distributions: the supplied prior conditioned on non-WT residues; a uniform distribution over the 19 substitutions; and an anti-prior proportional to $1 - \pi_{\mathrm{dense},j}$ and normalized over the non-WT residues, where π_{dense} is the untruncated ESM prior of Section 4. Comparing the prior and uniform results tests whether tilting produces prior-directed redistribution rather than generic flattening.