CONFIDENCE IS NOT COMPETENCE: INVERSE-FOLDING MODELS KNOW BINDING IN THE MIXED DERIVATIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Staged binder design generates a backbone and then inverse-folds a sequence, where current approaches read the model's *confidence* to judge interface choices. Recovery-based evaluations report a hotspot deficit taken as evidence that inverse-folding models cannot place binding-critical residues. We show this metric is misaligned with binding competence: while confidence evaluates the likelihood within a single bound state, competence requires comparing likelihoods across bound and unbound states, making confidence the wrong proxy for binding. Confidence is a within-state scalar of the bound-conditioned distribution P. Binding *leverage* is the *mixed second derivative*, the response to ablating the partner, which has the form of the model's **classifier-free-guidance direction**. Leverage is not determined by P, so every scalar of confidence is blind to it **by construction**. On SKEMPI natural complexes, confidence ranks hotspots near chance, whereas the mixed derivative, a zero-shot readout, tracks experimental binding free energy (Spearman -0.30 pooled, -0.21 per interface) and adds signal beyond geometry, conservation, and the one-pass log-odds. The effect holds across four architectures and beyond a fitted physics function (FoldX). But the advantage is regime-specific: on de-novo designed binders confidence overtakes it (interface AUROC 0.61 versus 0.57). Nonetheless, the signal is actionable: a small logit tilt on a frozen ProteinMPNN raises binding-favorability. The tilt beats a random tilt broadly, and a same-magnitude confidence tilt out of family on FoldX (best-of-k, lower CI $+0.72$ kcal/mol), while ipTM measures foldability rather than binding. Separately, the reported deficit is itself largely a burial and composition confound. The recipe is not protein-specific: to read an untrained quantity from a conditional generative model, ablate its conditioner and take the mixed derivative. We test it only on proteins.

1 INTRODUCTION

The dominant paradigm for designing protein binders is staged: a generative model proposes a backbone, and an inverse-folding model assigns a sequence by maximising $p(\text{sequence} \mid \text{backbone})$. This factorisation is convenient, but it omits the very quantity a binder exists to optimise. There is no binding-energy term anywhere in $p(\text{sequence} \mid \text{backbone})$. The concern this raises is sharpest at *interface hotspots*: the small set of residues that contribute most of the binding free energy, and which are frequently *frustrated*, buried polar residues or strained rotamers that are locally unfavourable for the monomer's fold yet selected because they contribute disproportionately to binding. On this reasoning, a model that optimises fold-compatibility should systematically miss exactly these residues, and a hotspot should sit in the tail of the model's distribution rather than at its mode. A prominent measurement appears to confirm this: ProBID-Net (Chen et al., 2024) reports inverse-folding recovery of 0.334 at hotspots versus 0.472 elsewhere, and suggests the gap may reflect dynamics.

This paper asks one question: when an inverse-folding model is used to build a binder, where, if anywhere, does its knowledge of *binding* live, and can a designer read it and act on it? The scope is deliberately narrow and practical: structure-conditioned inverse-folding models (ProteinMPNN, ESM-IF1, and three others), protein-protein interface binding, and readouts computable at design time. Our answer is that the binding signal is present, but the field has been reading it from the

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.

wrong place. It has read summaries of the model's output that miss binding **by construction** (the scalars of P: recovery and confidence) or by contraction (the two-pass complex-vs-monomer KL, which collapses the binding vector to a single number), when it lives in a **mixed second derivative** that is cheap to compute and directly actionable.

The consequence is immediate for practice, in three forms. As a **diagnostic**: inverse-folding confidence is a geometry detector, not a binding score, so a pipeline that distrusts it at the interface (as BindCraft (Pacesa et al., 2025) does, by freezing the interface) is justified in doing so, and can now do better. As a **training-free ranker**: the mixed derivative locates interface hotspots and adds on top of the exact feature set published predictors already use, at no training cost. As a **drop-in steering knob**: a $+\alpha \cdot L$ tilt biases a *frozen, off-the-shelf* inverse-folding model toward higher-binding interface residues with no retraining, corroborated out of family by a physics energy function, with a correlated cross-architecture judge a consistency check and ipTM a foldability check. The recipe is not protein-specific: the mixed derivative has the form of the model's **classifier-free-guidance direction**, so the same move applies to any conditional generative model.

There is a confound that no prior analysis controls, and it runs opposite to intuition. Hotspots are, on average, *more deeply buried* than other interface residues, and burial is precisely where inverse folding is *most* confident and accurate. An uncontrolled hotspot-vs-rest comparison therefore mixes a putative binding effect with a large burial effect of the opposite sign, so the naive comparison **hides** any real deficit rather than inventing one. Controlling for burial is therefore central to the design, not a detail.

Our contribution is conceptual and, to our knowledge, new. We identify binding as a specific functional of an inverse-folding model, the mixed derivative, show that no confidence readout can recover it, and establish it with the first control that survives beyond geometry and beyond evolutionary conservation (the full feature set published hotspot predictors use), a feature-class law across four architectures, and frozen-model steering whose binding-specificity rests out of family on a physics energy function. The operator itself is not new: the partner-ablated second forward pass is the inner difference of the BA-Cycle operator of Jiao et al. (2025), and decoders such as RedNet (Xie & Xu, 2026) tilt along the same direction. Our advance is what the operator means and what it delivers, which we position against this prior work in full below.

A single principle runs through all of this (Figure 1): confidence is not competence at every level of the stack. An inverse-folding model's confidence is blind to binding by construction, and one level up a structure predictor's ipTM is insensitive to binding-*specificity* (Section 5.1). The competence lives in the response to ablating the partner, whichever model does the reading.

Contributions.

- **A decomposition and a measured non-identifiability.** Confidence is a within-state scalar of the inverse-folding log-likelihood, leverage is the mixed second derivative, and leverage is not determined by the bound distribution: a flexible learner over all of P explains only about 37% of the variance in $|L|$ out of sample in both families.
- **A feature-class law.** On natural complexes (SKEMPI) every scalar summary (confidence and negentropy of the bound distribution, and even the two-pass scalar KL) does not clear a calibrated placebo floor, while the mixed derivative adds binding information beyond geometry, beyond evolutionary conservation, and beyond the one-pass log-odds, across four architectures.
- **The published deficit is largely a burial and composition confound.** A regression over 701 hotspots puts the crystal-backbone gap at $+0.059$ $[-0.051, +0.167]$, excluding both the -0.113 we measure for ProBID-Net's released model and the -0.138 its published rates imply. A matched-pair design finds no deficit under matching across all five architectures but lacks the resolution to test it.
- **Steering a frozen model.** A $+\alpha \cdot L$ tilt on a frozen ProteinMPNN raises binding-favorability, corroborated out of family by FoldX, with the four inverse-folding judges a correlated consistency check and ipTM a foldability check, on crystal and predicted backbones.

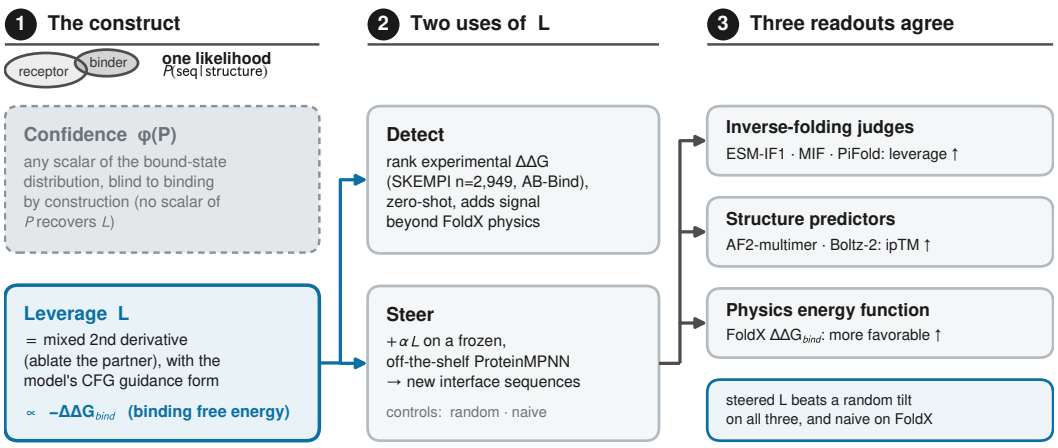


Figure 1: **Confidence is not competence.** Confidence and competence are two readings of one inverse-folding likelihood, one within a conditioning state and one across two. Confidence is a within-state scalar of the bound distribution and measures fold-stability constraint. Leverage is the mixed second difference, the response to ablating the partner, and has the model's classifier-free-guidance form, so it measures binding. The field has read the within-state scalar, where binding does not live. The pattern recurs one level up: a structure predictor's ipTM reads foldability, not binding.

- **Generality.** The insensitivity extends from binding to catalytic residues, and the recipe, ablating the conditioner and taking the mixed derivative, has the form of the model's classifier-free-guidance direction, which we conjecture transfers to any conditional generative model.

2 SETUP

Fixtures. Our primary fixture is SKEMPI 2.0 (Jankauskaite et al., 2019), from which we take single-mutation binding data and define a hotspot by alanine-scan $\Delta\Delta G_{\text{bind}}$, using the destabilizing convention $\Delta\Delta G_{\text{bind}} = \Delta G_{\text{mut}} - \Delta G_{\text{wt}} > 0$, with a null set of $|\Delta\Delta G| < 0.25$. The labeled analyses use the strict threshold > 2 kcal/mol (ProBID-Net's threshold, 327/13,401 interface positions, 2.4%, label entropy 0.115 nats), and a loose tier > 1 (701 interface hotspots) powers the regression estimator. This sets the scale for the effect sizes below. Because SKEMPI complexes are crystal structures of *natural* complexes, we add two fixtures of different character. The Cao-2022 de-novo designed binders (Cao et al., 2022) carry site-saturation binding measurements (73 designs in our processing of their supplementary data), a true design-regime test. Bennett et al. (2023) separately used Cao's large binder-versus-nonbinder screen, not the site-saturation set, to retrospectively evaluate deep-learning binder design. AB-Bind (Sirin et al., 2016) gives antibody-antigen $\Delta\Delta G$ over 27 analysed complexes, a second SKEMPI-class fixture with distinct biophysics, though it overlaps SKEMPI substantially (Section 9).

The matched-pair protocol, and leakage. The burial analysis uses within-complex optimal 1:1 matching of hotspots to null residues on relative complex SASA (± 0.05), secondary-structure class, and neighbour count (± 1), with all matching and cross-validation at the complex level, so a leak through near-identical interfaces cannot inflate a result. A regression estimator over all hotspots complements the matched design. Matched-pair retention and the absence of any survivorship bias toward exposed hotspots are in Appendix C.2.

Statistical stance. We fixed the headline falsifiers and analysis plan before computing them, so their tests are confirmatory. The conditional-predictive-impact (CPI) estimator and its calibrated placebo floor are defined in Appendix C: a feature counts only if its CPI lower bound clears the

floor. We write $\mathrm{CPI}_{\mathrm{pos}}$ for the per-interface-position aggregation and $\mathrm{CPI}_{\mathrm{mut}}$ for the per-mutation one (reconciled in Appendix C.3). Every interval is a complex-clustered bootstrap with a fixed seed (20260803), and the secondary and cross-architecture analyses carry no family-wise correction, as descriptive support.

Models. Five inverse-folding architectures span the design space: ProteinMPNN (Dauparas et al., 2022) (vanilla and soluble), ESM-IF1 (Hsu et al., 2022) (a 142M-parameter GVP-transformer), PiFold (Gao et al., 2023) (a one-shot GNN), MIF (Yang et al., 2023) (masked inverse folding), and ProBID-Net (Chen et al., 2024) (a voxel CNN). A positive control gates every scoring path (each pipeline reproduces a committed reference score before its outputs are trusted). These are inverse-folding *sequence* models, the stage our critique targets. De-novo backbone generators such as RFdiffusion (Watson et al., 2023) sit upstream and are not steerable by $+\alpha \cdot L$. The predicted backbones we study (Section 7) come from structure predictors of natural complexes, not from such a generator.

What the evaluation answers. Four questions organise the empirical sections: whether any scalar read off an inverse-folding model locates hotspots beyond cheap geometry (Section 3), where binding knowledge lives if not (Section 4), whether it is actionable and independently corroborated (Section 5), and whether it survives predicted backbones or is a burial confound (Sections 6 and 7).

3 Confidence is not competence

The quantity a practitioner is most tempted to trust, the model's own confidence that the native residue belongs at a position, does not locate hotspots. On SKEMPI interface positions ProteinMPNN's per-residue confidence attains a pooled AUROC of 0.538 for hotspots, barely above chance. Figure 2 shows the dissociation on barnase-barstar, the canonical hotspot system, where confidence ranks the alanine-scan hotspots at AUROC 0.31 while the partner-ablation mixed derivative (Section 4) ranks them at 0.77.

One might object that any single scalar can be rescued by combining it with structure. It cannot. A CPI test that holds cheap geometry (burial, neighbour count, ΔSASA) fixed finds confidence adds nothing: CPI 0.000 $[-0.0003, +0.0003]$ (all intervals 95% complex-clustered bootstraps), conditionally independent of hotspot-ness given structure, while ΔSASA, tested against burial and neighbour count alone, adds real information (CPI $+0.013$).

This is not specific to one network. Across all five architectures interface-hotspot confidence-AUROC lies in 0.50 to 0.54, each well below trivial burial, while free geometry predicts hotspots (burial 0.689, ΔSASA 0.673, a cheap-geometry combination 0.734) and the learned sequence-free KL detector captures only a small signal (CPI $+0.002$). None of this is SKEMPI-specific: on AB-Bind the identical pattern holds (confidence-AUROC 0.560, CPI spanning zero).

4 Detection: the confidence–leverage decomposition

Every quantity the field reads off these models (recovery, confidence, the KL) is a scalar summary of the distribution, and there is a specific reason they all fail at hotspots. This section builds one arc: the decomposition and its no-go, the feature-class law that only the mixed derivative clears, and the readout's fragility and reach.

4.1 The decomposition and the no-go

Decompose the model's per-position signal into two readings of the inverse-folding log-likelihood, one *within* a single conditioning state and one *across* two. **Confidence** is a within-state scalar functional of the single bound-conditioned distribution $p(\cdot \mid X_{\mathrm{complex}})$ ($\log p$(native), negentropy), and estimates fold-stability constraint. **Leverage** is the *mixed second difference* across the bound and partner-ablated states, the per-substitution response to ablating the partner:

$$L_i(a) = [\log p(a \mid X_{\mathrm{complex}}) - \log p(\mathrm{wt} \mid X_{\mathrm{complex}})] - [\log p(a \mid X_{\mathrm{monomer}}) - \log p(\mathrm{wt} \mid X_{\mathrm{monomer}})],$$

which by the standard binding thermodynamic cycle estimates $-\Delta\Delta G_{\mathrm{bind}}$ up to a single unknown positive scale. This has the algebraic form of the classifier-free-guidance direction (Ho & Salimans,

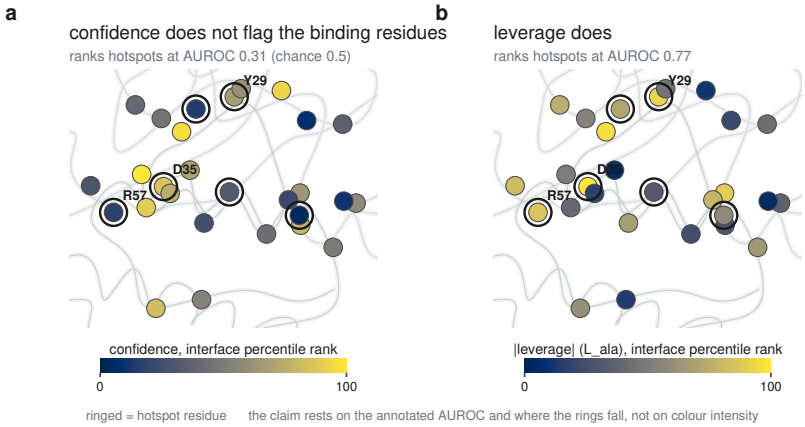


Figure 2: **Confidence does not flag binding, leverage does**, on barnase-barstar (1BRS). Interface residues are coloured by percentile rank on one shared scale. Confidence (**a**) ranks the hotspots (rings) at AUROC 0.31, the mixed derivative (**b**) at 0.77. This $+0.46$ gap exceeds that of 89% of the 108 SKEMPI hotspot complexes (median $+0.18$), an upper-tail case shown in context, not a cherry-pick.

2022): a guided logit $\text{logit}(\cdot \mid X_{\text{complex}}) + \alpha \cdot [\text{logit}(\cdot \mid X_{\text{complex}}) - \text{logit}(\cdot \mid X_{\text{monomer}})]$ has bracket equal to L up to a reference shift the softmax absorbs. It is an analogy, not an identity: the ablated state is a real partner-deleted structure, not a trained null, so it is not guaranteed in-distribution. The binding partner is the conditioner and confidence is the conditional marginal, so the recipe for reading an un-trained quantity off a conditional generative model is to ablate the conditioner and take this mixed derivative, not the marginal (Figure 2 shows the two readouts on a real interface).

> **Proposition 1 (Confidence is blind to leverage).** Fix a model and an interface position i, with $P_i = p(\cdot \mid X_{\text{complex}})$ and $Q_i = p(\cdot \mid X_{\text{monomer}})$. **(i, cycle)** under a shared-temperature energy readout and a rigid partner ablation (Appendix C), $L_i(a)$ estimates $-\Delta\Delta G_{\text{bind}}(i, \text{wt} \to a)$ up to one positive scale. **(ii, non-identifiability, and a separate floor)** L is not a function of P: two positions with the same bound distribution and native residue share every confidence functional $\varphi(P_i, \text{wt}_i) = \varphi(P_j, \text{wt}_j)$ (confidence, negentropy, recovery all depend on (P, wt)), yet can differ in leverage, so no confidence readout recovers L. This does not by itself create a binding floor: by the minimum-mean-squared-error identity the best P-only risk is $\inf_\varphi \mathbb{E}[(\Delta\Delta G_{\text{bind}} - \varphi(P))^2] = \mathbb{E}[\text{Var}(\Delta\Delta G_{\text{bind}} \mid P)]$, strictly positive as a separate measured fact (Section 4.1) that L partly recovers (Section 4.2). **(iii, readout, not model)** X_{monomer} is X_{complex} with the partner deleted, a deterministic map, so L is computable by a second forward pass yet by (ii) is not a function of P. The binding term is in the weights, reachable only by re-querying on X_{monomer}.

The no-go is not merely formal, and we measure it. A flexible learner (the better of gradient boosting and random forests, out-of-sample under complex-clustered CV) trained on the full 20-vector P explains only $R^2(|L|_{\text{rms}} \mid P) = 0.37$ $[0.34, 0.40]$ of the variance for ProteinMPNN and 0.36 $[0.34, 0.39]$ for ESM-IF1 (Figure 3), so these learners leave 63 to 64% of the leverage-magnitude variance unexplained in either family (adding the native identity to their inputs moves this by under 0.01). The same measurement rules out the second pass being a fixed reparameterisation of confidence (which would force $R^2(|L|_{\text{rms}} \mid P)$ to one) and lower-bounds what a P-only readout can recover, so the unexplained share upper-bounds the irreducible one and non-identifiability is a measured effect. Against experimental $\Delta\Delta G_{\text{bind}}$, P alone explains 12%, L alone 9%, and together 18%: the mixed derivative adds a $+0.06$ increment P does not carry (`r2_ddg_increment.csv`).

Table 1: The feature-class law on SKEMPI (per interface position, CPI_{pos} beyond burial, neighbours, and ΔSASA, in units of 10^{-3}, 13,401 positions). Every per-position aggregate scalar of P fails to clear the calibrated placebo floor, including the two-pass scalar KL. Only the two-pass mixed derivative, a vector contrast, clears it. The one-pass log-odds is a per-substitution readout rather than a per-position scalar summary, with real signal, and is compared against L in Section 4.2 and Figure 15.

Feature (scalar summaries vs. the mixed derivative)	CPI_{pos} beyond geometry ($\times 10^{-3}$)
Placebo floor (a deterministic function of geometry)	$+0.7$ (false-positive floor)
Confidence (within-state scalar)	$+0.2$ $[-0.2, +0.6]$ (below floor)
Negentropy (one-pass)	$+0.9$ $[+0.3, +1.5]$ (at floor)
Scalar KL (two-pass, a contraction of L)	$+1.0$ $[+0.3, +1.7]$ (at floor)
Leverage L (the mixed derivative)	$+\mathbf{4.9}$ $[+3.4, +6.5]$ (about $7\times$ floor)

4.2 The feature-class law on natural complexes

The feature-class law is the central empirical claim of this section: whether a readout carries binding information beyond geometry is set by its *class*, not its sophistication, so every per-position scalar summary of the bound distribution fails while the two-pass mixed derivative, a vector contrast, succeeds. We test it position by position on SKEMPI. On SKEMPI, the mixed derivative adds binding information beyond cheap geometry where every scalar summary does not (Table 1 and Figure 4, per interface position, CPI over burial, neighbours, and ΔSASA). At the mutation level the effect is large: $\text{Spearman}(L, \Delta\Delta G) = -0.30$ and $\text{CPI}_{\text{mut}}(L \mid \text{geometry}) = +0.059$ $[+0.046, +0.073]$. The second pass earns its keep against the standard one-pass log-odds, itself strong (marginal -0.29, near L's -0.30): controlling for it and geometry, leverage still tracks $\Delta\Delta G$ (partial Spearman -0.147 $[-0.190, -0.108]$), and the one-pass readout retains signal after controlling for leverage (-0.094, `w2_onepass_control.csv`), so neither pass subsumes the other. The increment is not conservation in disguise. We score each position with the field-standard ESM-2 (Lin et al., 2023) masked-marginal estimator (Meier et al., 2021), itself the stronger baseline (adding $+0.0064$ beyond geometry), and leverage's contribution is undiminished: $\text{CPI}_{\text{pos}}(L \mid \text{geometry} + \text{conservation}) = +0.0059$ $[+0.0031, +0.0097]$ (per interface position, the scale of Table 1, so the mutation-level $+0.059$ above differs only by aggregation). And as a zero-shot readout of a model never trained on binding, L adds $+0.030$ interface AUROC on top of a *supervised* geometry-plus-substitution baseline. On per-interface SKEMPI Spearman our zero-shot L is about 0.21 (`reviewer_round2.csv`), far below folding-energy and supervised methods: Deng et al. (2025) report StaB-ddG at 0.45 and its zero-shot variant at 0.43 on an 80-complex homology-clustered held-out split under their own filtering, not commensurable with our all-SKEMPI evaluation. We scope detection to the zero-shot regime, where our object is the decomposition and controls, not a better $\Delta\Delta G$ score.

All comparators run under identical settings, so the contrasts isolate the feature, not the protocol. The law holds in all four architectures: beyond ProteinMPNN, ESM-IF1 (Spearman -0.26, CPI $+0.035$), PiFold (-0.33, $+0.050$), and MIF (-0.27, $+0.058$), each surviving substitution, confidence, and scalar-KL controls, with confidence at chance throughout. The law is about a feature class, not a regime (Figure 10): on natural complexes scalar summaries reduce to geometry (confidence exactly, KL nearly), while the mixed derivative does not.

4.3 A backbone-error dose law, and two further readouts

The mixed derivative is read off a backbone and does not survive a large perturbation of one. Jittering the crystal backbone to a target interface RMSD and re-scoring (the monomer inheriting the same jitter, so the partner ablation stays clean), $\text{CPI}(L \mid \text{geometry})$ and $\text{Spearman}(L, \Delta\Delta G)$ hold through sub-Ångström error, then collapse past a model-dependent knee (about 1.0 Å for ProteinMPNN, 1.5 Å for ESM-IF1, Table 3 and Figure 5), not a universal cliff. The decisive test is which side real predicted backbones fall on, and they fall on the surviving side (Section 7): on OpenFold3 (The OpenFold3 Team, 2026; Abramson et al., 2024) and AF2-multimer (Evans et al., 2021; Jumper et al., 2021) backbones for 140 shared complexes, $\text{CPI}_{\text{mut}}(L \mid \text{geometry})$ is $+0.039$ $[+0.027, +0.050]$ and $+0.032$ $[+0.023, +0.041]$, 69 to 84% of the crystal CPI_{mut} on the same 140 complexes. Because

the collapse is driven by the backbone the derivative is read from, every method built on this mixed derivative inherits the same fragility on predicted backbones.

The signal also reaches the second mixed derivative (Figure 6): the partner-ablated pairwise coupling tracks measured binding epistasis (Spearman$(C, g) = -0.14$), and it carries into the de-novo regime, though the scalar leads there (Section 9, Appendix C).

5 STEERING: ACTING ON THE MIXED DERIVATIVE

Detection located the binding signal. Here we act on it, steering a frozen model, adjudicating against physics, and testing the predicted backbones a designer actually has. All of this is the natural-complex regime. On de-novo co-designed binders confidence overtakes the derivative, so the design regime itself is scoped out here and treated in Section 9.

5.1 STEERING A FROZEN MODEL

If L is the guidance direction, the direct test is to guide with it. We bias a frozen, off-the-shelf ProteinMPNN's interface logits by a static $+\alpha L$ term and sample (K=64, 271 complexes, over a fixed grid). This raises, monotonically, the mean binding-leverage a *different* model (ESM-IF1) assigns the sampled residues, from -0.20 at $\alpha = 0$ to $+0.27$ at $\alpha = 2$, while a matched-magnitude random direction does the opposite (paired $L -$ random reaches $+0.77$). The benefit is carried by the direction, not the perturbation. The steered sequences also transfer to an independent structure predictor: folded with AF2-multimer (120 complexes), the L-steered interfaces beat the random control on ipTM by $+0.235$ $[+0.199, +0.272]$, interface-localized (global pTM rises about $2.9\times$ less). The benefit reproduces on a second structure predictor (Boltz-2 (Passaro et al., 2025)), a second steered model judged by a third, and four cross-architecture judges (self-judged pairs excluded), all in Table 4 and Figure 7.

A specificity control bounds the interpretation. A *naive* tilt, the model's own log-odds direction rescaled to L's magnitude (be more confident, not bind better), recovers most of the ipTM gain and beats L there ($L -$ naive $= -0.034$ $[-0.064, -0.007]$), because ipTM is itself a structure-level confidence rewarding a foldable interface, the confidence-is-not-competence split one level up. The binding-specific increment survives where the mechanism predicts, at the inverse-folding-judge level (a correlated cross-architecture check, Appendix C.2), where L beats the confidence tilt (ESM-IF1 $+0.159$ $[+0.110, +0.207]$, MIF $+0.242$ $[+0.191, +0.292]$). The confidence tilt is thus a strong but binding-*correlated* baseline (64 to 77% of $L -$ random at the judge level), not binding-*specific*. In the classifier-free-guidance framing naive is the unguided marginal and L the guidance term, so naive, not random, is the demanded control.

A second control rules out the last trivial explanation, that the tilt merely restores the native residue. We add a native-reversion arm (a one-hot tilt toward wild type) and restrict every contrast to non-native positions, where reversion has nowhere to go (falsifier fixed in advance: judge-level $L -$ wttilt ≤ 0). On the independent ESM-IF1 judge L beats native reversion by $+0.39$ $[+0.26, +0.51]$ and the naive tilt by $+0.21$ $[+0.16, +0.27]$ (both $P(> 0) = 1.0$, 60 complexes), and FoldX agrees in direction at $+0.06$ to $+0.16$ kcal/mol, far smaller than the unrestricted Section 5.2 contrast. The tilt stays 64 of 64 distinct (Figure 8, and $L' = L$ bit-for-bit confirms the softmax-invariance of L (Section 4.1, Appendix C.1).

5.2 PHYSICS ADJUDICATION (FOLDX)

To close the gap between inverse-folding judges and a real binding term, we ran FoldX (Guerois et al., 2002; Schymkowitz et al., 2005; Delgado et al., 2019), a hand-fit physics energy function with an explicit binding term, (falsifier fixed in advance: $L -$ naive favorability ≤ 0). On the 60-complex steering set L beats the confidence tilt on physics: the design-relevant best-of-k $L -$ naive gain clears zero (Table 2), with the conservative mean-over-k in the appendix. Because FoldX is physics, not an inverse-folding proxy, it sides with the judges ($L >$ naive) against ipTM (naive $\gtrsim L$), so the binding-specific advantage of L is real, and the ipTM disagreement was the foldability effect the frustration thesis predicts. As a detector FoldX ranks better than the zero-shot L (Table 2). The narrower, more surprising claim is that L carries binding rank-signal beyond the fitted energy

Table 2: FoldX physics adjudication (out of family). Detection is Spearman vs experimental $\Delta\Delta G$ ($n = 2{,}948$ mutants, 284 complexes, both signs meaning "ranks binding correctly"). Steering specificity is the paired $L -$ naive favorability at $\alpha = 2$ (design-relevant best-of-k, 95% CI, $P(> 0)$, with the conservative mean-over-k and vs-random contrasts in Appendix A). The AB-Bind replication (Section 5.2) uses 22 of 27 complexes.

FoldX adjudication (out of family)	value
Detection, Spearman: FoldX (fit) vs L (zero-shot)	$+0.434$ vs -0.301
L beyond fitted FoldX (partial Spearman)	-0.17 $[-0.23, -0.12]$
$\mathbf{L} -$ naive steering, best-of-k (kcal/mol)	$\mathbf{+1.86}$ $[+0.72, +3.02]$, P=0.999

function (partial Spearman -0.17 $[-0.23, -0.12]$), replicated same-signed on AB-Bind (22 crystal complexes, underpowered).

5.3 Steering on predicted backbones of natural complexes

Every steering result so far tilts a crystal backbone, but staged design has only a predicted one. So we recompute leverage and apply the tilt on the predicted backbone (OpenFold3 and AF2-multimer, 106 complexes), judged anti-circularly by an independent model's leverage read off the predicted structure (falsifier fixed in advance: judge-level $L -$ random ≤ 0). The falsifier does not fire and the effect is barely attenuated: judge-level $L -$ random is $+0.70$ to $+0.76$ across both backbones and judges, matching the crystal control, and $L -$ naive stays positive ($+0.17$ to $+0.22$), so specificity is preserved. An out-of-family structure predictor (AF2-multimer) confirms it (interface ipTM $L -$ random $+0.18$ $[+0.14, +0.22]$ on 93 complexes, Figure 11). This is the predicted-backbone regime of *natural* complexes, not de-novo generation (Section 4.3).

6 On crystal backbones, the hotspot gap is a burial and composition confound

We now return to the published deficit and show that, on crystal backbones, it is a confound of burial and residue composition rather than a binding-specific failure. The confound is visible directly (Figure 12): as hotspot strength rises, recovery and burial rise together (recovery 0.347 to 0.529, relative SASA 0.218 to 0.080), so a raw hotspot-versus-rest comparison charges to hotspots what is really a burial effect.

We test the deficit two ways. The powered test is a regression over all 701 hotspots controlling for burial and composition, which puts the crystal-backbone gap at $+0.059$ $[-0.051, +0.167]$, excluding both the -0.113 deficit we estimate for ProBID-Net's released model and the -0.138 its reported rates imply. The matched-pair design agrees in direction but is underpowered (minimum detectable effects exceed the deficit), so it is a transparency check, not a test (Appendix C).

The result is consistent across models. No architecture retains a deficit under matching, most attenuate to zero, and two (MIF, PiFold) show a small positive surplus in the strict tier. The sharpest test uses ProBID-Net's own released voxel-CNN, which passes its positive control: the published deficit reproduces only in comprehensively-scanned complexes and dissolves once residue type is matched, so it is primarily a residue-composition confound (Appendix C).

7 The tax lives in the conditioning set

On crystal backbones the gap was a confound (Section 6). On the predicted backbones designers actually condition on, it returns (Figure 13). On backbones from two architecturally-independent predictors (OpenFold3 and AF2-multimer) a burial-matched deficit is present (-0.191 and -0.233, the crystal deficit about zero). The claim rests not on either number but on their agreement: the two predictors' per-complex deficits correlate at $\rho = 0.565$ $[0.40, 0.71]$ over 127 complexes, and it is not carried by a few outliers. The effect tracks how a backbone is non-native, not how far: on noised crystals at matched RMSD the deficit is absent, so it is a property of independent reconstruction.

Yet the binding signal a designer reads off these backbones survives: recomputed on the predicted structures, CPI($L \mid$ geometry) is $+0.039$ and $+0.032$ (pooled $+0.036$, $P(>0) > 0.999$), a 69 to 84% retention of the crystal CPI on the same complexes. The confidence readout degrades into the deficit while the mixed derivative survives, exactly where the decomposition says they must diverge. Two competing explanations for this residual are measured and refuted (Appendix C).

8 Related work and positioning

The nearest prior methods use this operator. The leverage operator L is the inner difference of BA-Cycle (Jiao et al., 2025), and RedNet (Xie & Xu, 2026), a preprint, independently operationalises a related contrast for selective binder redesign and a zero-shot $\Delta\Delta G$ score. We credit both, and our object differs: we **(i)** identify their score as the mixed derivative and confidence as the within-state scalar that cannot recover it, **(ii)** give the first beyond-geometry and beyond-conservation control for an inverse-folding binding signal (CPI (Watson & Wright, 2021) with a conditional permutation test (Berrett et al., 2020)), **(iii)** state the feature-class law, **(iv)** steer an off-the-shelf model with a drop-in tilt that preserves native recovery and trains nothing, and **(v)** validate anti-circularly with a confidence-tilt specificity control resting out-of-family on FoldX. Related work also predicts binding $\Delta\Delta G$ (StaB-ddG (Deng et al., 2025), USP-ddG (Yu et al., 2025)), but our object is the decomposition and steering, not a new predictor.

We also correct ProBID-Net's recovery-deficit attribution (Section 6): a burial and composition confound, not dynamics. On the beyond-X ladder (Figure 15) the mixed derivative survives geometry, substitution similarity, the one-pass log-odds, and a fitted physics function, while every per-position scalar of P stays at the placebo floor. Practice already leans on this split, BindCraft (Pacesa et al., 2025) withholding interface residues from ProteinMPNN redesign, and Janusz et al. (2026) independently report the same partner-insensitivity in antibody inverse folding.

9 Limitations

We evaluate on three primary fixtures (SKEMPI, Cao de-novo, AB-Bind), none a full wet-lab loop. The *detection* claim is experiment-anchored (L vs experimental $\Delta\Delta G$ at Spearman -0.30, beyond a fitted physics function at partial -0.17). L is conceptual, not a state-of-the-art $\Delta\Delta G$ predictor. The *steering* claim's specificity rests out-of-family on FoldX, strongest in the design-relevant best-of-k aggregation ($L -$ naive $+1.86$, lower CI $+0.72$ kcal/mol, the conservative mean-over-k in the appendix), pending a wet-lab test. The inverse-folding judges are a correlated model class (inter-model $\rho = 0.52$ to 0.64), a consistency check rather than independent evidence, and ipTM measures foldability, not binding.

The primary claims rest on one natural fixture, SKEMPI 2.0 (Appendix C), mitigated with two further fixtures, four architectures, ATLAS (Borrman et al., 2017), and catalytic residues (M-CSA (Ribeiro et al., 2018)), though only SKEMPI is powered for the geometry-controlled add-on and AB-Bind overlaps it (19 of 23 codes).

Three caveats bound the headline claims (more in Appendix C). Confidence cannot express leverage *by construction* (non-identifiability), which is not statistical independence. L estimates $-\Delta\Delta G_{\text{bind}}$ only up to a temperature, so every readout is rank-based. And the advantage is regime-specific: on de-novo co-designed binders confidence becomes the better readout (0.61 vs 0.57 interface AUROC). Natural complexes are the regime this paper targets, so this convergence bounds the readout's advantage rather than undercutting detection.

10 Conclusion

Confidence and competence are two readings of one inverse-folding likelihood, and the field has read the wrong one: every within-state scalar of the bound distribution is blind to the mixed derivative by construction, yet that derivative detects binding and steers a frozen model at no training cost. Shown here for proteins, the same recipe, ablate a conditioner and take the mixed derivative, should read an untrained quantity off any conditional generative model, and a wet-lab test of the steered designs is the natural next step.

REPRODUCIBILITY STATEMENT

Every numerical result traces to a committed script and CSV, each carrying its exact command and a fixed seed (20260803). A fresh pointer-only clone reproduces every headline number and figure from the committed CSVs alone, with no GPU, no Git-LFS pull, and no special hardware, which we verified directly. The analysis-ready data, the code, and this pointer-only reproduce script are available for review at an anonymized mirror, `https://anonymous.4open.science/r/confidence-is-not-competence-5DE2`. The $+\alpha L$ steering tilt is a drop-in reweighting of a frozen model's interface logits, released as a small reference implementation in that mirror. Re-running the GPU pipeline end to end (model inference, folding, and FoldX) additionally needs the model checkpoints, the FoldX binary, and the archived raw artifacts.

ETHICS STATEMENT

This work studies and steers existing, publicly released inverse-folding and structure-prediction models on public benchmark data (SKEMPI, AB-Bind, ATLAS, M-CSA, and released de-novo design measurements), and introduces no new experiments involving humans or animals. Its dual-use profile is that of the open models and public data it analyzes. The paper contributes a diagnostic, a training-free readout, and a bounded steering tilt. The tilt is a small reweighting of a frozen off-the-shelf model that reuses already-published partner-ablation operators (BA-Cycle, RedNet) rather than a new model or dataset, and its in-silico uplift is modest, so it does not meaningfully lower the barrier to designing harmful binders. All data are used under their original licenses, and the analysis-ready data and code are released to support scrutiny and reproduction.

AI USE STATEMENT

We designed the study, carried out the analysis, and independently verified all theoretical and empirical results (every number against committed code and data). Generative-AI tools assisted with literature search and reference checking, experiment and analysis implementation, and manuscript editing. All such output was reviewed and validated by the authors, who are solely responsible for the content.

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(8016):493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Nathaniel R. Bennett et al. Improving de novo protein binder design with deep learning. *Nature Communications*, 14(1):2625, 2023. doi: 10.1038/s41467-023-38328-5.

Thomas B. Berrett, Yi Wang, Rina Foygel Barber, and Richard J. Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B*, 82(1):175–197, 2020. doi: 10.1111/rssb.12340.

Tyler Borrman et al. Atlas: A database linking binding affinities with structures for wild-type and mutant tcr-pmhc complexes. *Proteins: Structure, Function, and Bioinformatics*, 85(5):908–916, 2017. doi: 10.1002/prot.25260.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *International Conference on Machine Learning (ICML)*, 2024.

Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M. Jude, Iva Marković, Rameshwar U. Kadam, Koen H. G. Verschueren, Kenneth Verstraete, Scott Thomas Russell Walsh, Nathaniel Bennett, Ashish Phal, Aerin Yang, Lisa Kozodoy, Michelle DeWitt, Lora Picton, Lauren Miller, Eva-Maria Strauch, Nicholas D. DeBouver, Allison

Pires, Asim K. Bera, Samer Halabiya, Bradley Hammerson, Wei Yang, Steffen Bernard, Lance Stewart, Ian A. Wilson, Hannele Ruohola-Baker, Joseph Schlessinger, Sangwon Lee, Savvas N. Savvides, K. Christopher Garcia, and David Baker. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910):551–560, 2022. doi: 10.1038/s41586-022-04654-9.

Zhihang Chen, Menglin Ji, Jie Qian, Zhe Zhang, Xiangying Zhang, Haotian Gao, Haojie Wang, Renxiao Wang, and Yifei Qi. Probid-net: a deep learning model for protein-protein binding interface design. *Chemical Science*, 15(47):19977–19990, 2024. doi: 10.1039/d4sc02233e.

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

Javier Delgado, Leandro G. Radusky, Damiano Cianferoni, and Luis Serrano. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics*, 35(20):4168–4169, 2019. doi: 10.1093/bioinformatics/btz184.

Arthur Deng, Karsten D. Householder, Fang Wu, K. Christopher Garcia, and Brian L. Trippe. Predicting mutational effects on protein binding from folding energy. In *International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 13129–13151, 2025. URL `https://proceedings.mlr.press/v267/deng25d.html`.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, and Demis Hassabis. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021. doi: 10.1101/2021.10.04.463034. Preprint, version 2 posted 10 March 2022.

Zhangyang Gao, Cheng Tan, Pablo Chacón, and Stan Z. Li. Pifold: Toward effective and efficient protein inverse folding. In *International Conference on Learning Representations (ICLR)*, 2023.

Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of Molecular Biology*, 320 (2):369–387, 2002. doi: 10.1016/S0022-2836(02)00442-4.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv*, 2022. doi: 10.48550/arXiv.2207.12598.

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 *International Conference on Machine Learning (ICML)*, 2022.

Justina Jankauskaite, Brian Jimenez-Garcia, Justas Dapkunas, Juan Fernandez-Recio, and Iain H. Moal. Skempi 2.0: an updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019. doi: 10.1093/bioinformatics/bty635.

Bartosz Janusz, Dawid Chomicz, Sonia Wrobel, Pawel Dudzic, Adithya Polasa, Kyle Martin, Steven Darnell, Stephen R. Comeau, and Konrad Krawczyk. Benchmarking antigen-aware inverse folding methods for antibody design. *Bioinformatics Advances*, 6(1):vbag081, 2026. doi: 10.1093/bioadv/vbag081.

Xiaoran Jiao, Weian Mao, Wengong Jin, Peiyuan Yang, Hao Chen, and Chunhua Shen. Boltzmann-aligned inverse folding model as a predictor of mutational effects on protein-protein interactions. In *International Conference on Learning Representations (ICLR)*, 2025.

John Jumper et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873): 583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Zeming Lin et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Martin Pacesa et al. One-shot design of functional protein binders with bindcraft. *Nature*, 646 (8084):483–492, 2025. doi: 10.1038/s41586-025-09429-6.

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

Antonio J. M. Ribeiro et al. Mechanism and catalytic site atlas (m-csa): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Research*, 46(D1):D618–D623, 2018. doi: 10.1093/nar/gkx1012.

Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The FoldX web server: an online force field. *Nucleic Acids Research*, 33(suppl_2):W382–W388, 2005. doi: 10.1093/nar/gki387.

Sarah Sirin, James R. Apgar, Eric M. Bennett, and Amy E. Keating. Ab-bind: Antibody binding mutational database for computational affinity predictions. *Protein Science*, 25(2):393–409, 2016. doi: 10.1002/pro.2829.

The OpenFold3 Team. OpenFold3-preview (version 0.5.0), August 2026. URL `https://github.com/aqlaboratory/openfold-3`. AlQuraishi Laboratory, Columbia University, and the OpenFold Consortium, research preview, concept DOI (latest version).

David S. Watson and Marvin N. Wright. Testing conditional independence in supervised learning algorithms. *Machine Learning*, 110(8):2107–2129, 2021. doi: 10.1007/s10994-021-06030-6.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Ziwei Xie and Jinbo Xu. Redesign selective protein binders using contrastive decoding. *bioRxiv*, 2026. doi: 10.64898/2026.05.09.722041.

Kevin K. Yang, Niccolò Zanichelli, and Hugh Yeh. Masked inverse folding with sequence transfer for protein representation learning. *Protein Engineering, Design and Selection*, 36:gzad015, 2023. doi: 10.1093/protein/gzad015.

Guanglei Yu, Xuehua Bi, Qichang Zhao, and Jianxin Wang. USP-ddG: a unified structural paradigm with data efficacy and mixture-of-experts for predicting mutational effects on protein-protein interactions. *bioRxiv*, 2025. doi: 10.1101/2025.11.09.687124.

A Extended figures and tables

This appendix collects the extended figures and tables in the order the argument uses them, so it reads as a guided extension rather than a bare reference list. Figures 3 and 4 give the decomposition and the feature-class law of Section 4: a learner explains only about 37% of the variance in the leverage from P, and only the two-pass mixed derivative clears the placebo floor, and Figure 2 shows the same dissociation on a real interface. Table 3 and Figures 5 and 6 establish robustness, the readout surviving sub-Ångström backbone error and reaching the partner-ablated pairwise coupling. Table 4 and Figure 7 give the steering matrix behind Section 5.1, and Figure 10 the regime dependence of the scalar. Figures 12 and 13 document the burial confound and the conditioning-set tax of Sections 6 and 7. Table 2 and Figures 14 and 15 give the physics adjudication of Section 5.2. A short note after each figure states, in prose, what it shows, how it relates to the main text, and how it was produced.

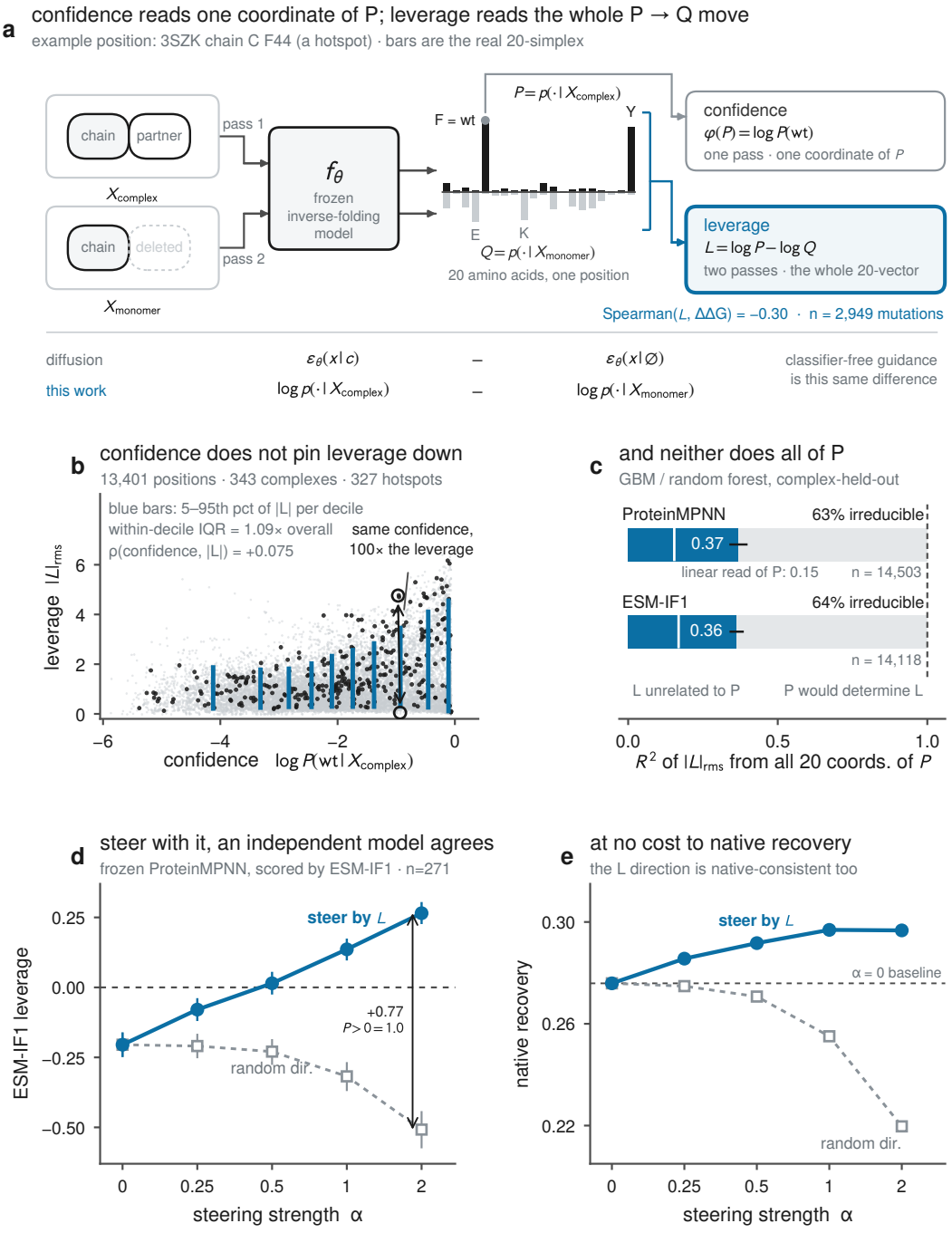


Figure 3: The Confidence-Leverage Decomposition. A flexible learner over the entire bound distribution recovers only about 37 percent of the variance in the leverage magnitude, so most is not recovered by these learners, in both inverse-folding families.

A.1 THE DECOMPOSITION AND THE FEATURE-CLASS LAW

Figure 3 in detail. This panel is the empirical content of the no-go in Section 4.1. For each interface position we take the model's full 20-dimensional bound-conditioned distribution P and ask how much of the leverage L a flexible learner can reconstruct from P alone. The learner is the better of gradient boosting and a random forest, scored out of sample under complex-clustered cross-validation so that no complex informs its own prediction, and it recovers only $R^2(|L|_{\mathrm{rms}} \mid P) \approx 0.37$ for ProteinMPNN and 0.36 for ESM-IF1. The gap between the identity line and the scatter is therefore the roughly 63 to 64% of the leverage-magnitude variance that these learners do not recover, and the exact statement is the vector-level non-identifiability of Proposition 1, shown here for both inverse-folding families.

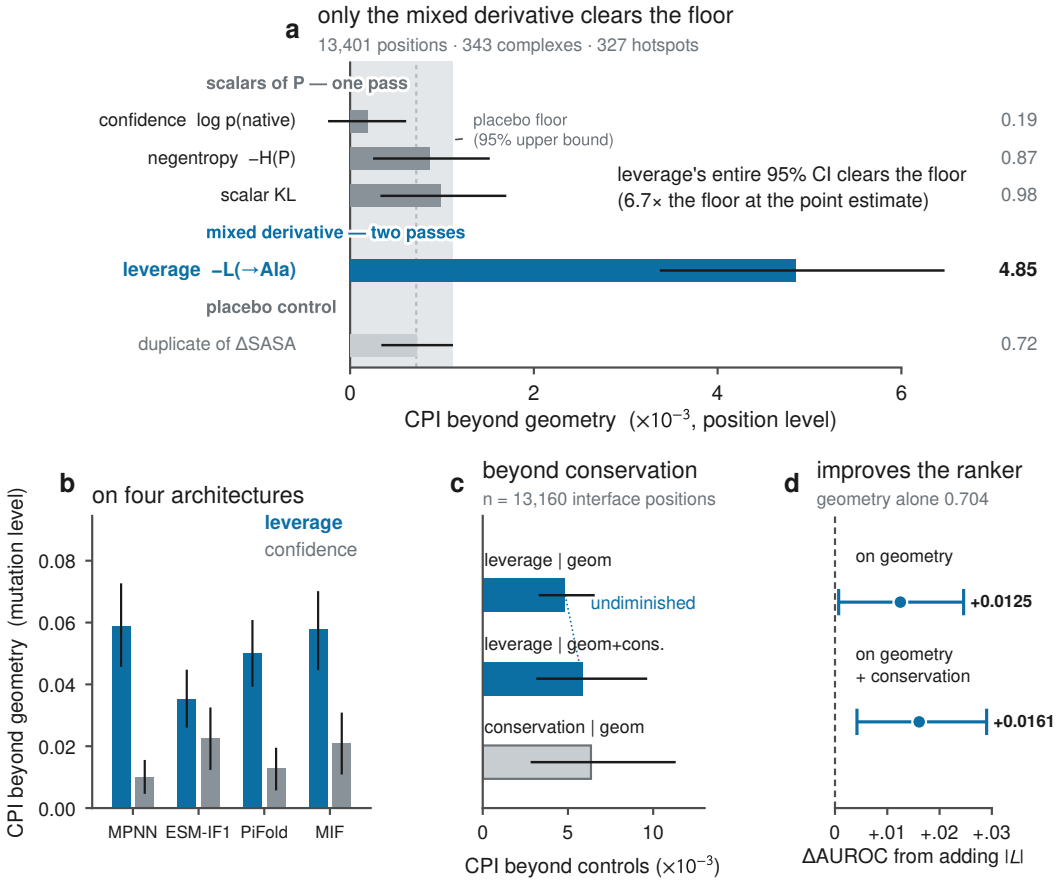


Figure 4: The feature-class law. **a**, every per-position aggregate scalar of the bound distribution (confidence, negentropy, the two-pass scalar KL) fails to clear the calibrated placebo floor, while the two-pass mixed derivative (leverage) clears it about sevenfold. **b**, the same split holds on four inverse-folding architectures. **c**, leverage's CPI is undiminished when masked evolutionary conservation is added to the geometry control, and leverage is nearly independent of conservation ($\rho(\text{conservation}, -L) = -0.08$ over 13,160 interface positions). **d**, adding $|L|$ improves a hotspot ranker over geometry ($+0.0125$) and over geometry plus conservation ($+0.0161$).

Figure 4 in detail. This is the visual form of the feature-class law of Section 4.2 and Table 1. Each bar is a feature's conditional predictive impact (CPI) over the cheap geometry baseline of burial, neighbour count, and ΔSASA, estimated per interface position on SKEMPI with a conditional permutation test, and the dashed line is the calibrated placebo floor, the same statistic computed for a deterministic function of geometry that carries no new information ($+0.0007$). Every per-position aggregate scalar of the distribution (confidence, negentropy, the two-pass scalar KL) fails to clear

that floor, while the mixed derivative clears it by about sevenfold. The figure makes concrete that the failure is a property of the whole scalar-summary class, not of any one readout.

Which leverage the hero (Figure 2) shows. Leverage is not one number per residue: it is the mixed derivative for each substitution, a vector over the amino acids. To colour a residue we aggregate to a scalar, and the hero uses L_{ala}, the leverage of the wild-type-to-alanine substitution. We choose it on principle, not to maximize the number: hotspots are *defined* by alanine scanning, so L_{ala} is the readout that matches the label a priori. It is not the aggregation that flatters us here, since the all-substitution RMS magnitude is comparable on this complex (interface AUROC 0.79, versus 0.77 for L_{ala}). L_{ala} asks how much binding is lost if the residue becomes alanine, aligned with the hotspot definition, whereas the RMS magnitude asks how much binding responds to any mutation, a more general but less aligned readout.

A.2 Backbone-error robustness and the second derivative

Table 3: Backbone-error dose law (crystal-jitter ladder). † marks a 95% CI excluding zero (above the $+0.0007$ placebo floor). $n = 2{,}949$ mutations (ProteinMPNN, matched sample), 2,809 (ESM-IF1, 285 complexes), seed 20260803. ProteinMPNN was run to 1.5 Å, past its $\sim$1.0 Å knee where the signal has already collapsed, and ESM-IF1 was extended to 2.0 Å to confirm the plateau, so the ProteinMPNN 2.0 Å cells are marked n/a. The $\sigma = 0$ crystal row is the jitter-ladder sample, and the Section 4.2 feature-class-law figures use the matched sample, so the two differ by rounding. The single-draw ProteinMPNN 1.0 Å cell ($+0.002$) is one realization near the knee, and the 10-draw multi-seed refinement in Appendix C gives $+0.009 \pm 0.005$ there.

Jitter σ (Å)	MPNN: CPI beyond geom	MPNN: $\rho(L, \Delta\Delta G)$	ESM-IF1: CPI beyond geom	ESM-IF1: $\rho(L, \Delta\Delta G)$
0.0 (crystal)	$+0.058^{\dagger}$	-0.30	$+0.036^{\dagger}$	-0.25
0.25	$+0.059^{\dagger}$	-0.29	$+0.035^{\dagger}$	-0.17
0.50	$+0.047^{\dagger}$	-0.29	$+0.027^{\dagger}$	-0.17
0.75	$+0.032^{\dagger}$	-0.19	$+0.011^{\dagger}$	-0.12
1.0	$+0.002$	-0.08	$+0.018^{\dagger}$	-0.17
1.5	-0.001	-0.06	$+0.002$	-0.10
2.0	n/a	n/a	$+0.001$	-0.10

Figure 5 in detail. This traces the backbone-error dose law of Section 4.3 and Table 3. We perturb the crystal backbone to a target interface RMSD and re-score both passes, with the monomer inheriting the same perturbation so the partner ablation stays clean, then plot CPI and Spearman against the jitter magnitude. The signal is flat through sub-Ångström error and collapses past a model-dependent knee, near 1.0 Å for ProteinMPNN and 1.5 Å for ESM-IF1, rather than at a universal cliff. The decisive marker is that real predicted backbones (Section 7) fall on the surviving, low-error part of the curve, so the collapse describes an error regime a designer avoids, not the one they operate in. Seed 20260803.

Figure 6 in detail. This extends the mixed derivative to its second order, the coupling result of Section 4.3. The partner-ablated pairwise coupling is the mixed third difference (a double-mutant cycle across the two states) $C_{ij}(a, b) = L_{ij}(a, b) - L_i(a) - L_j(b)$, the change in the wild-type-to-a leverage at position i when position j is set to b, read from the conditional model by flipping the input residue at j and measuring the shift in the teacher-forced log-odds at i over decode orders where j precedes i, symmetrised over the two directions. For a cross-interface pair the two residues share no monomer, so $C_{\text{lev}} = C_{\text{complex}}$, and for a same-side pair $C_{\text{lev}} = C_{\text{complex}} - C_{\text{monomer}}$ subtracts the intra-fold coupling. By the cycle C_{lev} estimates $-g$ up to L's positive scale, where $g = \Delta\Delta G_{ab} - \Delta\Delta G_a - \Delta\Delta G_b$ is the experimental binding epistasis. The figure plots it against measured g. It tracks the experimental sign (Spearman -0.14) and survives control for inter-residue distance (distance-controlled partial $\rho(|C|, |g| \mid \text{distance}) = +0.21$), while the un-ablated coupling computed within the bound state alone is null. The sign channel is weak but above chance (chance-corrected $\rho = +0.079$ $[+0.012, +0.172]$). The contrast shows that the binding information in the second derivative, like that in the first, is reachable only by ablating the partner, not by reading the bound distribution more finely.

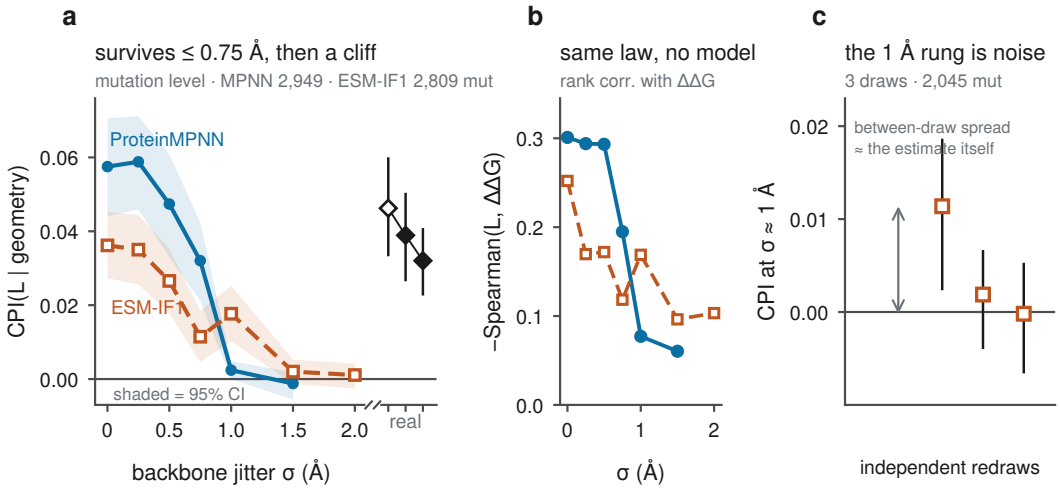


Figure 5: The backbone-error dose law. The mixed derivative survives sub-Ångström reconstruction error and collapses past a model-dependent knee. **a**, CPI(L |geometry) versus backbone jitter σ for ProteinMPNN and ESM-IF1, with the real predicted backbones (crystal, OpenFold3, AF2-multimer) in the broken-off slot at right, falling on the surviving low-error part. **b**, the same law in a model-free readout, the rank correlation of L with $\Delta\Delta G$. **c**, three independent noise redraws at $\sigma \approx 1$ Å: the between-draw spread is as large as the estimate, so the 1 Å rung is consistent with noise.

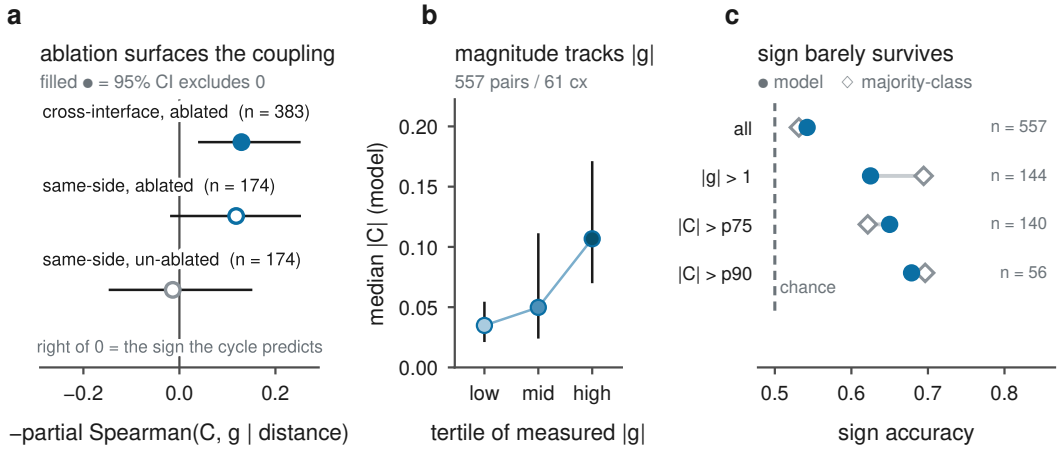


Figure 6: The second mixed derivative. **a**, the ablation forest for the partner-ablated pairwise coupling. The decisive contrast is matched: the same-side pairs scored with the partner ablated versus un-ablated use the identical $n = 174$ pairs, so ablation is the only thing that changes between them. The cross-interface pairs ($n = 383$) are a larger, independent arm that is ablated by construction (the two residues sit on different chains), which is why the two populations differ in size. **b**, the coupling magnitude tracks measured binding epistasis with the cycle-predicted sign, surviving control for inter-residue distance. **c**, the honest sign channel: model versus majority-class sign accuracy across subsets.

A.3 STEERING THE FROZEN MODEL, AND THE REGIME DEPENDENCE

Table 4: Steering benefit across independent readouts (paired contrasts at $\alpha = 2$, complex-clustered 95% CI). ipTM rows are within-folder paired effects (AF2 and Boltz-2 are differently calibrated, each read in its own scale). Judge rows are on the leverage scale. Self-judged cells (steer-X, judge-X) are excluded as trivially circular. The L − naive specificity control (the confidence tilt) was scored on the 60-complex ProteinMPNN-steered subset, so it is reported where available (dashes elsewhere) and the FoldX L − naive is in Table 2. On the foldability readout (ipTM) the naive tilt significantly wins (one of the two ipTM readouts carries this contrast), and on the binding readouts (the inverse-folding judges) L wins, the confidence-is-not-competence split one level up.

Steered model	Independent readout	Paired L − random	Paired L − naive
ProteinMPNN	AF2-multimer ipTM ($n = 120$)	$+0.235$ $[+0.199, +0.272]$	-0.034 $[-0.064, -0.007]$
ProteinMPNN	Boltz-2 ipTM ($n = 60$)	$+0.139$ $[+0.101, +0.177]$	n/a
ProteinMPNN	ESM-IF1 leverage (judge)	$+0.77$ $[+0.72, +0.82]$	$+0.159$ $[+0.110, +0.207]$
ProteinMPNN	MIF leverage (judge)	$+0.71$ $[+0.66, +0.76]$	$+0.242$ $[+0.191, +0.292]$
ProteinMPNN	PiFold leverage (judge)	$+0.89$ $[+0.84, +0.95]$	n/a
ESM-IF1 (reverse)	AF2-multimer ipTM ($n = 60$)	$+0.158$ $[+0.102, +0.215]$	n/a
ESM-IF1 (reverse)	ProteinMPNN leverage (judge)	$+0.43$ $[+0.37, +0.49]$	n/a
ESM-IF1 (reverse)	MIF leverage (judge)	$+0.57$ $[+0.48, +0.66]$	n/a
ESM-IF1 (reverse)	PiFold leverage (judge)	$+0.73$ $[+0.62, +0.84]$	n/a

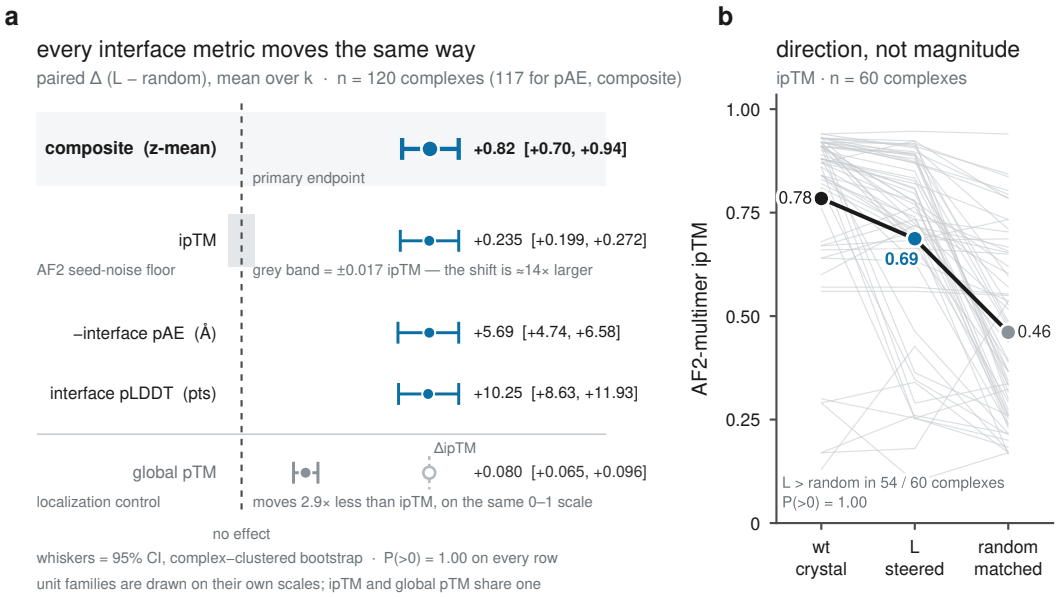


Figure 7: Steering a frozen ProteinMPNN by the mixed derivative raises AF2-multimer interface ipTM against a matched-magnitude random control, and the gain is interface-localized (global pTM rises far less). **a**, the aggregate forest over the full 120-complex judge set. **b**, the per-complex arms on the 60 complexes that also carry a wild-type fold, so specificity (L vs random, $+0.226$) and recovery (L vs wt, -0.097) read at once. Each faint grey line is one complex and the bold line is the arm mean.

Figure 7 in detail. This is the structure-level view of the steering experiment of Section 5.1 and Table 4. We bias a frozen ProteinMPNN's interface logits by $+\alpha L$, sample sequences, and fold them with AF2-multimer, then compare the leverage-steered interfaces to a matched-magnitude random direction on interface ipTM over 120 complexes. Steering raises interface ipTM by $+0.235$ against the random control, and the gain is interface-localized, since global pTM rises about 2.9 times

less, so the effect concentrates where the guidance acts rather than improving the fold as a whole. Because the direction and not the perturbation size carries the benefit, random is the fair null for the perturbation and the naive tilt (Table 2) is the fair null for the direction.

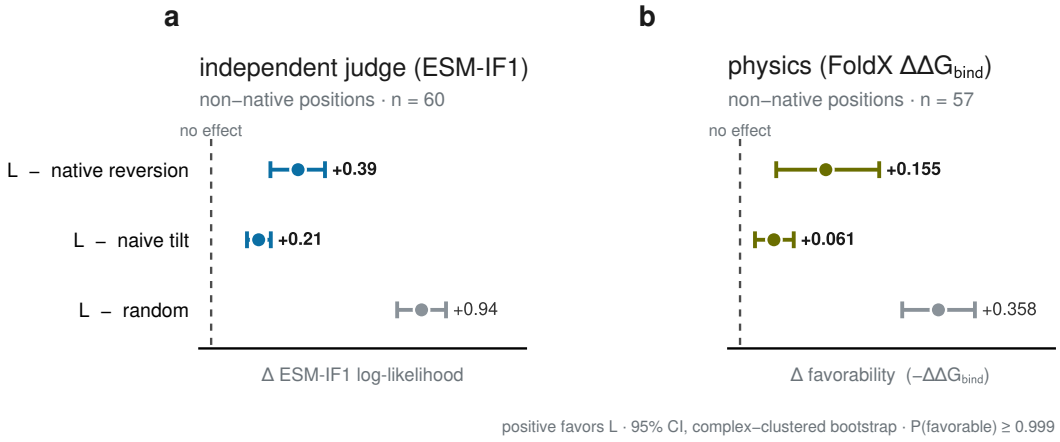


Figure 8: The native-reversion control (Section 5.1). At non-native positions, where reverting to the wild-type residue has nowhere to go, the leverage tilt L beats both a native-reversion arm and the naive confidence tilt on an independent sequence model (**a**, ESM-IF1 judge leverage) and on a physics energy function (**b**, FoldX $\Delta\Delta G_{\text{bind}}$, sign flipped so a positive bar favors L), on 60 complexes with every $P(\text{favorable}) \geq 0.999$. The wt-centered direction L' equals L bit-for-bit by softmax invariance, and the L arm stays 64 of 64 distinct at 44% mean pairwise identity, so the gain is neither native recovery nor mode collapse.

Figure 10 in detail. This panel explains why confidence sometimes appears to work, and is the regime-dependence result behind Sections 4.2 and 4.3. Confidence-AUROC for flagging binding-critical residues climbs a fold-coupling gradient across interface classes and only clears chance on de-novo binders, where the fold and the binding site are most nearly the same constraint, whereas the mixed derivative clears chance in every class, including de-novo (interface AUROC 0.57, and on de-novo it adds binding signal beyond geometry, CPI $+0.011$, though confidence edges it there, Section 9). The figure shows that both readouts are regime-dependent and their ordering reverses: on natural complexes the derivative leads, and on de-novo the scalar overtakes it, the precise sense in which confidence is not competence in the natural regime this paper targets.

Figure 11 in detail. This is the predicted-backbone steering result of Section 7. Leverage is recomputed and the tilt applied on OpenFold3 and AF2-multimer backbones rather than crystals, judged anti-circularly by an independent model's leverage read off the predicted structure. The judge-level L − random effect ($+0.70$ to $+0.76$) matches the crystal control and L − naive stays positive, so both the readout and its specificity survive the swap to the backbone a designer actually has.

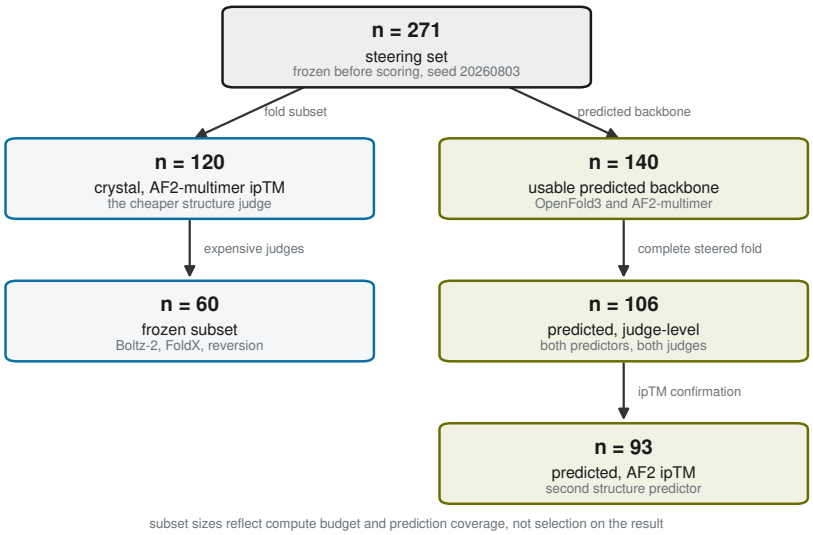


Figure 9: Complex counts through the steering pipeline (CONSORT-style). The steering set (271 complexes) splits into a crystal-backbone branch (folded and judged) and a predicted-backbone branch (the design regime). The counts fall from stage to stage for reasons unrelated to the outcome: the crystal branch narrows to the 120 complexes folded by the cheaper AF2-multimer judge and then to the 60-complex frozen subset carried through the expensive judges (Boltz-2, FoldX, native-reversion). The predicted branch is the 140 complexes with a usable backbone from both predictors, of which 106 yielded a complete steered fold judged anti-circularly and 93 also received an AF2-multimer ipTM confirmation. Every count is read from a committed CSV, so the subset sizes reflect compute budget and prediction coverage, never selection on the result.

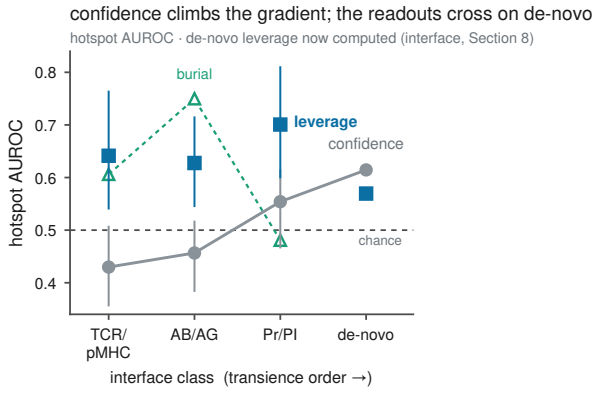
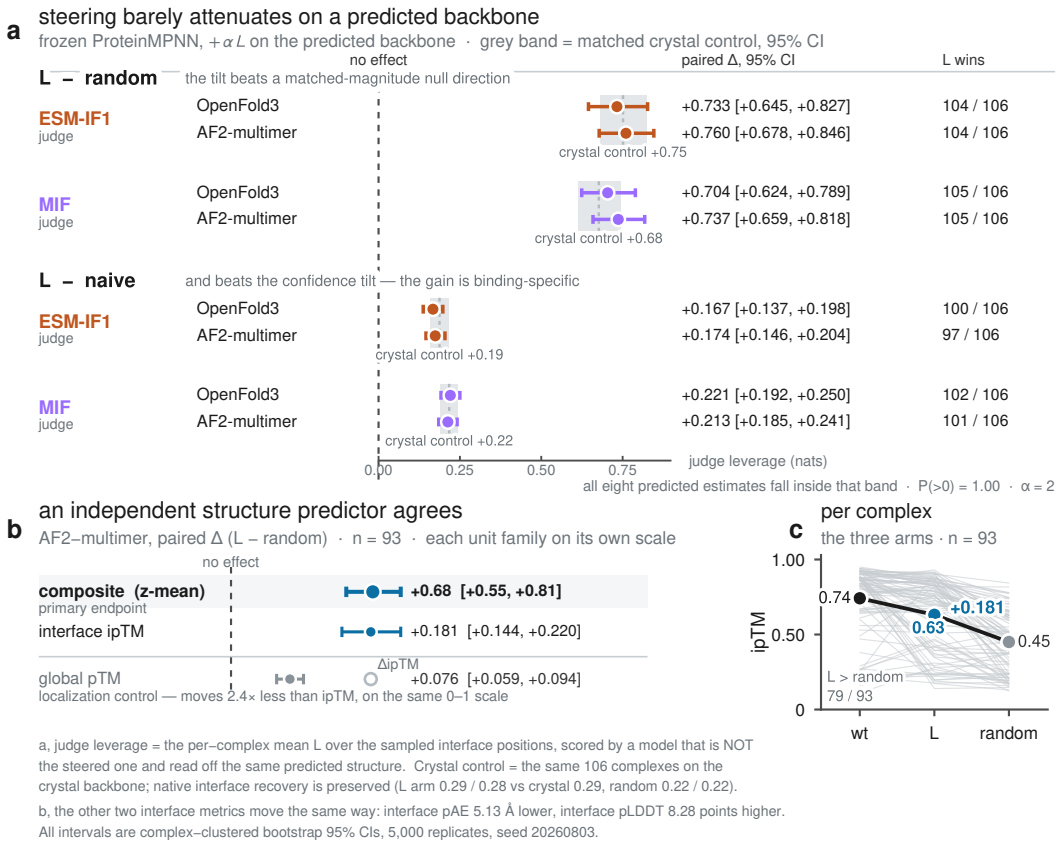


Figure 10: Confidence-AUROC climbs a fold-coupling gradient across interface classes (TCR/pMHC, antibody-antigen AB/AG, protein-protein interfaces Pr/PI, and de-novo binders) and only clears chance on de-novo binders, while the mixed derivative clears chance in every class, including de-novo (interface AUROC 0.57, adding CPI $+0.011$ over geometry). Both readouts are regime-dependent and their ordering reverses: the scalar only clears chance on de-novo binders, where it also overtakes the derivative.

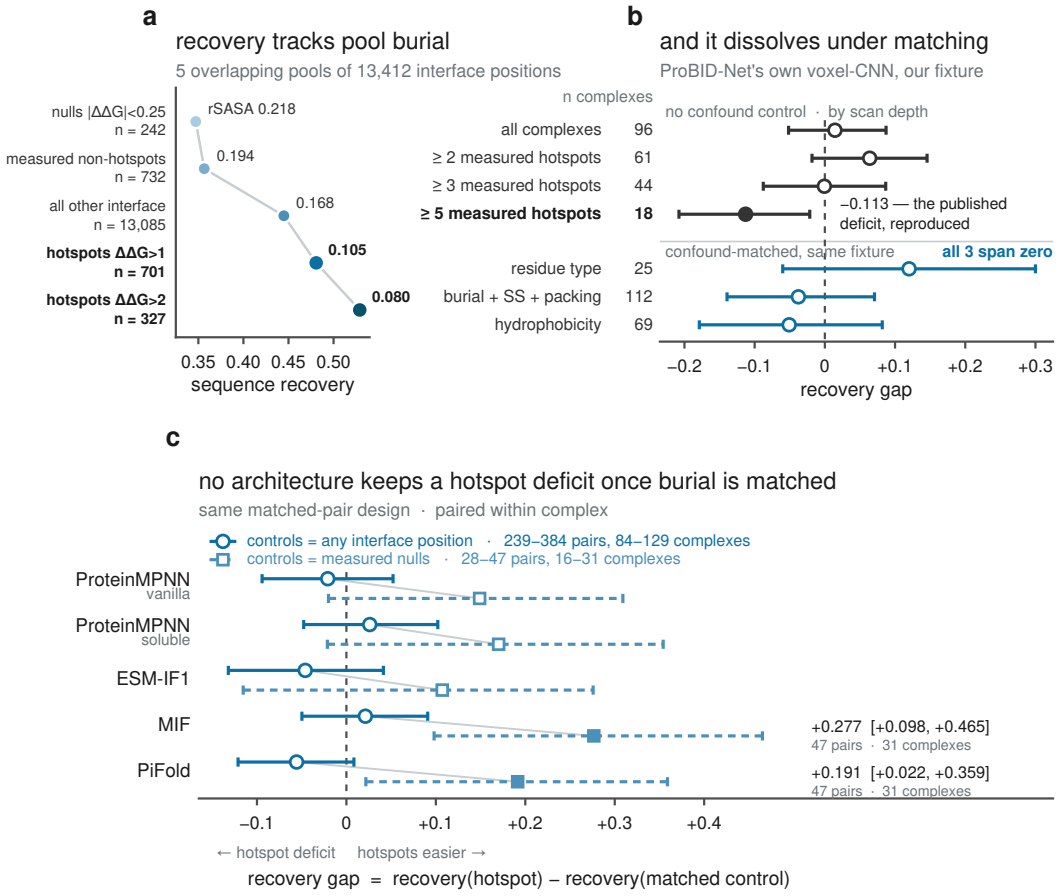


a, judge leverage = the per–complex mean L over the sampled interface positions, scored by a model that is NOT the steered one and read off the same predicted structure. Crystal control = the same 106 complexes on the crystal backbone; native interface recovery is preserved (L arm 0.29 / 0.28 vs crystal 0.29, random 0.22 / 0.22).

b, the other two interface metrics move the same way: interface pAE 5.13 Å lower, interface pLDDT 8.28 points higher. All intervals are complex–clustered bootstrap 95% CIs, 5,000 replicates, seed 20260803.

Figure 11: Steering survives the swap to a predicted backbone (Section 7). **a**, judge-level paired L − random and L − naive on OpenFold3 and AF2-multimer backbones under two anti-circular judges. The grey band is the matched crystal control's 95% CI on the same 106 complexes, and all eight predicted estimates fall inside it. **b**, an instrument outside the inverse-folding family agrees, with global pTM as the localization control. **c**, the same ipTM effect per complex.

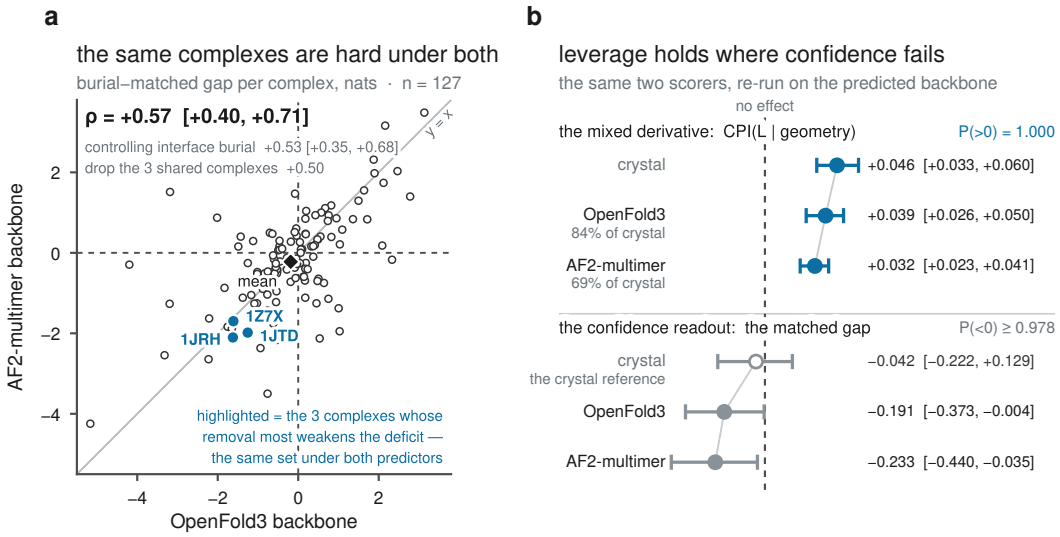
A.4 THE BURIAL CONFOUND AND THE CONDITIONING-SET TAX



whiskers = 95% complex–clustered bootstrap CI (10,000 replicates, seed 20260803); filled mark = CI excludes zero.
ΔSASA is the one cheap hotspot feature the matching does not control, so it is checked separately: the 384–pair tier carries a ΔSASA imbalance of +0.072 [+0.045, +0.100], and adjusting the gap for it moves it only –0.042 → –0.028 nats.

Figure 12: On crystal backbones the raw hotspot-versus-rest recovery gap tracks burial, a burial-driven surplus for the strongly buried hotspots rather than a deficit. **a**, the confound itself: sequence recovery for the two hotspot pools (measured $\Delta\Delta G > 1$, and the strict > 2) and the three candidate control pools (assay nulls with $|\Delta\Delta G| < 0.25$, measured non-hotspots, and all other interface positions), each pool's size shown as n and its mean relative SASA beside it. Marker colour encodes burial (darker = more buried = lower rSASA). Recovery rises monotonically with burial across all five pools, so the sign of an uncontrolled "deficit" is set by which control pool is chosen, before any binding term enters. **b**, ProBID-Net's own released voxel-CNN on our fixture: the published deficit reproduces only in deeply-scanned complexes and dissolves under residue-type, burial, and hydrophobicity matching. **c**, the same matched-pair design across five inverse-folding architectures on two matched tiers, and no architecture keeps a deficit once burial is matched. The published deficit is thus largely a burial and composition confound.

Figure 12 in detail. This documents the burial confound of Section 6, the first of the two corrective results. On crystal backbones the raw hotspot-versus-rest recovery gap tracks burial closely, so an uncontrolled comparison attributes to hotspots what is really a burial effect. Under the matched-pair design, which compares hotspot and non-hotspot positions within the same complex at matched relative SASA (± 0.05), matched secondary-structure class, and matched neighbour count (± 1), the deficit attenuates to statistical indistinguishability. The figure is the evidence that the published crystal-backbone deficit is largely a burial artifact rather than a binding-specific failure.



b, upper block: CPI(L | burial + neighbours + ΔSASA), mutation level, 2,520 mutations over 140 shared complexes; lower block: the burial–matched paired design in nats, 382 pairs / 127 complexes predicted (384 / 129 crystal). The two blocks are different quantities: each is on its own scale and every value is printed, so only the sign is compared. 95% complex–clustered bootstrap CIs.

Figure 13: On predicted backbones a burial-matched deficit reappears, and the two architecturally-independent predictors agree per complex, so the residual tax lives in the conditioning set, not in the benchmark. **a**, the per-complex burial-matched gap under OpenFold3 against AF2-multimer, with the $y = x$ guide, the Spearman correlation, and the partial correlation controlling for interface burial. The three complexes whose removal most weakens the deficit are the same set under both predictors. **b**, the divergence the decomposition predicts on those predicted backbones: the mixed derivative (CPI(L |geometry)) stays positive while the confidence-type matched gap goes negative, each block on its own scale with every value printed.

Figure 13 in detail. This is the complementary result of Section 7 and the reason the paper is a correction in both directions. Repeating the burial-matched comparison on predicted rather than crystal backbones, a matched deficit reappears, and the two architecturally independent predictors (AF2-multimer and OpenFold3) agree per complex on where it falls. Because the deficit is absent on crystal backbones but present on predicted ones, and the two predictors concur, it cannot be a benchmark artifact and must live in the conditioning set the designer actually has, which is the conditioning-set tax of Section 7.

A.5 Physics adjudication and the beyond-X ladder

Figure 14 in detail. This is the physics adjudication of Section 5.2 and the companion to Table 2. Every structure is FoldX RepairPDB-relaxed once before its binding $\Delta\Delta G$ is computed, and the figure reports two readings. As a detector the fitted physics energy ranks experimental $\Delta\Delta G$ better than the zero-shot leverage, as expected of a hand-fit function, yet the leverage retains binding rank-signal beyond it in partial correlation, so it is not merely re-deriving the physics. As a steering readout the mixed derivative beats the confidence tilt on the same physics energy, which places FoldX with the inverse-folding judges against ipTM and shows the advantage is binding-specific rather than foldability. The full FoldX steering contrasts, in kcal/mol: $L-$ naive best-of-k $+1.86$ $[+0.72, +3.02]$ with the conservative mean-over-k $+1.24$ $[+0.14, +2.33]$, $L-$random $+6.27$ $[+5.03, +7.49]$, and naive $-$ random $+4.79$ $[+3.46, +6.11]$ (both vs-random best-of-k).

Figure 15 in detail. This is the synthesis figure for the "beyond-X" claims of Sections 4.2 and 5.2, drawn in one matched metric so the controls are comparable. Each bar is the partial rank-correlation of leverage with experimental $\Delta\Delta G$ after controlling for one competing feature class in turn: cheap geometry, substitution similarity (BLOSUM), the one-pass complex log-odds, and a fitted physics energy function. Leverage clears every rung, while every per-position aggregate scalar of the bound

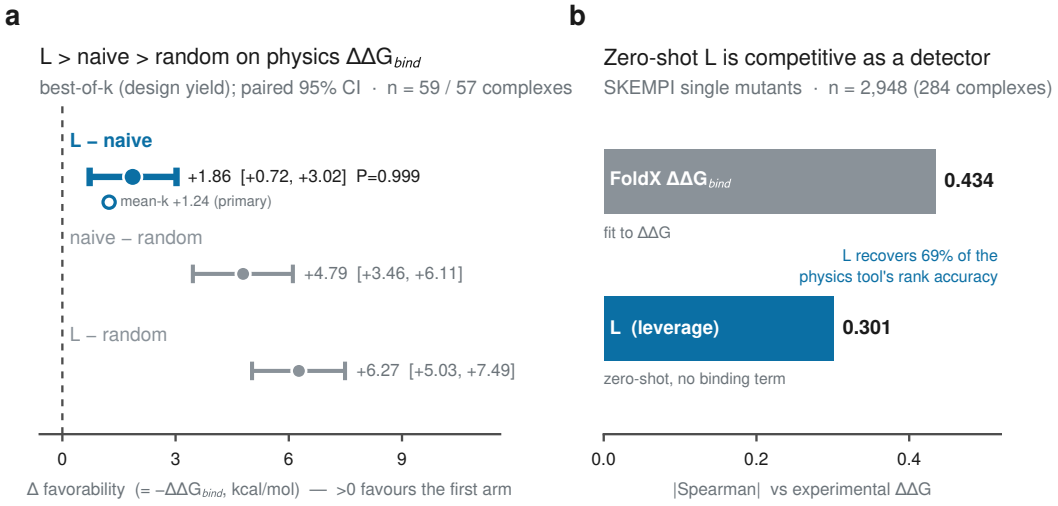


Figure 14: FoldX adjudication (companion to Table 2). **a**, steering specificity on the physics energy: the mixed derivative beats the confidence tilt (L − naive, the decisive contrast) as well as both vs-random controls, so the advantage is binding-specific, not simply foldability. **b**, detection: the fitted physics ranks experimental $\Delta\Delta G$ better than the zero-shot leverage, which still recovers about 69% of its rank accuracy with no binding-energy term.

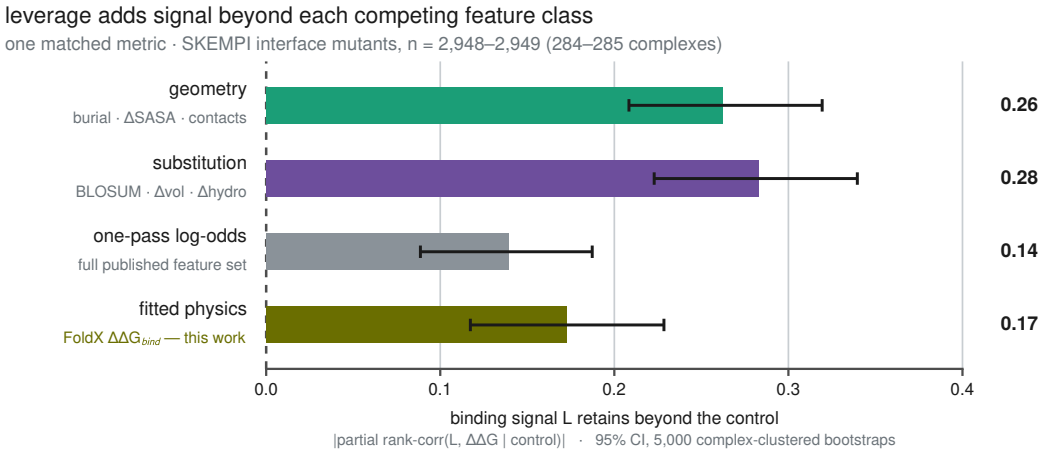


Figure 15: The beyond-X ladder in one matched metric (partial rank-correlation of L with experimental $\Delta\Delta G$, controlling for each competing feature class). L clears geometry, substitution similarity, the one-pass log-odds, and a fitted physics energy function. Every per-position aggregate scalar of P stays at the floor.

distribution stays at the floor. The figure collects in one place the claim that the mixed derivative adds binding information beyond each alternative a reviewer would propose, not just beyond geometry.

B FALSE-POSITIVE MODES AND THEIR CONTROLS

Six ways to get a false positive were named in advance, before any number was computed (Table 5). One (assay heterogeneity) is controlled only in part, and we say so. The pattern worth noting is that the two largest confounds, burial and training leakage, both run against the hypothesis, so the burden is to show the effect despite them, which is why the matched-pair design, not the raw deficit, is the experiment.

Table 5: False-positive modes named in advance, and their controls.

Mode	Control	Result
Burial (buried positions are where inverse folding is most confident, so it hides the effect)	within-complex matched pairs (rSASA, secondary structure, neighbour count) plus CPI over geometry at every step	deficit attenuates across all five architectures, including ProBID-Net, and leverage CPI $+0.0049$ survives full geometry
Native amino-acid identity	per-type breakdown, alanine-only subset, substitution-similarity partialled out	Spearman negative in 18/19 wt-types and -0.25 on Ala-only
PDB training leakage	the double difference cancels a native-confidence bias shared by both passes, and the signal holds within-complex, not only pooled (Appendix C)	inflating native probability cannot inflate leverage, only single-pass recovery and confidence, which then cancel
Assay heterogeneity	strict (> 2) and loose (> 1) hotspot thresholds, both reported	conclusions hold under both (partial: no temperature/pH stratification, disclosed)
Positional independence	measure the second mixed derivative directly	couplings track epistasis, Spearman -0.14
Decoding-order variance	committed L is order-free, and a conditional scorer symmetrised over 16 orders	stable (disjoint-half Spearman 0.93 MPNN, 1.00 ESM-IF1, tracking the order-free L at 0.88), oracle order inert (-0.002)

C PROOFS, PROTOCOLS, EXTENDED RESULTS, AND CAVEATS

C.1 DECOMPOSITION: PROOFS AND SCOPE OF PROPOSITION 1

Precise scope of Proposition 1 (assumptions and logic). We separate what is mathematics from what is measured, because the two are easy to conflate.

The cycle, (i). $L_i(a)$ estimates $-\Delta\Delta G_{\text{bind}}$ up to one positive scale under three conditions. First, an energy readout: the logit gap $\text{logit}(a \mid X_{\text{complex}}) - \text{logit}(a \mid X_{\text{monomer}})$ equals $-\beta\,\Delta E(a)$ up to an a-independent constant, with a single inverse temperature β shared across the two passes. A bare "Gibbs form" assumption is empty, since any strictly positive distribution is Gibbs for some energy. The content is that the two passes share one temperature: were β to differ between them, the self-energy would not cancel. Second, a rigid ablation: X_{monomer} is X_{complex} with the partner removed and the remainder held fixed, needed only up to an a-independent shift of the self-energy, so a uniform relaxation is harmless but residue-specific reorganization adds a term we do not model. Third, common support, so the log-ratios are finite. Only the sign and rank of the energy contrasts are recovered, as β is not identified from L. We call L the mixed second *derivative* in the abstract and introduction by analogy to the continuous classifier-free-guidance score, and on the discrete amino-acid alphabet it is the finite difference of Section 4.1.

Non-identifiability, (ii). L equals a measurable function $\psi(P)$ almost surely (that is, with probability one) if and only if $Q = p(\cdot \mid X_{\text{monomer}})$ is a Borel function of P, which fails exactly when two

inputs with the same bound conditional P can differ in the ablated conditional Q, that is when the conditional law of L given P is not a point mass (equivalently, for $L \in L^2$, when $\mathrm{Var}(L \mid P) > 0$ on a positive-probability set). At fixed P the map from Q to L is a bijection, with inverse $Q_a \propto \exp(\log P_a - L_a)$ after normalization, so (P, L) and (P, Q) generate the same information and L varying at fixed P is exactly Q varying at fixed P. This exact-complement property is what licenses replacing $\mathrm{Var}(\Delta\Delta G \mid P, Q)$ by $\mathrm{Var}(\Delta\Delta G \mid P, L)$ in the floor decomposition below. Identifiability holds trivially, with L carrying no information beyond P, in the degenerate cases where partner ablation does not move the logits, or where the second pass is a fixed reparameterization of the first (temperature or mixture), which the measured $R^2(|L|_{\mathrm{rms}} \mid P) = 0.37$ excludes. The readout is fixed only up to the wild-type reference.

The floor is three separate statements. Non-identifiability of L from P bounds only the prediction of L itself and says nothing on its own about predicting binding. The binding floor is the separate, measured fact that $\mathbb{E}[\mathrm{Var}(\Delta\Delta G \mid P)] > 0$ (Section 4.1, 88 to 90% unexplained by these learners), which by the minimum-mean-squared-error identity equals the best P-only risk $\inf_\varphi \mathbb{E}[(\Delta\Delta G - \varphi(P))^2]$, attained at $\varphi(P) = \mathbb{E}[\Delta\Delta G \mid P]$. That L closes part of it is the third, empirical claim, read off the variance decomposition $\mathbb{E}[\mathrm{Var}(\Delta\Delta G \mid P)] = \mathbb{E}[\mathrm{Var}(\Delta\Delta G \mid P, L)] + \mathbb{E}[\mathrm{Var}(\mathbb{E}[\Delta\Delta G \mid P, L] \mid P)]$, whose second term is the part of the floor attributable to L and which Section 4.2 measures to be strictly positive. Conditioning on any feature weakly reduces variance, so only a strict reduction is evidence, which is why we report it as a measured effect and not as a corollary of (ii).

Why the scalar KL stays at the floor. The two-pass scalar KL does query the monomer, so it is not insensitive to binding, yet $D_{\mathrm{KL}}(p(\cdot \mid X_{\mathrm{complex}}) \,\|\, p(\cdot \mid X_{\mathrm{monomer}})) = \langle P, L \rangle$ up to the reference shift, a single inner product that collapses the signed per-substitution leverage vector to one non-negative number and down-weights exactly the low-probability substitutions where leverage is often largest. The contraction is rank-one, since at fixed P the leverage vector has $|A| - 1$ free coordinates against the single scalar of the KL, so it discards $|A| - 2$ of them, 18 of 19 for the 20-letter alphabet. Its failure is this contraction, not insensitivity to the ablated state, which is why a two-pass *scalar* sits at the floor while the two-pass *vector* clears it.

Why training leakage does not inflate leverage. Leverage is a within-model double difference of log-odds across two conditioning states. A memorized native-residue preference enters the complex pass and the monomer pass as the same per-position, per-substitution bias, and any bias shared by the two passes cancels exactly in the double difference, dropping out of both within-state contrasts equally. So the mechanism by which leakage would inflate a single-pass metric, an inflated native probability, does not inflate leverage: it inflates recovery and confidence, which are single-pass and which we show fail. Even an asymmetric native-residue spike, present in the complex pass but not the monomer pass, only shifts a position's leverage by a constant across its non-native substitutions, leaving every within-position rank statistic unchanged. The one channel that could survive is pooling across complexes without within-complex centering, and we close it directly: the leverage-versus-experiment relationship holds within complexes, not only pooled (within-complex-centered Spearman -0.29 versus pooled -0.30, and negative in 78% of the 116 complexes with at least six interface mutations, Wilcoxon $p < 10^{-11}$, from `leverage_within_complex.csv`). The residual concern is genuine label contamination, a partner-conditioned memory of which substitutions are damaging held more strongly in the bound than the unbound pass, which is a far stronger claim than memorizing the native residue and is bounded by the predicted-backbone result, where the exact crystal is not the model's input.

C.2 Protocols and implementation details

Matched-pair retention and the hides-versus-invents logic. The within-complex 1:1 matching pairs only the hotspots that have an intra-complex null at the same relative SASA, secondary-structure class, and neighbour count, which is a minority of the 327 SKEMPI hotspots (42 pairs in the loose-null primary, 21 in the strict tier). Crucially the matched hotspots are not the exposed ones: 66 to 67% of them are buried below relative SASA 0.10, matching the 69% of all 327 hotspots, so there is no survivorship bias toward less-buried hotspots. We therefore do not rest the conclusion on the matched subset: the higher-powered regression estimator uses all interface hotspots with no 1:1 matching and agrees. The logic of the control is one-directional. Because buried positions are where

inverse folding is most confident and accurate, an uncontrolled hotspot-versus-rest comparison hides a real deficit rather than inventing one, so controlling burial can only reveal a deficit that survives it, never manufacture one. The attenuation to statistical indistinguishability therefore means the crystal deficit is largely burial, not that matching erased a real effect. This hides-not-invents direction is the one we verify for ProteinMPNN and ESM-IF1, where buried positions are recovered more accurately, so it makes matching conservative and their raw within-complex signal is a burial-driven surplus rather than a deficit. It is not a universal law across models. ProBID-Net's published deficit is driven primarily by residue composition rather than burial (matching residue type alone flips its sign to positive), and for its voxel-CNN the burial term need not run the ProteinMPNN direction, so the unifying claim is that the published deficit does not survive matching its confounds, not that burial always hides one.

CPI and the placebo floor. The conditional predictive impact of a feature is estimated by cross-fitting a model of hotspot-ness on cheap geometry (burial, neighbour count, ΔSASA), then measuring the loss in out-of-sample skill when the feature's information within geometry strata is destroyed by a conditional permutation test (Watson & Wright, 2021; Berrett et al., 2020). The placebo floor is that same statistic computed on a deterministic pure-noise function of geometry, so it reports the estimator's false-positive level under the null, measured at $+0.0007$. A pure-noise feature itself yields CPI indistinguishable from zero (-0.00002), confirming the estimator does not leak, and a feature counts only if its CPI clears that floor rather than merely exceeding zero. All CPI intervals are complex-clustered bootstraps.

FoldX and backbone-jitter protocols. For the physics adjudication (Section 5.2) each crystal is FoldX RepairPDB'd once, the standard relaxation that removes baseline steric clashes, and cached. A steered or single-point sequence is scored by encoding its differences from the wild type as FoldX point mutations and calling BuildModel, which repacks the mutated side chains and builds a matched wild-type remodel through the identical procedure, at five runs. $\Delta\Delta G_{\text{bind}}$ is then $\text{mean}(\text{IE}_{\text{mut}}) - \text{mean}(\text{IE}_{\text{wt}})$, where IE is the AnalyseComplex interaction energy between the two chain groups. Because the mutant and the wild-type reference pass through the same repair and repack, the reported gain is a difference of relaxed states, not an unrelaxed-clash artifact. For the dose law (Section 4.3) the crystal backbone is perturbed by adding independent Gaussian noise to each of the N, Cα, C, O backbone-atom coordinates, with per-coordinate standard deviation scaled so the per-atom RMSD equals the target ($\sigma/\sqrt{3}$), and the monomer inherits the identical perturbation so the partner ablation stays clean. This perturbation does not enforce ideal bond geometry, which is one more reason the jitter ladder is only a controlled diagnostic and not the load-bearing claim: the decisive test is that real, physically-valid predicted backbones fall on the surviving part of the ladder, so the collapse past the model-dependent knee reflects reconstruction error a designer actually incurs, not an artifact of the perturbation geometry.

Steering implementation and evaluation subsets. The steering bias is static. The per-position leverage 20-vector L is computed once from the committed crystal-conditioned P and partner-ablated Q, then added as a constant term $+\alpha L$ to the interface logits of a frozen ProteinMPNN during autoregressive sampling, and it is not recomputed from the generated prefix at each step. Sampling uses temperature 1.0 and $K = 64$ samples per setting, and the matched random control is a per-position permutation of L, so it carries the identical per-position magnitude and only the direction differs. The native-reversion control adds a one-hot bias toward the wild-type residue and centers every contrast on non-native positions. A wt-centered variant $L' = L - \text{mean}_a L_a$ that subtracts each position's mean leverage before tilting yields a sampled distribution identical to the raw L tilt bit-for-bit, because a per-position additive shift cancels in the softmax, so the reported advantage is a property of L's shape and carries no native-anchored offset. The tilt diversifies rather than collapses: at $\alpha = 2$ the L arm returns 64 of 64 distinct sequences at 44% mean pairwise identity (entropy 1.27 nats), against the native-reversion arm's near-degenerate 58 of 64 at 85% (L' is identical to L, as above). The non-native restriction is what makes the reversion control fair: on *total* $\Delta\Delta G_{\text{bind}}$ native reversion is strong because it restores wild-type contacts ($L - \text{wttilt} = +0.77$, n.s.), so the design-relevant question is at positions a designer cannot simply revert. There FoldX sides with L: $\Delta\Delta G_{\text{bind}}$ favors L over the naive tilt by -0.06 $[-0.10, -0.03]$ and over native reversion by -0.16 $[-0.25, -0.07]$ kcal/mol (lower binds better), matching the ESM-IF1 judge. Because a cross-architecture judge reads the same functional, we report the baseline inter-model leverage

correlation: per-position Spearman between ProteinMPNN and ESM-IF1 leverage is $+0.52$ (L_{ala}) to $+0.64$ ($|L|_{\text{rms}}$) on 13,037 shared interface positions (`reviewer_controls.csv`), correlated but far from identical, so the inverse-folding judge is a consistency check rather than an independent instrument and the binding-specific claim rests on FoldX, whose energy function shares no training data with the sequence models. The evaluation subsets were fixed before any structure was folded or scored (rule 1, timestamped in the anonymized repository), and Figure 9 traces the counts. The folding sets are a seed-shuffled (seed 20260803) first-N of the complexes that carry all seven sampled conditions and fall under a residue-length cap, so they are not chosen by any outcome. The cheaper judge (AF2-multimer, 120 complexes) covers more than the expensive ones (Boltz-2 and FoldX, a frozen 60-complex subset). On predicted backbones the detection set is the 140 complexes with a usable prediction from both predictors, and the steering sets are those of the 140 that also yielded a complete steered fold: the judge-level effect is measured on 106 complexes across both predicted backbones, and 93 of them also received an AF2-multimer ipTM confirmation, with the crystal arm scored on the matched subset as a positive control. Subset size therefore reflects compute budget and prediction coverage, not selection on the result.

C.3 Robustness of the feature-class law

Table 6 first pins down the samples and aggregations that "the crystal CPI" denotes in different parts of the paper, so the numbers below and in the main text are read on a common footing.

Table 6: Crystal-CPI sample definitions. "The crystal CPI" denotes different samples and aggregations in different places, and all trace to committed CSVs. The feature-class law (Table 1) is the position-level CPI_{pos} over all SKEMPI interface positions. The predicted-backbone comparison (Section 7) uses the mutation-level CPI_{mut} on the 140 complexes with a usable predicted backbone, so its retention rate is a matched crystal-versus-predicted contrast (OpenFold3 84%, AF2-multimer 69% of the crystal value). Sources: the feature-class row from `w_placebo_ladder.csv`, the predicted-comparison rows from `leverage_predicted.csv`.

Sample	Aggregation	CPI(L | geometry)
Feature-class law, all SKEMPI	position, 13,401 pos	$+0.0049$ $[+0.0034, +0.0065]$
Predicted-comparison, crystal	mutation, 140 cx	$+0.046$ $[+0.033, +0.060]$
Predicted-comparison, OpenFold3	mutation, 140 cx	$+0.039$ $[+0.027, +0.050]$
Predicted-comparison, AF2-multimer	mutation, 140 cx	$+0.032$ $[+0.023, +0.041]$

Leverage-robustness controls. Four controls (committed CSVs, seed 20260803) confirm the binding signal is not an artifact of decoding order, backbone-error resolution, evolutionary conservation, or homology. **Decoding order:** the committed L is order-free (an unconditional or teacher-forced marginal), and an order-*dependent* conditional scorer averaged over 16 random decode orders is stable (disjoint-half Spearman 0.93 for ProteinMPNN, 1.00 for ESM-IF1) and tracks the committed order-free L at 0.88 (`decode_order_spread.csv`). **Backbone-error resolution:** under 10 independent Gaussian backbone-jitter draws per level, CPI(L | geometry) decays gracefully from $+0.057 \pm 0.002$ at 0 Å (reproducing the committed crystal value) to $+0.009 \pm 0.005$ at 1.0 Å and $+0.001 \pm 0.002$ at 1.5 Å (`jitter_ladder_multi.csv`), the multi-seed refinement of the single-draw Table 3 (whose $+0.002$ at 1.0 Å is one noisier realization), both placing the knee near 1.0 Å. **Conservation:** on the MSA-covered subset, L's increment survives a real MSA-derived conservation covariate, CPI(L | geometry + MSA-conservation) $= +0.0096$ $[+0.0066, +0.0128]$ against a $+0.0030$ floor, and the conservation covariate is itself weak beyond geometry ($+0.0004$, n.s.) (`cpi_msa_conservation.csv`). **Homology:** clustering SKEMPI at 30% interface identity (343 complexes into 81 clusters) and resampling by cluster, CPI(L | geometry) $= +0.0042$ $[+0.0022, +0.0063]$ still clears the floor ($+0.0010$) and Spearman($L, \Delta\Delta G$) $= -0.30$ $[-0.34, -0.21]$ holds, so the signal is not homology leakage (`homology_cluster_spearman.csv`).

C.4 THE COUPLING AND THE DE-NOVO REGIME

The second mixed derivative. The mixed derivative extends to second order. The partner-ablated pairwise coupling tracks measured binding epistasis with the cycle-predicted sign (Spearman$(C, g) = -0.14$ $[-0.23, -0.07]$) and survives control for inter-residue distance (partial -0.12). The effect is modest and reports coupling magnitude more than sign, and partner ablation is what surfaces it: the un-ablated coupling is null.

The de-novo regime. On the Cao de-novo designs (Cao et al., 2022) with site-saturation labels the regime differs from natural complexes. Here even the scalar complex-conditioned distribution ranks substitutions by binding (interface AUROC 0.615, above every substitution baseline), it answers the stability question better than the binding question (0.721 vs 0.615, the control a fold-trained model should pass), and it adds $+0.018$ beyond an all-atom rotamer-repacked occlusion baseline, reproduced by a non-parent scorer. The mixed derivative also clears chance at de-novo interface positions (leverage AUROC 0.570 $[0.563, 0.576]$ against the binding label, CPI $+0.011$ over geometry), but it no longer dominates the scalar there, the regime convergence of Section 9 (`reviewer_round2.csv`, `leverage_bennett_denovo.csv`).

C.5 RULING OUT COMPETING MECHANISMS FOR THE CONDITIONING-SET TAX

Ruling out competing mechanisms. Two alternative explanations for the residual conditioning-set tax of Section 7 are separately measured and refuted. A *sample-budget* effect (at low sampling temperature a multi-residue hotspot constellation has astronomically small joint probability) is identical at burial-matched control constellations (median difference 0.000, 95% CI $[-1.67, +2.09]$ spanning zero): it is a generic property of low-temperature sampling, not a hotspot tax, and no oversampling recovers it. A *commitment-ordering* effect (the autoregressive schedule paints hotspots into a corner) is inert on ProteinMPNN (oracle-order difference-in-differences -0.002, `p2_ordering_stage1_summary.csv`), and on MultiFlow (Campbell et al., 2024) (a discrete flow-matching model for protein sequence-structure co-design) varying the unmasking order has only a marginal effect (order-span 0.012, far below the 0.065 sample SD across seeds and reference structures, `expB_ordering.csv`).

C.6 CAVEATS AND SCOPE

Caveats specific to the decomposition. Orthogonal is not independent: confidence cannot express leverage, but the two are weakly-to-moderately correlated and the correlation is model-dependent (Spearman(confidence, $|L|$) $= +0.075$ for ProteinMPNN, $+0.30$ for ESM-IF1), so we claim insensitivity by construction, not statistical independence. L estimates $-\Delta\Delta G_{\text{bind}}$ only up to a temperature, and all headline readouts are rank-based, but a single 2-parameter out-of-fold affine reads one unit of L as about 0.42 kcal/mol (OOF RMSE 1.82, next to fitted FoldX's 1.79, and evidence for the single shared temperature the cycle assumes), a good ranker and a rough absolute predictor, kept out of every headline. The decomposition replicates under a second architecture (ESM-IF1) on the 337 complexes both models cover, of which the 285 carrying single-mutation $\Delta\Delta G$ form the dose-law and detection subset.

Why an in-silico result is informative now. The detection half rests on real $\Delta\Delta G$ across three fixtures, and the steering half on instruments from three classes (a correlated second sequence model, two structure predictors reading foldability, and out of family a physics energy function with an explicit binding term, on which the binding-specific claim rests, Appendix C.2), the standard by which comparable design methods are judged. A wet-lab loop is the one step we scope as pending.

One natural primary fixture. That the primary claims rest on one natural fixture is structural to the field's data rather than a gap in our search: SKEMPI 2.0 is the largest curated affinity resource, with the later resources built around it.

Predicted-backbone leakage. OpenFold3 and AF2-multimer were trained on PDB subsets that likely include the SKEMPI wild-type structures, so the absolute leverage-retention rates on predicted backbones (Section 7) may be optimistic. Two features of the design mitigate this. The

conditioning-set tax rests on the per-complex agreement between two architecturally-independent predictors ($\rho = 0.565$), which a shared training set would inflate within each predictor but not obviously in their agreement pattern. Nor is it a prediction-quality artifact: residualizing each predictor's per-complex deficit on its interface RMSD barely moves the agreement (0.565 to 0.544), and the deficit correlates only weakly with interface RMSD (Spearman -0.06 for OpenFold3, -0.21 for AF2-multimer, `reviewer_controls.csv`). And a noised-crystal control at matched RMSD shows no deficit, so the signal tracks independent reconstruction rather than memorized coordinates. The genuinely de-novo regime, which no predictor could have memorized, is checked separately on Cao-2022 designs (interface AUROC 0.615), and the dose law scopes the claim to the low-error regime, so higher-error de-novo predictions past the model-dependent knee fall outside it.