PRESTO: integrating protein language models and structural features for protein-complex affinity prediction

Ivan Meić[1], Buğra Alptekin Sarı[1], Steve Muller[1], Fedor Chishchin[1], Gabriel Richman[1]*

[1]Omic Inc., Seattle, Washington, USA

*Correspondence: Gabriel Richman, gabe@omic.ai

Abstract

Binding affinity prediction, de novo binder ranking and mutational effect prediction are treated as one problem by practitioners and as three by the tools available, and a method built for one is rarely tested on the others. PRESTO pairs gradient-boosted trees over engineered biophysical features with an ensemble of interface cross-attention networks over protein language model embeddings, and we evaluate it on all three tasks. For absolute binding free energy it roughly halves the error of the established contact-based predictor on the complexes both methods score, with an applicability domain we measure rather than assume: accuracy degrades where a target is unrelated to any training target, and the reported figures are development estimates because model selection and evaluation share the same predictions. The other two tasks yield negative results that we report as findings. An absolute affinity does not rank designed binders and degrades structural confidence when combined with it, while an objective trained on design outcomes composes with confidence instead. Mutational effects require a separate additive term, interchangeable with a published specialist. Structural confidence, absolute affinity and mutation effects carry different information and do not substitute for one another.

Introduction

Most of the human proteome lacks small-molecule binding pockets[1,2], and the protein–protein interactions that govern those proteins make PPI disruptors, peptide therapeutics in particular, an attractive modality[3–6]. The need is acute for targets whose oncogenic activity stems entirely from aberrant PPIs rather than catalytic pockets — the fusion oncoproteins EWS-FLI1 and PAX3-FOXO1, and MYC, p53–MDM2 and β-catenin[1,16–18]. Development against them depends on affinity: of an enormous design space, only sequences binding tightly enough to compete with endogenous partners are viable.

Current methods are too imprecise for this. PRODIGY[7], the established contact-based ΔG predictor, reports RMSE 1.89 on its own 81-complex benchmark but reaches 3.87 kcal/mol on our crystal structures — roughly 700-fold uncertainty in K_d, so a prediction of 100 nM is equally consistent with a potent drug and an inactive compound. Physics-based methods such as FoldX[8] and Rosetta[9] target relative ΔΔG instead, and FoldX is uncorrelated with absolute affinity on SKEMPI crystal structures. Recent machine learning approaches[10–13] improve on energetic methods but remain limited by small training sets, single databases, or evaluation protocols that overestimate generalization.

Computational design has raised the stakes. BindCraft[14] and RFdiffusion[15] generate thousands of candidates in days, and experimental validation is now the bottleneck. Pipelines rank candidates by AlphaFold2 confidence (ipTM, pDockQ), which assesses whether a predicted structure is plausible rather than whether the complex would bind with therapeutic affinity.

Here we describe PRESTO (PRotein-protein Energy from STructural and sequence Operations). For absolute affinity it combines gradient-boosted trees on engineered biophysical features with an ensemble of 14 Interface Cross-Attention Networks (ICANs) over three protein language models (PLMs). We then ask what

that model transfers to, and report the answers whichever way they fall: an outcome-trained head ranks designed binders slightly better than ipTM on average but not reliably on a new target, and absolute ΔG cannot rank mutational variants of one complex at all, so a separate additive ΔΔG term is required.

Intended use and applicability domain. The endpoint is the absolute binding free energy of a two-chain protein or peptide complex, in kcal/mol, from a structure and its two sequences. The domain is complexes resembling the training distribution: natural or naturally-derived binders, predominantly antibody–antigen and enzyme–inhibitor families, affinities between −1.7 and −21.4 kcal/mol, and targets related to those in SKEMPI v2 and PPB-Affinity. Three limits bound it, each measured below. Accuracy degrades where the target is unrelated to any training target. Predictions on de novo designed binders do not rank experimental success, and combining them with structural confidence worsens selection. And the model returns one value per complex, so it distinguishes no variant of a fixed fold.

Results

Absolute binding free energy on natural complexes, and where it holds

On 5-fold cross-validation grouped by PDB identifier over the 2,154 complexes for which all 14 ensemble members have predictions and whose measured ΔG does not also appear in another fold, PRESTO achieves RMSE 1.82 kcal/mol (r = 0.72) against 2.73 for a mean baseline. Including the 94 complexes that do reappear gives 1.79 (r = 0.75) over 2,248, and on those 94 alone RMSE falls to 0.98: each is the same complex deposited under a second PDB id, which a split grouped by PDB identifier cannot separate. We use the leak-free value throughout.

Every published ΔG predictor considered here reads a structure, so comparison is restricted to the complexes both methods score. On 174 SKEMPI complexes with deposited crystal structures PRESTO reaches 1.78 (r = 0.88) against PRODIGY's 3.87 (r = 0.32), a 2.2-fold difference (paired Wilcoxon on absolute error, $p = 7 \times 10^{-16}$); on 909 PPB-Affinity complexes with ColabFold structures, 1.91 (r = 0.69) against 5.33 (r = 0.20), 2.8-fold ($p = 4 \times 10^{-60}$). PRODIGY was designed for crystal structures, so the second figure is partly a property of its input (Table 1, Fig. 1d). One qualification belongs with the comparison: PRODIGY is also an input, supplying 27 of the 93 tabular features, so these blocks compare a fitted model with the descriptor set it was given rather than two independent methods (Supplementary Note 3).

Rescaling RMSE through exp(RMSE/RT) yields a multiplicative factor, not a calibrated interval. Split-conformal intervals fitted on development data and evaluated on complexes disjoint by canonical identity give ±3.15 kcal/mol at nominal 90% (89.1% empirical coverage) — a 200-fold range in K_d, not the 20-fold that RMSE rescaling implies. Residual bias also changes sign across the range, so ranking across a wide affinity span requires accounting for it (Supplementary Note 1). Replicate measurements give a noise floor, and it does not explain the model's error. Independent measurements of the same complex differ by 1.79 kcal/mol RMSE in the cleanest PPB-Affinity subset, indistinguishable from the model's own error, which invites the conclusion that the label-noise ceiling has been reached. The data does not support it: RMSE where repeats disagree is 1.75 against 1.88 where they are identical, and replicate spread correlates with absolute error at ρ = 0.10. Label noise is real, substantial for a minority of complexes, and not what limits this model (Supplementary Note 1).

The hybrid exploits complementary representations. A CatBoost tree over 93 engineered features — AlphaFold2 confidence metrics, 14 arpeggio contact types[21], ESM-2 pair embeddings, ESM CNN scalars and physics descriptors — gives a well-calibrated baseline (RMSE 1.91) whose summary features discard per-residue information. Each ICAN instead operates on per-residue PLM embeddings, with cross-attention logits biased by arpeggio contact types, CB–CB distances, pairwise PAE and per-residue pLDDT, so the

model attends preferentially to residue pairs in physical contact (Methods). The 14 members draw on ESM-1v[22], MINT[23] and SaProt[24]. Residuals decorrelate across PLMs only modestly — 0.76–0.85 between models from different PLMs against 0.82–0.89 within one (Figure S4) — so most of the ensemble benefit is averaging rather than complementarity. The ensemble (1.84) beats every individual ICAN (1.98–2.07), and the calibrated hybrid beats both components (Table 2, Figure S1).

These figures are development estimates. Each ICAN was trained with early stopping on its cross-validation fold, and the predictions returned for that fold are those from which ensemble membership, the hybrid weight, the calibration and every RMSE above were computed — so model selection and performance estimation share the same complexes, and the resulting optimism cannot be removed by reweighting (Supplementary Note 2). The gradient-boosted component can be refitted cheaply, and doing so bounds the effect. Under checkpoint selection on a disjoint inner split, grouping by canonical biological identity costs 0.002 kcal/mol against grouping by PDB identifier, which localises the split question to the cross-attention ensemble. A second leak sits in the feature pipeline rather than the model: the tree's most important input is itself a cross-validated neural prediction whose folds align with the outer split for only 7.2% of complexes. Regenerating that feature with outer-aligned folds and refitting gives **RMSE 1.945 (r = 0.676)**, the leakage-free estimate for the tabular arm, against 2.067 with the feature dropped — so it contributes 0.122 kcal/mol legitimately and 0.022 came from leakage (Supplementary Note 2).

Accuracy is also not uniform, and the variation is not random. Grouping by PDB identifier separates crystal forms rather than proteins, so a quarter of the evaluation set was scored against targets already seen in another complex: those 555 complexes reach RMSE 1.56 (r = 0.85) against 1.86 (r = 0.70) for the 1,693 whose target is unique to their fold. Clustering targets at 30% sequence identity makes the comparison stricter and widens it — 1.73 (r = 0.79) where the target family recurs across folds, against **1.93 (r = 0.61)** for the 686 complexes whose family appears in no other fold. That last figure is the one describing prospective use against an unfamiliar target, and it is 0.11 kcal/mol worse than the pooled result (Supplementary Note 2).

A peptide screen: ranking is partly recovered, absolute affinity is not

Benchmark accuracy need not translate into useful ranking, so we ranked 414 candidate peptides against EWS-FLI1 (208) and PAX3-FOXO1 (206): 14 literature sequences, 240 ESM-2-generated variants and 160 PepMLM designs, two of which are composition-matched scrambles of literature peptides.

Within this screen the literature sequences rank highly: an EWS-FLI1 peptide and its variants take positions #1, #14 and #15 of 414 while its composition-matched scramble falls to #117 (Supplementary Table S8, Fig. 1b–c), and a method reading only amino acid composition would not separate them. The separation does not hold on the second target, where the PAX3-FOXO1 scramble at #247 sits beside its parent at #246: two controls on two targets, one separating and one not.

The absolute predictions on these peptides are wrong by five orders of magnitude. The top-ranked peptide is predicted at −15.61 kcal/mol, or 3.6 pM, against a literature value of 2.3 μM — an error of 7.9 kcal/mol, and the scramble is predicted tighter than the real peptide's measurement. Neither these sequences nor the EWS-FLI1 construct appear in training, making this a genuine external test: on short peptides against large disordered constructs, outside the applicability domain above, ranking signal partly survives while the absolute scale fails. These values carry no affinity interpretation, and the sequence-to-affinity pairing itself is unverified pending the primary records (Supplementary Table S8).

What transfers to designed binders, and what does not

We next asked how PRESTO discriminates experimentally characterised designed binders, starting with 378 designed EGFR binders from the Adaptyv Bio competition[26]: 53 confirmed binders, 325 non-binders (Fig. 2a).

On this target, sequence-level ESM features outperform structural metrics. Seven span AUC 0.69–0.76, the best being ESM euclidean distance at 0.763, against pLDDT 0.656 and ipTM 0.636 (Supplementary Table S9); PRESTO's calibrated ΔG reaches 0.639, indistinguishable from ipTM. The advantage does not generalise: across the 15 targets of the Overath meta-analysis[33] the same distance has median per-target AUC 0.525 and falls below chance on 5 of 12 scorable targets. We therefore do not recommend a sequence-only pre-screen.

A contemporaneous method sharpens the point. BindPred[38] predicts absolute log10(K_d) from ESM-2 embeddings and reports r = 0.86 under a random split, but PPB-Affinity holds 8,045 entries over only 3,099 unique sequence pairs, so a random split tests on near-identical complexes. On the EGFR designs, external to both methods, it does not discriminate (AUC 0.43) while the euclidean distance between the two halves of its own input vectors recovers 0.767: fitting a regressor to measured affinities discarded discriminative power already present in its representation.

This bounds what affinity regressors, ours included, can do outside the distribution they are fitted to, and it motivates training on design outcomes directly. The designed sets carry binary bind/no-bind labels rather than dissociation constants — only 55 of 1,482 sampled complexes have a measured K_d — so we trained that objective, leave-one-target-out across the 12 Overath targets with at least 45 labelled designs, which is the regime a campaign faces.

The outcome head reaches a median per-target AUC of 0.754 across the 12 targets, against 0.737 for ColabFold ipTM on the same designs. The available summaries differ substantially and are named throughout: the median paired per-target difference against ColabFold ipTM is +0.049, whereas the median difference against a model given ipTM as its only feature is +0.125 (p = 0.005, paired Wilcoxon). Only the former compares against ipTM as a selection score (Supplementary Note 4).

The improvement over ipTM is not established at the level that matters. Within-target resampling measures stability on these twelve targets; a campaign needs to know how the head behaves on a thirteenth. Resampling targets rather than designs, the median paired difference is +0.049 with a 95% interval of [−0.035, +0.068] — including zero — and the head is worse than ipTM on 4 of 12 targets. On the metric closest to practice, precision among the top 10% of a target's designs, ipTM is the better selector (3.19-fold enrichment against 2.81), while average precision favours the head (0.396 against 0.337). These do not support replacing confidence screening with this head; they support treating the two as comparable filters whose relative advantage varies by target in a way twelve targets cannot resolve.

The head does establish that the signal is not a confidence metric predicting itself, which matters because the feature set contains ipTM: removing every structural-confidence feature costs nothing measurable (−0.008, p = 0.47, 149 of 204 features retained). Importance is diffuse — the top five features account for 13.5%, led by interface geometry and packing terms rather than confidence.

The result that bears on practice is the combination. Rank-averaged with ColabFold ipTM within each target, the head reaches 0.783 (95% CI 0.730–0.823) against 0.747 (0.674–0.796) for ipTM alone, better in 90.6% of paired bootstrap resamples. This is the opposite sign to the calibrated ΔG on the same benchmark, which degrades ipTM in 99.6% of resamples. The head alone beats ipTM alone in only 70.2% of resamples, so the result rests on the composition rather than on the head in isolation.

Mutational variants require a separate, interchangeable ΔΔG term

An absolute ΔG cannot say whether a substitution helps or hurts. The quantity is defined per complex, so within a fixed fold the calculator returns one value for every variant of it; the training table reflects this, holding one row per SKEMPI crystal structure — 184 rows across 184 PDB identifiers — with 2,357 mutation measurements reduced to a single label per structure and no wild-type/variant pair ever seen. The features are also the wrong kind, PRESTO's ESM descriptors being whole-sequence pooled statistics whose binder embedding norm varies by 0.2% within a structure while measured ΔG varies by 39–64%.

ΔΔG enters as a separate additive term, ΔG(variant) = ΔG(wild type) + ΔΔG, and that term is interchangeable. On 2,119 variants across 173 complexes with measured variant ΔG, a term assembled from FoldX interaction energies and masked-marginal language model likelihoods lifts pooled Pearson from 0.709 to 0.755; substituting the published specialist RDE-Network[39] gives 0.767, against 0.805 for measured ΔΔG inserted directly. Within a structure the ordering reverses: 0.472 for the assembled term against 0.444 for the specialist. Either serves, so a pipeline can adopt whichever ΔΔG method is current without altering the affinity model.

Two properties make that substitution safe. Physics and evolutionary likelihood are near-orthogonal ($\rho = -0.087$), and their combination is what produces a usable term: median within-structure Spearman 0.442 across 2,147 mutations, against 0.130 for FoldX and 0.109 for ESM-2 alone. And the gap to the specialist is smaller than the literature implies — the 0.74–0.76 commonly quoted for RDE-Network on SKEMPI v2.0 requires pooling across complexes and augmenting with reverse mutations, while its own per-complex Spearman is 0.401 against 0.405 for the assembled term (Supplementary Note 4). Neither is state of the art, and neither needs to be: the term composes and can be replaced.

Prospective design campaigns

On the SARS-CoV-2 receptor-binding domain, standard BindCraft failed completely: across 362 trajectories no design passed the ipTM ≥ 0.5 threshold (peak 0.15–0.21). Of 12 configurations varying hotspot subsets, length and AF2 presets, only the focused 8-residue core set matching the LCB1 footprint[29] produced accepted designs, reaching ipTM 0.82–0.86 and ΔG −11.62 kcal/mol. PRESTO scores LCB1 itself at −10.88 kcal/mol, about 11 nM against a measured affinity below 200 pM — at least fiftyfold weaker than the measurement, an instance of the range compression above rather than a validation. We therefore claim only that a quantitative score differentiated configurations in a regime where every configuration failed the confidence threshold.

That rule does not transfer. On MYC–WDR5 we derived an 8-residue set from the 4Y7R footprint by the same procedure, registering the prediction and its falsifier before running it against a contemporaneous 14-residue control. The focused set produced no accepted designs; the control produced four. Across all seven configurations not one trajectory was rejected for missing its hotspots: here the binding site was never the constraint, steric clash was, and a narrower epitope does not relieve it. The RBD outcome is a property of that target, and affinity-guided reconfiguration is worth testing per campaign rather than assuming.

We then ran campaigns against three undruggable cancer PPIs (Supplementary Table S11, Fig. 3g): p53–MDM2[30] (1YCR, 16 accepted designs, best −11.12 kcal/mol ≈ 7 nM, against ALRN-6924's reported 4.7 nM), MYC–WDR5 (4Y7R, 3 accepted; the closest comparison is Pomplun and colleagues[31], six WBM-site binders at 219–650 nM and a stapled variant at 39 nM), and PCNA–PIP box (7KQ0, 4 accepted after relaxing hydrophobicity filters for that groove).

Across the three targets the best predicted ΔG runs −10.1 to −11.1 kcal/mol, below the −8.19 kcal/mol corresponding to 1 μM. Within each target the two rankings are weakly related and not sign-stable ($\rho = -0.317$

on p53–MDM2, +0.521 on PCNA–PIP; MYC–WDR5's three designs support no estimate), and no pooled correlation is reported, since pooling mixes between-target level differences into a within-target question.

Three properties of the design bound every campaign result here. The configurations varied hotspot sets, lengths, presets and acceptance filters simultaneously, so none isolates the causal contribution of affinity feedback; that requires matched campaigns differing only in the affinity criterion, judged on metrics other than PRESTO. The scores come from a different model than Table 1 — one ICAN configuration inside the design loop, which approximates aligned error from pLDDT as $30 \times (1 - \text{pLDDT}/100)$, biased low by more than half against real interface PAE (Supplementary Note 3) — and that configuration is unbenchmarked. No design here has been synthesised or assayed. These campaigns therefore demonstrate the feasibility of an integrated design loop, not improved binding.

Discussion

Ensembling gradient-boosted trees with cross-attention networks over multiple PLMs improves protein–protein affinity prediction relative to the established contact-based predictor, roughly halving its error on the complexes both methods score. We do not convert that into a fold-uncertainty in K_d: the conversion is arithmetic on an error magnitude rather than a calibrated interval, and the error is neither symmetric nor uniform (Supplementary Note 1). What can be said is narrower and still useful — the model ranks complexes well within the affinity band and target families it was fitted to, while its absolute scale compresses toward the dataset mean at both extremes. The gain comes from components that make partly different errors, though mixing PLMs decorrelates residuals by only ~0.05 in mean r, so most of the ensemble's benefit is averaging.

The most useful negative results concern what does not transfer. On the EGFR benchmark, ESM embedding features discriminate validated binders better than AlphaFold2 confidence (0.69–0.76 against 0.636 for ipTM), while PRESTO's own calibrated ΔG (0.639) does not — and the ESM advantage itself collapses across fifteen targets. This is not isolated: fine-tuning Boltz-2 for protein–protein affinity on PPB-Affinity yields a model that underperforms sequence-based alternatives, its authors concluding that "current structure-based representations are not primed for performant affinity prediction"[37]. An independent group reaching that conclusion from a state-of-the-art structure model is stronger evidence than our EGFR result alone.

PLMs do encode binding-relevant information that structural confidence does not, but the useful form of that claim is narrow. Position-specific likelihoods predict mutational effects at buried interface positions, where confidence metrics are silent; pooled whole-sequence statistics transfer on one target and not across fifteen. Within every target the rank correlation between calibrated ΔG and ipTM is weak and not sign-stable, which invites reading it as unused discriminative information. It is not: combining the two degrades ipTM in 99.6% of paired resamples, and conditioning the tree on ipTM collapses it toward chance. What design pipelines do leave unused is an objective trained on design outcomes rather than dissociation constants, which composes with confidence instead of degrading it — though its advantage is not established for an unseen target. Structural-confidence screening should stay the default for prospective binder selection, with an outcome-trained model alongside it where per-target validation data exist, and an absolute affinity substituting for neither.

Four limitations bound these conclusions. Ultra-tight binders ($\Delta G < -16$ kcal/mol) remain hard: 47 of 2,248 complexes sit there with error 2.5-fold worse, in both components, which indicates a limit of training data at the extreme rather than of either architecture. The method needs a predicted structure but is not tied to the predictor — re-predicting 314 complexes with Boltz-2 leaves accuracy unchanged (2.15 against 2.19, $p = 0.22$). Antibody–antigen and enzyme–inhibitor families dominate the training data. And every prediction here requires experimental validation: no design reported in this paper has been synthesised or assayed.

Methods

Training data and preprocessing

From SKEMPI v2[19] we drew on 2,412 wild-type binding affinity records across 192 crystal structures, converting dissociation constants by $\Delta G = RT \cdot \ln(K_d)$ at a nominal 298 K. Three properties of these labels follow from the extraction record, and each is measured rather than assumed. The training table contains one SKEMPI row per crystal structure — 190 rows, one per PDB identifier, with no mutation column — so a SKEMPI training label is a single affinity per structure rather than a per-variant measurement, and the 2,412 records are the pool those labels derive from. Censoring is rarer than previously reported for this dataset: inequality-bounded measurements are 41 of 7,085 wild-type entries (0.6%) and 159 of 7,085 mutant entries (2.2%) in the SKEMPI v2 release, and none survive into the extraction record, so the earlier figure of 585 truncated entries (24%) is not supported by either source. Temperature is the larger issue: 955 of 2,410 records (40%) were measured away from 298 K, spanning 273–310 K, and converting them at 298 K biases ΔG systematically within each temperature — −0.21 kcal/mol at 293–294 K, +0.35 kcal/mol at 310 K, with a mean absolute error of 0.09 kcal/mol and a maximum of 1.43 across 49 of 194 complexes. Experimental PDB structures were used directly. From PPB-Affinity[20] we obtained 7,892 entries, which contain extensive duplication (PDB 1CHO appears 296 times for the same complex), and deduplicated by grouping identical (PDB ID, ligand sequence, receptor sequence) tuples and taking the median ΔG, yielding 3,126 unique complexes whose structures we predicted with ColabFold[34] (AlphaFold2-Multimer[32], MMseqs2 MSAs). After removing prediction failures the unified dataset held 3,074 complexes spanning $\Delta G = -1.7$ to -21.4 kcal/mol. Identical pipeline code computed every feature for every sample, with one documented exception: the nine AlphaFold2 confidence features do not exist for crystal structures, and for the 190 SKEMPI rows they were assigned fixed values (ipTM and pTM 0.95, PAE 2.0, pLDDT 95.0) rather than computed (Supplementary Note 5).

Features for CatBoost

93 features per complex, in eight groups: PRODIGY-derived interface descriptors (27)[7], inter-chain contact types from pdbe-arpeggio v1.4.4 (15)[21], interface geometry (11), ProtBERT and ProtTrans embedding statistics (10), AlphaFold2 confidence metrics (9), FoldX AnalyseComplex energies (9)[8], ESM-2 650M pair-embedding statistics (8)[28], charge descriptors (3), and one ESM CNN scalar. Supplementary Note 3 lists the full inventory, regenerated from the shipped model's own feature list. Three entries differ from earlier descriptions of this feature set: the ProtBERT and ProtTrans features are present, no ESM-2 principal components exist, and PRODIGY contributes 27 features. SKEMPI contacts were recomputed with this arpeggio version; legacy counts differ by 2–5×.

Two properties of this set matter more than its composition. One feature carries the model: `esm_cnn_cv` accounts for 22.9% of importance and removing it costs 0.079 kcal/mol, whereas removing all 27 PRODIGY features or the entire AlphaFold2 confidence block costs nothing measurable. That feature is itself a learned predictor, and its folds do not match the outer cross-validation (Supplementary Note 2), which is why the leakage-free tabular result excludes or regenerates it.

ICAN architecture and training

Each ICAN (3.87M parameters) has four modules: a residue encoder, physics-biased cross-attention, gated pooling and a prediction head. Per-residue PLM embeddings are projected to 256 dimensions; for ESM-1v, a learned attention pools across the 5 pre-trained models per residue. Two bidirectional cross-attention layers let binder residues attend to target residues and vice versa, with attention logits receiving an additive bias from a 17-dimensional per-residue-pair vector: 14 log1p arpeggio counts, inverse CB–CB distance, inverse

cross-chain PAE and mean pLDDT. The PAE channel was inert for most of the training set: it is read from the AlphaFold2 pickle alone, and where that file had been pruned the extractor's uniform maximum-uncertainty default stood, giving a constant plane for 2,083 of the 2,248 scored complexes, so the reported models were trained with that channel carrying no signal for the large majority of complexes (Supplementary Note 5). Gated pooling compresses variable-length chains to fixed vectors, which are concatenated with 25 scalar features for an MLP head. Interface residues are those within CB–CB 12 Å of the partner chain (relaxed to 24 Å if fewer than 5), capped at 128 per chain. Exact layer dimensions, dropout and activation choices are in Supporting Information and the released code.

Training ran in two phases: 30 epochs with the encoder frozen (AdamW, lr 3×10^{-4}, weight decay 0.05, cosine annealing with warm restarts), then full fine-tuning with the early-stopping patience counter reset, up to 120 epochs with patience 15 on validation RMSE, gradients clipped to 1.0, batch size 16. The default loss was Huber ($\delta = 2.0$ kcal/mol), with variants adding a concordance term or a pairwise ranking term. Augmentations were chain swap ($p = 0.5$), Gaussian noise on embeddings ($\sigma = 0.01$) and random masking of 10% of interface residues. The 14-model ensemble was chosen by greedy forward search from 19 candidates to minimise RMSE on the out-of-fold predictions — the same predictions from which the reported RMSE is computed — and comprises 5 ESM-1v, 6 MINT and 3 SaProt models when counted by language model.

Two properties of this protocol limit what the resulting numbers support. Early stopping monitored RMSE on each fold's held-out half, and the predictions reported for that fold come from the checkpoint scoring best on it, so checkpoint selection and evaluation used the same complexes; the saved predictions cannot bound the resulting optimism. Folds were also grouped by PDB identifier, which does not separate the same complex deposited under two identifiers: grouping by canonical biological identity — the order-insensitive pair of chain sequences, collapsing chain-swapped deposits — finds 73 identity groups spanning more than one PDB identifier. The released refit driver corrects both, splitting each outer fold's training indices again by identity group, early-stopping on that inner split alone, and scoring the outer fold exactly once.

Hybrid ensemble, calibration and evaluation

The ensemble switches on the gradient-boosted model's own output, which requires no measured quantity: where $\Delta G\_CatBoost \geq -16$ kcal/mol, $\Delta G\_raw = 0.22 \times \Delta G\_CatBoost + 0.78 \times \text{mean}(\Delta G\_ICAN)$; below it, where the cross-attention models have few training examples (47 complexes in the training set sit there), the tree prediction is used alone. This is the rule implemented in the deployed scorer. An earlier evaluation applied the threshold to the measured affinity instead, which is target leakage; on the same saved predictions the two rules differ by 0.001 kcal/mol (1.7932 against 1.7941) because they route 39 complexes differently, and the switch is worth 0.006 relative to applying the hybrid everywhere. We report the deterministic rule throughout and treat the branch as an implementation convenience rather than a contribution. The weight was grid-searched in steps of 0.01 on the out-of-fold predictions. A post-hoc linear calibration corrects ~14% range compression: $\Delta G\_calibrated = 1.139 \times \Delta G\_raw + 1.387$. Weight, slope, intercept and ensemble membership are single values fitted on the full out-of-fold set rather than per fold, on the same predictions from which the headline RMSE is measured; we quantified the resulting optimism across 10 random 80/20 PDB holdout splits with weight and calibration refitted on the development half (dev–test gap -0.04 ± 0.08 kcal/mol leak-free, $+0.00 \pm 0.11$ overall; $w = 0.218 \pm 0.020$, slope 1.125 ± 0.012; test RMSE 1.68–1.87). Ensemble membership and the trained models were not re-selected per split, so this bounds the optimism from weight and calibration only.

Evaluation used PDB-disjoint 5-fold cross-validation (GroupKFold by PDB identifier), separating crystal-packing artifacts, symmetry mates and homologous chains within one crystal. It does not separate the same complex deposited twice: 46 identical (binder, target) sequence pairs covering 129 samples span folds, and

94 carry a ΔG within 0.1 kcal/mol of their duplicate, placing the label in the training set. Reported metrics exclude those 94 unless stated. CatBoost and all ICANs used identical fold assignments. Metrics are RMSE and Pearson r, with significance by paired test on per-sample errors. The measurement floor is the within-PDB-entry ΔG spread, 0.48 kcal/mol RMSE over 85 entries with repeated measurements, computed after median-ΔG deduplication and therefore a lower bound.

Baselines

PRODIGY[7] ran with default parameters (IC-based model) on both structure sources: SKEMPI crystal structures (180 of 192 PDBs produced a prediction) and ColabFold structures for PPB-Affinity. Table 1 reports the sources separately and restricts comparison to complexes scored by both methods, 174 and 909. **FoldX 5.1**[8] ran AnalyseComplex with default force field parameters on the 192 SKEMPI structures (three seeds × five folds), giving RMSE 8.08 ± 1.87 and r = 0.05 ± 0.17. Because FoldX targets ΔΔG rather than absolute ΔG, we confirmed this on the present dataset by re-running 190 structures through the documented RepairPDB-then-AnalyseComplex protocol: interaction energies correlate with measured ΔG at r = −0.06 (Spearman −0.03), indistinguishable from zero. FoldX is therefore absent from Table 1, a baseline at zero correlation being uninformative.

Peptide screen, EGFR benchmark and ΔΔG term

The 414 peptides comprised 23 literature sequences (including E9R and variants[25,36]), 187 ESM2-generated variants sampled at interface positions with ESM-2 650M, 196 PepMLM designs, and 8 scrambled controls. Complexes were predicted with ColabFold and ranked by predicted ΔG.

The EGFR benchmark used 378 designed binders from the Adaptyv Bio competition[26], experimentally tested by bio-layer interferometry (53 binders, 325 non-binders). We computed PRESTO sequence features for all 378; structural metrics were taken from the competition's ColabFold predictions where available (n = 44 complete). AUC came from scikit-learn, with distance-like features negated so lower values mean tighter predicted binding.

The ΔΔG term combines FoldX ΔΔG_binding — AnalyseComplex interaction energy differenced between wild-type and rebuilt mutant structure, distinct from FoldX's folding ΔΔG — with masked-marginal likelihood ratios log P(mutant) − log P(wild type) at the mutated position from ESM-1v and ESM-2, plus substitution and interface-location descriptors. Multi-point variants sum per-position ratios, an additive approximation that ignores epistasis. Evaluation is GroupKFold by structure with the median within-structure Spearman as headline. RDE-Network[39] was scored from its released per-fold checkpoint using its own evaluation script, which pairs each fold with its own validation split by complex; predictions were joined to ours one-to-one after collapsing SKEMPI's repeated measurements of the same variant (892 of 7,085 keys, one appearing 19 times).

Design campaigns

We integrated ICAN v34c into BindCraft[14] for affinity scoring inside the design loop, run on CPU to avoid GPU contention with AlphaFold2 (~5–10 s per design, dominated by ESM-1v inference). Per design we extracted per-residue ESM-1v embeddings, CB coordinates and pLDDT from the AF2 PDB, PAE estimated as 30 × (1 − pLDDT/100), and inter-chain contacts via pdbe-arpeggio; the 5-fold ensemble prediction was averaged and calibrated. Three pediatric campaigns targeted FLI1 ETS (5E8G, n = 635), DHX9 HA2 (8SZP, n = 786) and CBP KIX (2LQH, n = 208) under BindCraft's default 4-stage multimer protocol with ProteinMPNN sequence optimization.

For the cancer targets, structures were prepared by extracting the target chain and removing co-crystallized peptides: MDM2 chain A from 1YCR (85 residues), WDR5 chain A from 4Y7R (residues 31–334), PCNA chain A from 7KQ0 (257 residues). Hotspots came from inter-chain contacts computed with BioPython NeighborSearch at 5 Å (core) and 8 Å (extended). Four configurations per target varied hotspot subset, binder length (40–80 residues) and AF2 preset. Filters were relaxed relative to standard BindCraft (ipTM $\geq$ 0.3, pTM $\geq$ 0.45, shape complementarity $\geq$ 0.45, iPAE $\leq$ 0.45, unsatisfied H-bonds $\leq$ 10, binder pLDDT $\geq$ 0.7, binder RMSD $\leq$ 6.0 Å), with hydrophobicity and lysine limits relaxed further for PCNA. Each configuration ran 20 trajectories with a 2-hour timeout.

For SARS-CoV-2 RBD (chain E of 6M0J) we explored 12 configurations spanning hotspot subsets (8 core residues matching the LCB1[29] footprint; 14 medium; all 25 ACE2-contacting; none), binder lengths (40–55, 55–70, 70–90 aa) and AF2 presets (default, hardtarget with predict_initial_guess, and beta-sheet-biased), under the same relaxed filters, for 240 trajectories plus 122 from expanded hotspot configurations.

Computation, software and availability

Feature extraction takes ~25 minutes per complex, dominated by ColabFold on an RTX 3090; PLM embedding extraction 3–15 s; CatBoost inference < 0.1 s; ICAN inference ~0.1 s per model on GPU. Software: Python 3.10, PyTorch 2.1, TensorFlow 2.20 (CNN models only), CatBoost 1.2, fair-esm, pdbe-arpeggio 1.4.4, gemmi, BioPython, scikit-learn, scipy, ColabFold 1.5.5. Hardware: NVIDIA RTX 3090 and RTX 4090 (24 GB).

A reproduction repository with the data and analysis code needed to rebuild every table and figure is at https://github.com/omic/presto-reproduction: the training set with all 93 features, out-of-fold predictions from all 14 ICANs and CatBoost, the peptide screen, the EGFR benchmark, per-design scores from every campaign, and the scripts that read those files and write the reported results. Every number in Tables 1 and 2 was recomputed from those predictions by the distributed evaluation script. The PRESTO implementation and trained weights will be released under an MIT licence on acceptance; they are held privately during review and are available to editors and reviewers on request.

One category of data is withheld: the amino-acid sequences of binders designed in this work, which are the subject of active drug discovery programmes at Omic Inc. For every such design we release the identifier, target, configuration, length, PRESTO ΔG, ipTM, pTM, pLDDT, interface PAE and shape complementarity — all quantities from which the reported results are computed — and Tables 4–6 and Figure 3 rebuild from these columns alone. Sequences may be made available for non-commercial research under a material transfer agreement; requests to the corresponding author. The accuracy results, the paper's principal claim, involve no withheld data: SKEMPI v2, PPB-Affinity and the 378 EGFR binders are public and released with sequences intact.

Tables

Table 1. Performance comparison on absolute ΔG prediction

Method	Type	n	RMSE (kcal/mol)	Pearson r	p vs PRESTO
Crystal structures, both methods scored					
PRODIGY	Contact	174	3.87	0.32	7×10^{-16}
PRESTO (hybrid)	**Tree + Attention**	174	**1.78**	**0.88**	—

Method	Type	n	RMSE (kcal/mol)	Pearson r	p vs PRESTO
ColabFold-predicted structures, both methods scored					
PRODIGY	Contact	909	5.33	0.20	4×10^{-60}
PRESTO (hybrid)	**Tree + Attention**	909	**1.91**	**0.69**	—
PRESTO and its components, full evaluation					
Mean baseline	Null	2,248	2.73	—	
CatBoost only	Tree	3,074	1.91	0.69	
CatBoost only	Tree	2,248	1.92	0.71	
ICAN ensemble (14 models)	Attention	2,248	1.84	0.75	
PRESTO (hybrid), including cross-fold duplicates	Tree + Attention	2,248	1.79	0.75	
PRESTO (hybrid), label-similar duplicates excluded	Tree + Attention	2,154	1.82	0.72	—
PRESTO (hybrid), canonical-identity exclusion	**Tree + Attention**	2,110	**1.823**	**0.72**	—
CatBoost, refitted: inner-val early stopping, canonical grouping	Tree	3,074	1.923	0.685	—
CatBoost, refitted and leakage-free (CNN feature regenerated)	**Tree**	3,074	**1.945**	**0.676**	—
CatBoost, refitted with the CNN feature dropped	Tree	3,074	2.067	0.622	—

5-fold cross-validation grouped by PDB identifier on the unified SKEMPI + PPB-Affinity dataset. Every row names its own n because they are not the same set; the two paired blocks are the only comparisons in which both methods see the same complexes, and p is a paired Wilcoxon test on absolute error. ***All PRESTO and ICAN rows are development estimates*** *— each ICAN's checkpoint was selected on the fold scored here, so differences between rows are meaningful while their absolute level is optimistic. The two exclusion rows differ in criterion rather than degree, and PRODIGY supplies 27 of PRESTO's own*

features (Supplementary Notes 2, 3). Recomputed by `code/eval_headline_metrics.py`*; audited by* `code/build_jcim_review_audit.py`*.*

Table 2. Ablation study

Component	RMSE	r	Δ RMSE
CatBoost only (93 features)	1.92	0.71	+0.12
ICAN ensemble (14 models)	1.84	0.75	+0.04
ESM-1v ICANs only (5 models)	1.92	0.72	+0.12
MINT ICANs only (6 models)	1.86	0.73	+0.07
SaProt ICANs only (3 models)	1.92	0.71	+0.13
Hybrid, uncalibrated	1.81	0.75	+0.02
Hybrid + calibration	**1.79**	**0.75**	—

Every row on the same 2,248 complexes, which includes the 94 cross-fold duplicates so the ablation compares like with like; the figure after strict canonical-identity exclusion is 1.823 over 2,110 complexes (Table 1). Recomputed from the saved out-of-fold predictions by `code/eval_headline_metrics.py`*.* ***All rows are development estimates:*** *each ICAN's checkpoint was selected on the fold these predictions come from, so the comparison between rows is sound but their absolute level is optimistic by an unbounded amount. The gradient-boosted row refitted with inner-validation early stopping and canonical grouping is 1.923 (*`results/catboost_nested_refit_predictions.csv`*).*

Figure Legends

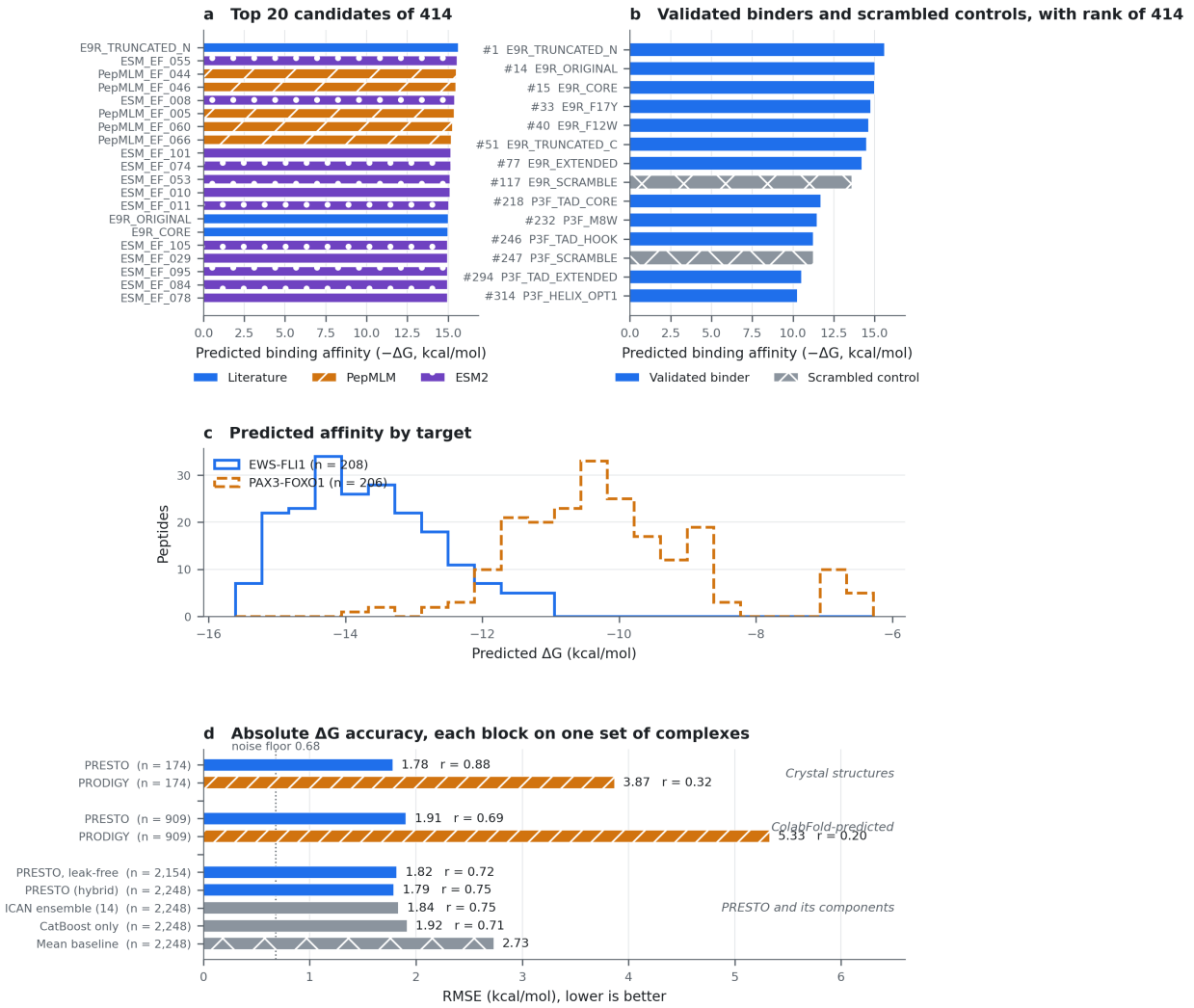

Figure 1. Peptide screen and absolute ΔG accuracy. (a) The 20 highest-affinity of 414 screened peptides, coloured by source. (b) The 14 literature peptides with their rank in the full screen: E9R[25] and its variants take #1, #14 and #15 and the scrambled E9R control falls to #117, while for PAX3-FOXO1 the scrambled control at #247 is not separated from its parent at #246, so the control works on one target and not the other. (c) Predicted ΔG by target. (d) Absolute ΔG accuracy; each block is measured on one set of complexes, the two paired blocks are the only comparisons in which PRODIGY and PRESTO scored the same complexes, and the dotted line is the 0.68 kcal/mol measurement noise floor. Drawn by `code/plot_figure1.py`.

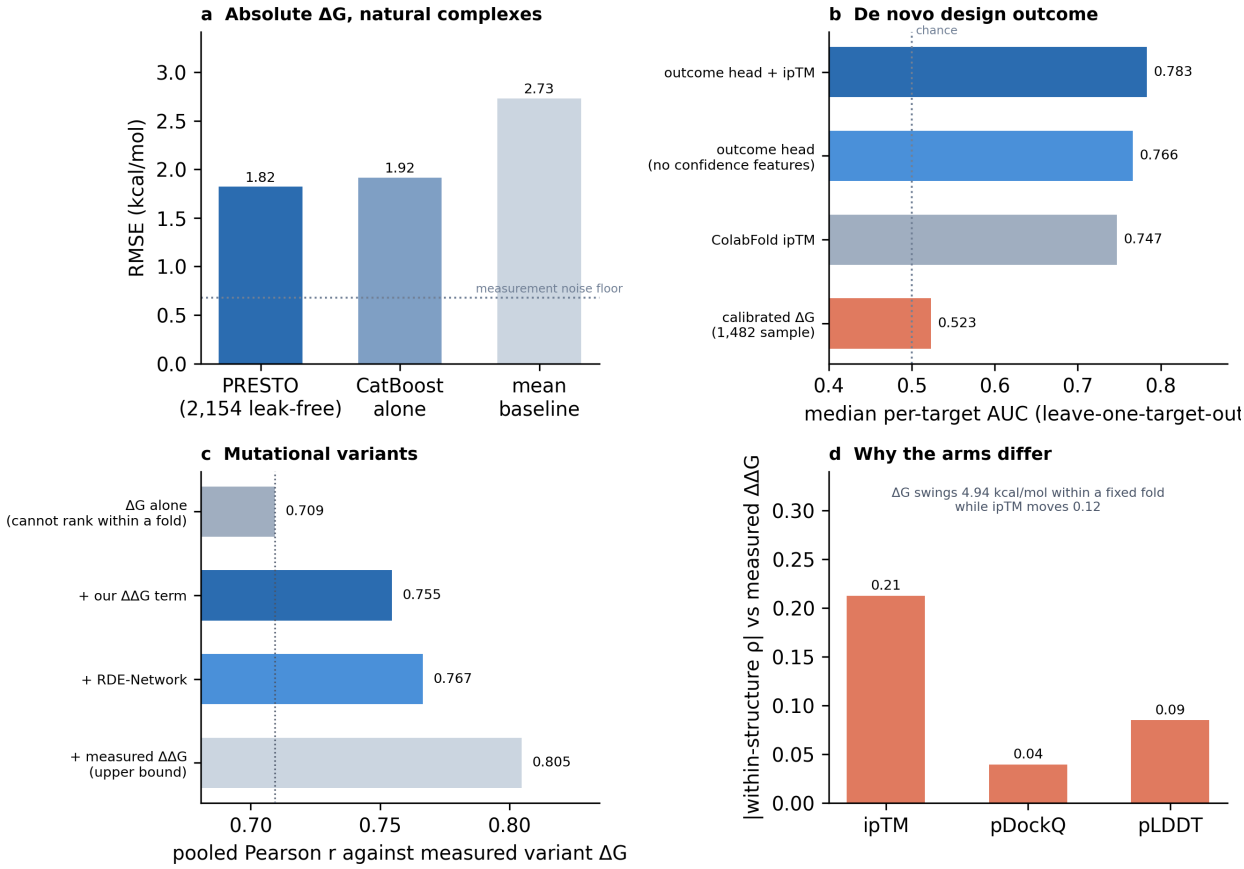

Figure 2. Coverage across the three binding questions. (a) Absolute ΔG on natural complexes: RMSE for PRESTO and PRODIGY on the complexes both score, 174 crystal and 909 predicted structures, with the noise floor marked. (b) Design outcome: median per-target AUC, leave-one-target-out over 12 targets, for the outcome head with all confidence features removed, ColabFold and Boltz-1 ipTM, the rank-averaged composition, and the calibrated ΔG, with 95% intervals from 2,000 within-target resamples. (c) Mutational variants: pooled Pearson against measured variant ΔG over 2,119 variants for ΔG alone, ΔG plus our ΔΔG term, ΔG plus RDE-Network, and ΔG plus measured ΔΔG as an upper bound. (d) Why the arms differ: within-structure Spearman against measured ΔΔG for ipTM, pDockQ and pLDDT beside the median within-structure ΔG range — confidence moves ~0.12 while affinity swings 4.94 kcal/mol at fixed fold. Drawn by `code/plot_figure2.py`.

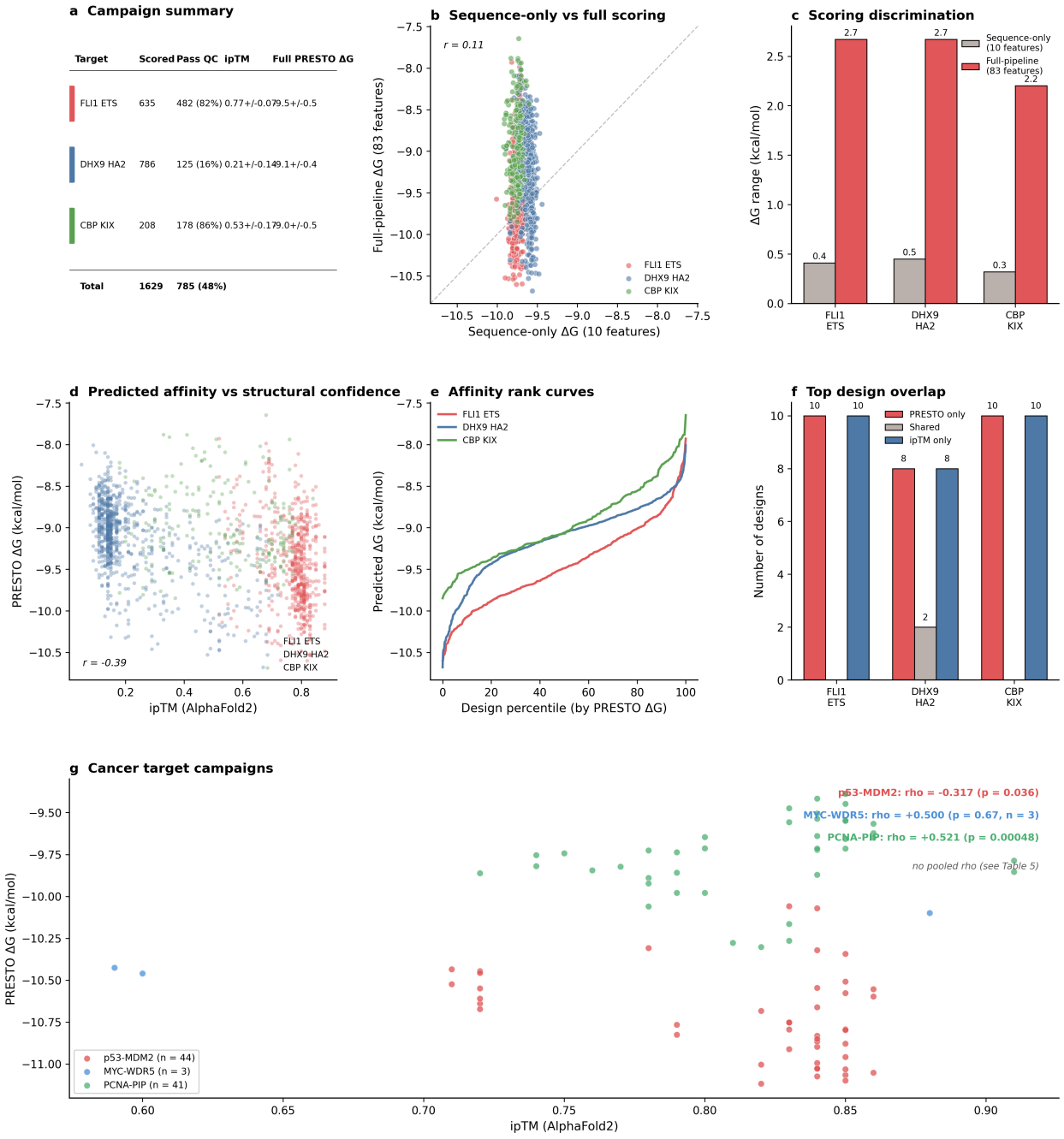

Figure 3. Prospective binder design campaigns. Drawn by `code/figure_binder_designs.py` from `data/binder_designs/`. (a) The three pediatric campaigns: 635, 786 and 208 designs, 1,629 in total, with the fraction passing QC and mean ipTM and ΔG. (b) Sequence-only ΔG (the 10 features available inside the design loop) against full-pipeline ΔG: nearly unrelated, so the score used during design is not the score used to rank. (c) Dynamic range of each: 2.2–2.7 kcal/mol per target against 0.3–0.5 for sequence-only. (d) Full-pipeline ΔG against ipTM by target; the pooled appearance is driven by targets sitting at different levels on both axes, and the within-target correlations are those in Supplementary Table S10. (e) Designs ranked by ΔG within each target. (f) Top-ten overlap between PRESTO and ipTM: 0/10 for FLI1 ETS and CBP KIX, 2/10 for DHX9 HA2. (g) The three cancer campaigns, ΔG against ipTM for the 88 designs carrying both, one ρ per target and none pooled. No design here has been assayed, so the sign of a correlation does

not establish which ranking would have selected better binders.

Associated Content

Supporting Information. The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim

Error distribution, calibration intervals and the withdrawn noise floor; split identity, exclusion criteria and development estimates; feature inventory and feature-block ablations; design-arm estimators with per-target results; training dataset, out-of-fold predictions, method comparison, binder design scores, EGFR benchmark and per-model ICAN predictions (Tables S1–S7); ensemble ablation, error across the ΔG range, CatBoost feature importance and ICAN residual diversity (Figures S1–S4) (PDF)

Funding

This work was funded by Omic Inc. The funder had no role in the design of the study, the analysis, or the decision to publish.

Acknowledgements

We thank the maintainers of SKEMPI, PPB-Affinity and the RCSB PDB, whose openly published measurements and structures this work is built on.

Author contributions

All authors contributed to the design of the study, the analysis and the manuscript, and approved the submitted version.

Competing interests

G.R. is the founder of Omic Inc. All other authors are employees of Omic Inc. and declare no other competing interests.

References

1. Dang, C.V. et al. Drugging the 'undruggable' cancer targets. Nat. Rev. Cancer 17, 502–508 (2017).
2. Lu, H. et al. Recent advances in targeting the 'undruggable' proteins. Signal Transduct. Target. Ther. 8, 393 (2023).
3. Muttenthaler, M. et al. Trends in peptide drug discovery. Nat. Rev. Drug Discov. 20, 309–325 (2021).
4. Scott, D.E. et al. Fragment-based drug discovery supports drugging 'undruggable' protein-protein interactions. Trends Biochem. Sci. 48, 339–352 (2023).
5. Al Musaimi, O. et al. 2024 FDA TIDES (peptides and oligonucleotides) harvest. Pharmaceuticals 18, 291 (2025).
6. Wang, L. et al. Therapeutic peptides: current applications and future directions. Signal Transduct. Target. Ther. 7, 48 (2022).
7. Xue, L.C. et al. PRODIGY: a web server for predicting the binding affinity of protein-protein complexes. Bioinformatics 32, 3676–3678 (2016).
8. Schymkowitz, J. et al. The FoldX web server: an online force field. Nucleic Acids Res. 33, W382–W388 (2005).
9. Stranges, P.B. & Kuhlman, B. A comparison of successful and failed protein interface designs. Protein Sci. 22, 74–82 (2013).

10. Myung, Y. et al. AREA-AFFINITY: machine learning-based prediction of protein-protein binding affinities. J. Chem. Inf. Model. 63, 90–97 (2023).
11. Wang, Y. et al. PPAC: predicting protein-protein affinity changes using protein large language models. ACS Omega (2025).
12. Kim, J. et al. BindPred: predicting protein-protein binding affinity from language model embeddings. bioRxiv 2025.09.27.678407 (2025).
13. Qian, Y., Yang, C., Duan, L., Wang, J. & Qi, Y. PPAP: protein-protein affinity predictor with interfacial contact-aware attention. J. Chem. Inf. Model. 65, 9987–9998 (2025).
14. Pacesa, M. et al. One-shot design of functional protein binders with BindCraft. Nature 646, 483–492 (2025).
15. Watson, J.L. et al. De novo design of protein structure and function with RFdiffusion. Nature 620, 1089–1100 (2023).
16. Erkizan, H.V. et al. A small molecule blocking oncogenic protein EWS-FLI1 interaction with RNA helicase A inhibits growth of Ewing's sarcoma. Nat. Med. 15, 750–756 (2009).
17. Gryder, B.E. et al. PAX3-FOXO1 establishes myogenic super enhancers and confers BET bromodomain vulnerability. Cancer Discov. 7, 884–899 (2017).
18. Harlow, M.L. et al. One oncogene, several vulnerabilities: EWS/FLI targeted therapies for Ewing sarcoma. Mol. Ther. Oncolytics 24, 313–327 (2022).
19. Jankauskaitė, J. et al. SKEMPI 2.0: an updated benchmark of changes in protein-protein binding energy. Bioinformatics 35, 462–469 (2019).
20. Liu, H. et al. PPB-Affinity: protein-protein binding affinity dataset for AI-based protein drug discovery. Sci. Data 11, 1316 (2024).
21. Jubb, H.C. et al. Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. J. Mol. Biol. 429, 365–371 (2017).
22. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. NeurIPS 34 (2021).
23. Ullanat, V., Jing, B., Sledzieski, S. & Berger, B. Learning the language of protein-protein interactions. Nat. Commun. 16, 2688 (2025).
24. Su, J. et al. SaProt: protein language modeling with structure-aware vocabulary. In Proc. ICLR 2024 (spotlight).
25. Barber-Rotenberg, J.S. et al. Single enantiomer of YK-4-279 demonstrates specificity in targeting the oncogene EWS-FLI1. Oncotarget 3, 172–182 (2012).
26. Overath, T. et al. Crowdsourced protein design: lessons from the Adaptyv EGFR binder competition. bioRxiv 2025.04.17.648362 (2025).
27. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. PNAS 118, e2016239118 (2021).
28. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130 (2023).
29. Cao, L. et al. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. Science 370, 426–431 (2020).
30. Saleh, M.N. et al. Phase 1 trial of ALRN-6924, a dual inhibitor of MDMX and MDM2, in patients with solid tumors and lymphomas bearing wild-type TP53. Clin. Cancer Res. 27, 5236–5247 (2021).
31. Pomplun, S.J. et al. Evaluating BindCraft for generative design of high-affinity peptides. ACS Chem. Biol. (2025). doi:10.1021/acschembio.5c00774
32. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. bioRxiv 2021.10.04.463034 (2022).
33. Overath, T. & Rygaard, A. Predicting experimental success in de novo binder design: a meta-analysis of 3,766 experimentally characterised binders. bioRxiv 2025.08.14.670059 (2025).

34. Mirdita, M. et al. ColabFold: making protein folding accessible to all. Nat. Methods 19, 679–682 (2022).
35. Bryant, P. et al. Improved prediction of protein-protein interactions using AlphaFold2. Nat. Commun. 13, 1265 (2022).
36. Selvanathan, S.P. et al. Targeted therapy for EWS-FLI1 in Ewing sarcoma. Cancers 15, 4035 (2023).
37. King, J., Cornwall, L., Nica, A.C., Day, J., Sim, A., Dalchau, N., Wollman, L. & Meyers, J. On fine-tuning Boltz-2 for protein-protein affinity prediction. arXiv 2512.06592 (2025). Machine Learning in Structural Biology (MLSB) 2025.
38. Piao, H., Boorla, V.S., Santra, S. & Maranas, C.D. BindPred: a framework for predicting protein–protein binding affinity from language model embeddings. Bioinformatics 42, btag309 (2026). doi:10.1093/bioinformatics/btag309
39. Luo, S. et al. Rotamer density estimator is an unsupervised learner of the effect of mutations on protein-protein interaction. In Proc. ICLR 2023.