Abstract

Generative and deep-learning tools for protein–ligand modelling are increasingly used, but it is unclear whether they can reliably detect structural changes caused by clinical mutations and how an inhibitor engages its target. This study benchmarks co-folding, docking, and affinity-rescoring tools on PIK3CA, the catalytic p110α subunit of PI3Kα and one of the most frequently mutated oncogenes in human cancer. A 300-pair cross-docking matrix was built from fourteen clinical p110α variants and the wild type, with twenty ligands spanning FDA-approved inhibitors, clinical candidates, allosteric and isoform-selective compounds, and inactive controls. Performance was assessed on four independent axes: experimental affinity ranking, cross-docked specificity, mutation-driven affinity change (Δp), and pose correctness against crystallographic references.

GatorAffinity rescoring on Boltz-2 poses gave the best affinity ranking (Spearman $\rho = 0.85$) and the best separation of binders from controls (AUROC = 0.88). CheapNet ($\Delta p\ \rho = 0.75$) and GatorAffinity ($\Delta p\ \rho = 0.67$) on Boltz-2 poses both signed all five significant mutational shifts correctly, so current models partly capture mutation-dependent changes (preliminary). Performance depended heavily on the pose source, and some generative models produced chemically invalid geometries.

Most critically, every pose tool placed the allosteric inhibitors in or beside the orthosteric ATP pocket (LRMSD > 19 Å), often with high confidence, mirroring the near-absence of allosteric PIK3CA structures in training data. Mutation-aware affinity is partly captured; allosteric pocket recognition remains an open problem.

Chapter 1: Introduction

1.1 Clinical and Biological Importance

PIK3CA encodes the catalytic p110α subunit of class IA phosphoinositide 3-kinase (PI3Kα), a lipid kinase that phosphorylates phosphatidylinositol-4,5-bisphosphate (PIP2) to generate phosphatidylinositol-3,4,5-trisphosphate (PIP3). This single phosphorylation event recruits AKT to the plasma membrane and activates a downstream signalling cascade governing cell growth, survival, metabolism, and proliferation. Under normal physiology, the pathway is tightly regulated through the autoinhibitory activity of the p85 regulatory subunit and the opposing phosphatase PTEN; in cancer, recurrent somatic mutations in PIK3CA render p110α constitutively active, decoupling growth signals from their upstream regulatory inputs. The oncogenic relevance of PIK3CA was established by (Samuels et al., 2004), who identified high-frequency somatic mutations across colon, breast, brain, and lung cancers. The mutations were clustered in two hotspot regions of p110α corresponding to the helical and kinase domains. Structural characterization by (Huang et al., 2007) subsequently showed how these alterations modify the enzyme at the molecular level. The most common substitution, H1047R, increases the membrane affinity of the kinase domain and expands accessibility of the catalytic site. These findings established PIK3CA as one of the most frequently mutated oncogenes in human cancer and a high-priority therapeutic target.

Drug discovery targeting PI3Kα has evolved through three broad phases. Early pan-PI3K inhibitors; including copanlisib and buparlisib, blocked all four class I isoforms (α, β, δ, γ), confirming that the pathway was druggable but produced metabolic toxicity through off-target suppression of insulin signaling. The field moved to isoform-selective compounds, with alpelisib (BYL719) becoming the first FDA-approved PI3Kα inhibitor in 2019, validated by the phase III SOLAR-1 trial in hormone receptor-positive, HER2-negative advanced breast cancer with PIK3CA mutations (André et al., 2019). More recently, inavolisib (GDC-0077) improved on this approach by coupling potent PI3Kα inhibition with selective degradation of mutant p110α protein; the INAVO120 trial demonstrated it.

Despite these advances, a fundamental limitation persists. Orthosteric ATP-site inhibitors suppress both mutant and wild-type PI3Kα, causing dose-limiting hyperglycaemia by disrupting insulin-mediated glucose uptake in non-tumour tissue. This toxicity restricts dosing and, in some patients, prevents full pathway inhibition. (Buckbinder et al., 2023a; Hopkins et al., 2018; Varkaris, Fece de la Cruz, et al., 2024).

A third generation of compounds addresses this by exploiting structural distinctions between wild-type and mutant p110α. RLY-2608, identified through a DNA-encoded library screen combined with cryo-EM structure, binds a novel allosteric pocket accessible only in PIK3CA-mutant conformations. In clinical proof-of-concept, it demonstrated good tumour responses while preserving normal insulin signalling (Varkaris, Fece de la Cruz, et al., 2024; Varkaris, Pazolli, et al., 2024). STX-478, a structurally distinct allosteric inhibitor, achieves 14-fold selectivity for helical and kinase-domain PI3Kα mutants over the wild-type enzyme, eliminating the metabolic adverse effects that limit first-generation agents (Buckbinder et al., 2023b).

This structural and clinical diversity defines the central challenge for computational modelling. Two distinct questions arise depending on binding mode: whether predicted affinity scores track experimental potency across wild-type and mutant receptor forms, and whether generated poses identify the correct binding site.

1.2 Structure of the PIK3CA Catalytic Subunit

Phosphoinositide 3-kinases are grouped into three structural classes (I, II, III) based on substrate specificity and domain organisation. Class I enzymes are the most directly relevant to cancer biology and are further divided into class IA (p110α, p110β, and p110δ, each paired with a p85-family regulatory subunit) and class IB (p110γ, which pairs with p101 or p84). All four isoforms share a conserved C-terminal kinase domain containing an ATP-binding cleft and a hinge region, a short backbone segment whose carbonyl and amide groups form hydrogen bonds with the adenine ring of ATP and, by extension, with ATP-competitive inhibitors (Vanhaesebroeck et al., 2010). Despite this conserved scaffold, the isoforms differ in tissue distribution, regulatory mechanism, and the geometry of the selectivity pocket flanking the ATP cleft, differences that have been exploited to develop isoform-selective inhibitors with reduced off-target effects (Knight et al., 2006). Among the

four catalytic isoforms (Figure 1.1, Figure 1.2), p110α is the most frequently mutated in solid tumours and is the exclusive catalytic target of all compounds evaluated in this study.

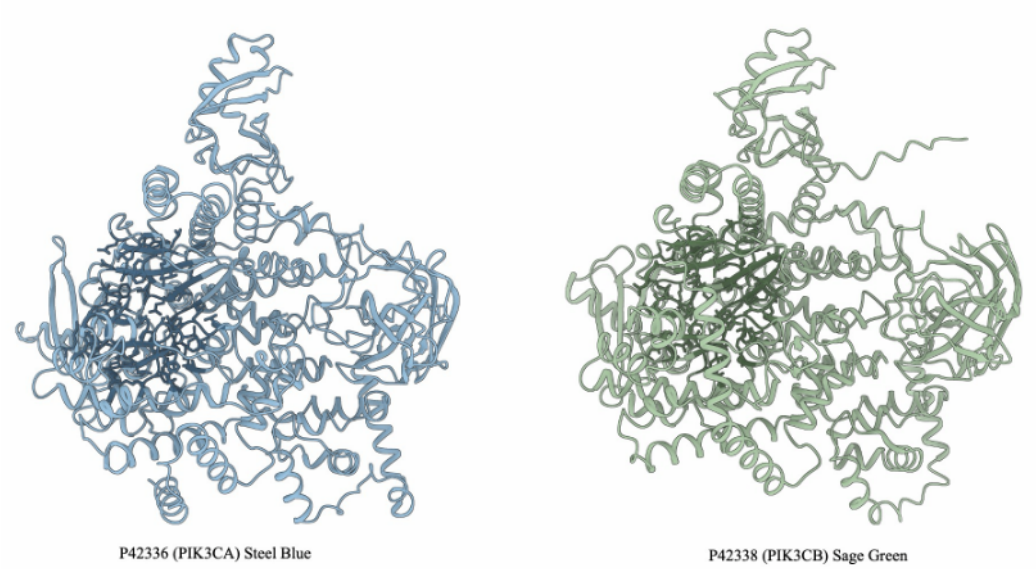


Figure 1.1. Ribbon representations of the class I PI3K catalytic isoforms p110α and p110β. Left: p110α (PIK3CA, UniProt P42336, steel blue), Right: p110β (PIK3CB, UniProt P42338, sage green), a class IA isoform sharing the conserved kinase domain and hinge region with p110α but differing in the geometry of the selectivity pocket, regulatory subunit interactions, and tissue-level activation context. Structures generated by AlphaFold3 and rendered in ChimeraX.

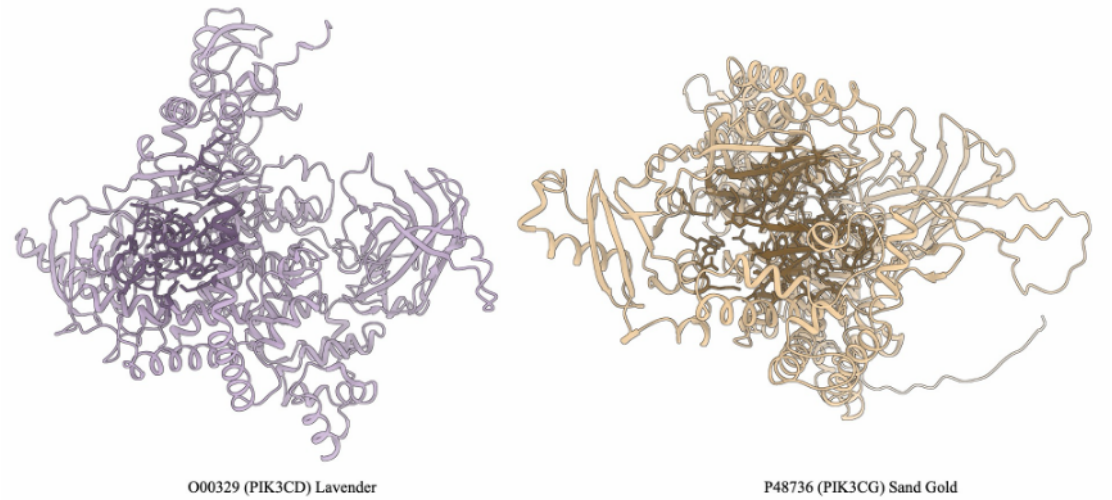


Figure 1.2. Ribbon representations of the class I PI3K catalytic isoforms p110δ and p110γ. Left: p110δ (PIK3CD, UniProt O00329, lavender), a class IA isoform. Right: p110γ (PIK3CG, UniProt P48736, sand gold) Despite sharing the conserved kinase scaffold with p110α, all four isoforms differ at the selectivity pocket bordering the ATP cleft, providing the structural basis for isoform-selective inhibitor development. Structures generated by AlphaFold3 and rendered in ChimeraX.

The catalytic p110α subunit (UniProt P42336) is a 1,068-residue, multi-domain enzyme. From N- to C-terminus, it comprises five structurally and functionally distinct domains. The adaptor-binding domain (ABD, residues 1–108) forms an obligate, high-affinity interface with the inter-SH2 coiled-coil domain of the p85 regulatory subunit, keeping p110α structurally stable and catalytically suppressed under basal conditions. The Ras-binding domain (RBD, residues 140–262) mediates membrane recruitment by engaging GTP-bound Ras G-proteins following receptor tyrosine kinase activation. The C2 domain (residues 300–478) provides a membrane-interaction surface through contacts with acidic phospholipids, supplementing Ras-dependent localisation. The helical domain (residues 517–696) acts as a structural scaffold and forms the autoinhibitory contact with the nSH2 domain of p85α, the interaction that restrains constitutive catalytic activity. The kinase domain (residues 697–1068) contains the ATP-binding cleft and carries out the phosphorylation of PIP2.

The fifteen clinical mutation variants evaluated in this benchmark span all five domains (Table 3.1) and act through mechanistically distinct pathways, some disrupt regulatory contacts far from the ligand-binding site, others alter the ATP pocket directly, and others shift membrane affinity or electrostatic surface properties. This heterogeneity makes the panel a good stress test: a tool that produces plausible scores across variants with such different structural consequences is being asked to generalize well beyond the narrow orthosteric geometry it was likely trained on. It also allows the benchmark to probe whether predicted scores track the underlying physics of each variant class or simply reflect the dominant signal from the ligand.

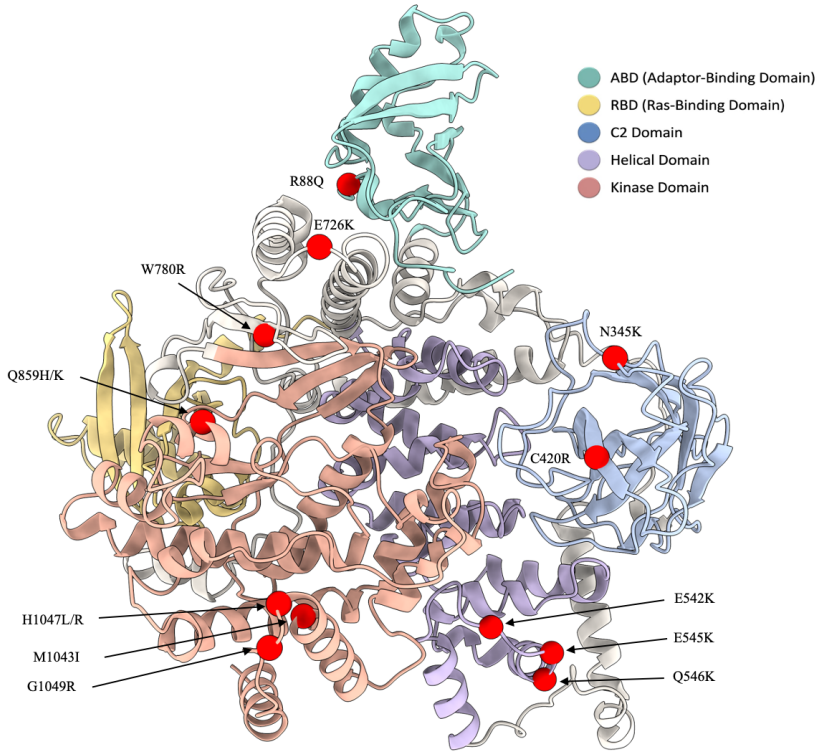


Figure 1.3 Mutations in different domains of PIK3CA

The clinical mutations evaluated in this benchmark are distributed across multiple structural domains of p110α (Figure 1.3). This study models p110α as an isolated catalytic subunit, without the p85 regulatory subunit, for several convergent reasons. All six crystal structures used as strict structural references in this benchmark, 7PG6 (alpelisib), 8AM0 (inavolisib), 8EXL (taselisib), 8TSD (RLY-2608), 8TDU and 8TGD (STX-478) confirm that inhibitor contacts are exclusively within the kinase domain of p110α; p85 does not form part of the drug-binding surface for any compound in this study. The inhibitors in the panel target either the ATP-binding cleft or the allosteric pocket on the kinase domain C-lobe, both of which are physically separated from the ABD–iSH2 and helical–nSH2 regulatory interfaces. The computational tools used and the rescoring models works better on single-chain receptor inputs and therefore p85 subunit wasn't taken into consideration.

1.3 The Computational Challenge

Mutation-aware binding prediction is a more demanding task than conventional docking. In standard docking, a ligand is placed into a single receptor structure and scored; success means reproducing a known crystallographic pose or ranking a known

binder above random decoys. In mutation-aware prediction, the comparison of interest is conditional: the relevant question is whether the same ligand changes recognition when the protein sequence is altered by a clinical mutation. This requires the model to encode sequence-to-structure-to-affinity relationships and not simply to recognize a familiar chemical scaffold in a familiar pocket.

Recent advances in structural Artificial Intelligence(AI), AlphaFold 2 established near-experimental accuracy for single-chain protein structure prediction, and AlphaFold 3 extended to biomolecular interactions including protein–ligand complexes (Abramson et al., 2024; Jumper et al., 2021). Boltz-2 further extended this framework, jointly predicting complex structure and binding affinity from sequence and ligand SMILES in a single inference pass (Passaro et al., 2025). Diffusion-based docking models such as DiffDock-L reframe pose prediction as generative sampling over translational, rotational, and torsional degrees of freedom, offering rapid conformational exploration without grid-based constraints (Corso et al., 2022). These tools have substantially accelerated structure-based drug discovery workflows. The AlphaFold-predicted p110α structure used here as the receptor input agrees closely with the experimental crystal structure (Figure 1.4), supporting its use as a modelling template.

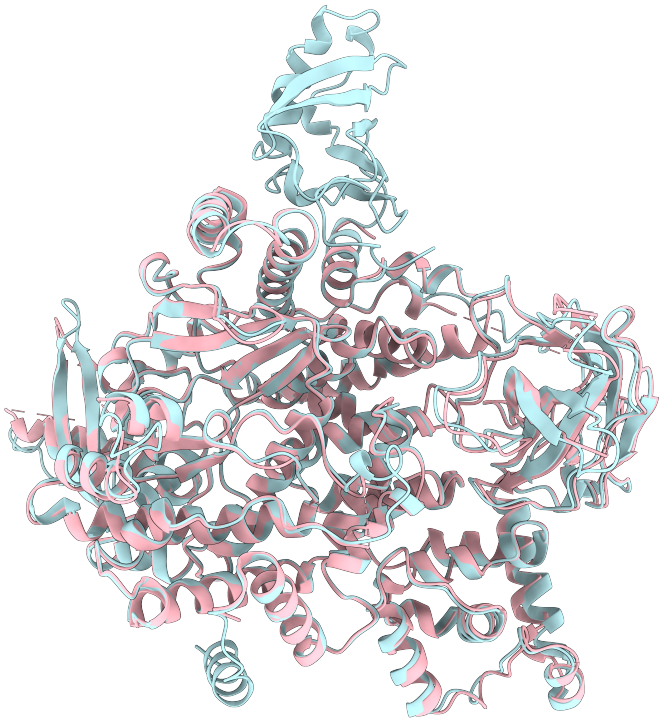


Figure 1.4. Comparative structural alignment of the AlphaFold WT prediction (blue, AF-P42336-F1) on the experimental p110α crystal structure (rose, PDB 8EXL). The AlphaFold model has a mean pLDDT of 92.38 % with over 82 % of residues in the highest-confidence band and a global Cα RMSD of 0.62 Å against the crystal structure.

A co-folding model trained predominantly on wild-type crystallographic data can place a ligand in the correct pocket of a mutant structure while assigning a confidence score that does not reflect whether the mutation actually changes binding affinity. Similarly, a rescoring model that achieves strong Spearman rank correlation across a curated affinity set may be recovering ligand-class information distinguishing high-affinity PI3Kα binders from unrelated kinase inhibitors rather than encoding the subtle, mutational variant-specific differences within the same chemical class.

This study evaluates tools across three methodological paradigms that differ in their input requirements and output modalities.

A) Co-folding models, represented here by Boltz-2 and Protenix, generate protein–ligand complex structures directly from sequence.
B) Structure-based docking and affinity models, represented by FlowDock, DiffDock-L, GNINA, AutoDock Vina, and TankBind, receive an explicit receptor structure alongside the ligand and sample binding conformations through learned or physics-guided search.
C) Affinity rescoring models, including CheapNet, GatorAffinity, DualBind, and GIGN, receive pre-generated protein-ligand complexes and return affinity estimates conditioned on the pose source. PLAPT is a sequence-only affinity model that takes protein sequence and ligand information without an explicit complex. Evaluating these paradigms across the 300-pair matrix therefore tests whether signal is carried by the ligand, by the generated pose, by the variant sequence, or by the scoring model itself.

1.4 Aims and Objectives

Current docking, co-folding, and rescoring tools do not capture inhibitor recognition equally well across all evaluation axes. Affinity scores are partly driven by the

potency of the ligand itself rather than by variant-specific binding, mutation-response signals appear only in certain tool and pose combinations, and pose accuracy varies independently of scoring performance.

This study pursued four objectives. First, a 300-pair cross docking benchmark matrix was constructed by crossing 15 p110α variants with 20 ligands spanning approved inhibitors, clinical candidates, allosteric compounds, isoform-selective agents, and inactive controls. Second, a curated affinity subset of 19 biochemical measurements drawn from primary literature and BindingDB (Gilson et al., 2016) was assembled to serve as ground truth for rank-ordering analysis. Third, all generated poses were assessed for structural quality against six crystallographic references using local root mean square deviation (LRMSD) (Su et al., 2019) and PoseBusters (Buttenschoen et al., 2023) physical-validity checks.

1.5 Scope and Limitations

The benchmark evaluated under a standardized protocol. The affinity ground truth is small (n = 19) because only direct biochemical measurements from recombinant PI3Kα assays were included, matched IC_{50} and K_i values for this exact variant panel are scarce in the literature, and cell-based or computationally derived values were excluded to avoid conflating biological endpoint with computational estimate. The specificity label set is broader (n = 202) but strict active/inactive classification is available only for the 19-row biochemical subset. Mutant structures were generated by PyMOL's mutagenesis module from the wild-type model rather than from experimentally determined coordinates, so backbone differences between variants are minimal.

Chapter 2: Literature Review

2.1 Free Energy Methods as the Accuracy Reference

Alchemical free energy perturbation (FEP) is the most physically rigorous computational approach to predicting how a protein mutation alters ligand binding affinity. By coupling molecular dynamics simulations to non-physical transformations FEP estimates the change in binding free energy ($\Delta\Delta G$) through a thermodynamic cycle. (Aldeghi et al., 2018) showed that consensus FEP approaches achieve RMSE values of approximately 1.01 kcal/mol on the PLATINUM database, a manually curated set of over 1,000 experimentally measured mutation effects on structurally defined protein–ligand complexes (Pires et al., 2015a). FEP's primary limitation is throughput: each mutation requires tens of nanoseconds of simulation, making systematic screening of large variant-by-ligand matrices impractical. Additionally, (Schindler et al., 2020) showed that starting pose quality has a larger effect on FEP accuracy than simulation length, motivating independent assessment of structural quality in any pipeline that uses faster pose generators as input to downstream scoring.

2.2 Docking, Co-folding, and Pose Prediction

The Comparative Assessment of Scoring Functions (CASF-2016) standardised the evaluation of docking methods across 285 protein–ligand complexes, reporting four performance axes: scoring power, ranking power, docking power, and screening power. The analysis found that docking power, reproducing the crystallographic pose was more reliably captured than affinity ranking across diverse targets (Su et al., 2019).

Generative AI docking and co-folding models have substantially changed the performance landscape. Co-folding models extend this further by jointly predict complex structure from protein sequence and ligand SMILES in a single inference pass, incorporating learned sequence-to-structure relationships that fixed-backbone docking cannot access. The PoseX cross docking benchmark (Jiang et al., 2026) evaluated these methods across 1,312 cross-docking entries spanning 109 protein targets and found that AI models consistently outperform physics-based methods on cross-docking pose reproduction, the scenario most relevant to multi-variant

benchmarking where an experimental receptor structure for every variant does not exist. PoseBusters (Buttenschoen et al., 2023) introduced a complementary validity test suite demonstrating that geometric accuracy alone is insufficient: a large fraction of AI-generated poses that pass RMSD thresholds still violate basic chemical and stereochemical constraints.

2.3 Structure-based Affinity Scoring and the Domain-Shift Problem

Deep learning scoring models trained on PDBbind commonly report Spearman correlations above 0.7 on the CASF-2016 core set, but performance degrades substantially when the target or chemical series falls outside the training distribution, a well-documented domain-shift problem. Models that achieve strong benchmark numbers through memorisation of PDBbind training chemistry cannot be expected to generalise to a target-specific, mutation evaluation. Classical scoring functions (empirical, force-field, and knowledge-based) trade physical grounding for speed and have been progressively displaced on PDBbind leaderboards by machine learning approaches from Random Forest to flow matching models. The benchmark gains, however, are partly artefactual: CASF-2016 test complexes share high sequence and scaffold similarity with PDBbind training entries, and performance drops substantially when evaluation is restricted to a previously unknown dataset and independent evaluations have found that affinity output of many models does not deliver the claimed precision in practice. (Wan et al., 2026) showed that confidence scores were compressed above 0.80 for all compounds tested across two large target libraries, providing effectively no discriminatory power between strong and weak binders, and that predicted affinities clustered within a narrow range regardless of true potency

2.4 Existing Benchmarks and the Gap This Study Addresses

The CASF-2016 and PoseX (Jiang et al., 2026) frameworks evaluate docking and scoring across diverse, unrelated targets but include no mutated proteins: each entry is a single protein sequence paired with a single ligand, also known as Self Docking. PLATINUM specifically addresses mutation effects on binding but aggregates mutations from hundreds of proteins and ligands, so it cannot probe how a given set of tools ranks binding across a defined clinical variant panel for a single target. No existing benchmark simultaneously evaluates pose quality, affinity ranking, selectivity

discrimination, and mutation-delta prediction for a single oncogenic kinase across a curated panel of approved and investigational inhibitors. Table 2.1 summarizes these differences.

Table 2.1 Comparison of established benchmark frameworks and this study. CASF-2016; PoseX; PLATINUM.

Feature	CASF-2016	PoseX	PLATINUM	This Study
N entries / pairs	285	1,312 (cross-dock)	>1,000 mutations	300 (15 × 20)
Single target	No	No	No	Yes (PI3Kα)
Variant / mutation panel	No	No	Multi-protein	15 clinical variants
Affinity ground truth	Yes	No	ΔΔG only	p-scale = pIC_{50} / pK_i (n = 19)
Pose correctness	Yes	Yes	No	Yes (6 crystal refs)
Specificity labels	No	No	No	Yes (n = 202)
Delta p-scale affinity (Δp)	No	No	Yes (cross-target)	Yes (9 matched pairs)
Allosteric ligands	No	No	No	Yes
Approved / clinical drugs	No	No	No	Yes

This study fills that gap by constructing a 300-pair matrix that fixes the target and varies both sequence (15 clinical variants of PI3Kα) and chemistry (20 ligands spanning orthosteric and allosteric binding modes). The design isolates variant-specific recognition from ligand-identity priors, a confound that multi-target benchmarks cannot control for and provides a structured test of whether current tools can distinguish mutation-sensitive from mutation-insensitive binding events in a clinically relevant setting.

3. Materials and Methods

3.1 Benchmark Design

The benchmark included 300 cross docking protein–ligand test cases, each corresponding to one p110α variant. The wild-type p110α sequence was obtained from UniProt entry P42336, and downloaded from AlphaFold Database (AF-P42336-F1-v6) and all mutant models were derived via PyMOL's mutagenesis module. The variant panel includes wild-type p110α, common oncogenic hotspot mutations, rare clinical variants, and ATP-pocket resistance-associated mutations. The ligand panel includes PI3Kα-selective orthosteric inhibitors, mutant-selective allosteric inhibitors, pan-PI3K inhibitors, isoform-selective PI3Kδ/γ controls, kinase decoys, and negative controls. Unlike classical scoring benchmarks such as CASF-16 which doesn't ranks scoring-functions from docking, or mutation-affinity database such as PLATINUM (Pires et al., 2015b) which aggregate mutation effects across many protein–ligand systems, this benchmark evaluates mutation-aware ligand recognition within a single clinically relevant oncogenic target. It is also complementary to recent benchmarks such as PoseX, which emphasizes self-docking and cross-docking pose accuracy, and DAVIS-Complete, which evaluates affinity prediction across modified kinase sequences (Jiang et al., 2026; Wu et al., 2025).

Table 3.1 Variant panel used in the p110α benchmark.

Variant	Domain	Category	Rationale
WT	—	Wild-type reference	Baseline receptor for all ligand and mutation comparisons.
H1047R	Kinase domain	Hotspot oncogenic mutation	Major PIK3CA hotspot (~20% of all PIK3CA-mutant tumors); important for mutant-selective inhibitor response and allosteric inhibitor evaluation (Menteş et al., 2022; Samuels et al., 2004).
E542K	Helical domain	Hotspot oncogenic mutation	Common helical-domain mutation, included to test whether models capture non-local mutational effects on inhibitor binding, particularly the reported reduction in inavolisib and serabelisib binding (Menteş et al., 2022; Samuels et al., 2004).
E545K	Helical domain	Hotspot oncogenic	Common hotspot close to E542K; included to test whether models distinguish related helical-

Variant	Domain	Category	Rationale
		mutation	domain variants (Menteş et al., 2022; Samuels et al., 2004).
H1047L	Kinase domain	Clinical hotspot variant	Alternative H1047 substitution (~4% of PIK3CA-mutant tumours); tests sensitivity to different amino-acid replacement at the same oncogenic position (Samuels et al., 2004).
M1043I	Kinase domain	Rare clinical variant	Located near the activation-loop region; with gain-of-function via activation-loop hinge alteration (Gymnopoulos et al., 2007).
G1049R	Kinase domain	Rare clinical variant	Introduces a charged residue into the catalytic loop; one of the strongest activating rare variants. Tests structural and electrostatic sensitivity. Sensitive to alpelisib, BEZ235, MK2206, trametinib; resistant to buparlisib (Dogruluk et al., 2015)
Q546K	Helical domain	Rare/helical clinical variant	Helical-domain mutant; disrupts inhibitory PIK3R1-nSH2 interaction. Included to expand beyond E542K/E545K hotspots (Gymnopoulos et al., 2007).
E726K	Kinase domain	Distal variant	Resistance to RLY-2608. Most frequent co-occurring "second hit" mutation in cis with E545K/E542K/H1047R (Varkaris, Fece de la Cruz, et al., 2024).
N345K	C2 domain	Rare clinical variant	Tests whether models respond to mutations outside the ATP pocket. C2 domain mutation that increases positive surface charge, enhancing membrane recruitment (Gymnopoulos et al., 2007).
C420R	C2 domain	Rare clinical variant	Disrupts disulfide bond, alters membrane interaction surface (Gymnopoulos et al., 2007).
R88Q	ABD	Rare clinical variant	Increases Akt/S6 phosphorylation despite being distal to catalytic domains, included to test distal mutation sensitivity (Oda et al., 2008).
W780R	Kinase / ATP-pocket region	Resistance-associated variant	Direct ATP-pocket steric perturbation. W780 is conserved across all 4 PI3K isoforms; W780R replaces bulky tryptophan with charged arginine, directly occluding orthosteric inhibitor binding. Causes universal resistance to alpelisib, inavolisib, pictilisib, and taselisib (Varkaris, Fece de la Cruz, et al., 2024).
Q859H	Kinase / ATP-pocket region	Resistance-associated variant	Unique to PI3Kα among isoforms. Q859 forms dual reciprocal H-bonds with the amide group in α-selective inhibitors (alpelisib, inavolisib). Q859H disrupts these H-bonds, selectively impairing α-selective but not pan-PI3K inhibitors. Tests whether models detect isoform-specific variants (Varkaris, Fece de la Cruz, et

Variant	Domain	Category	Rationale
			al., 2024).
Q859K	Kinase / ATP-pocket region	Resistance-associated variant	Same position, different positive substitution (lysine vs histidine). Disrupts the same H-bond network. Tests whether models distinguish distinct steric/geometric perturbations at the same pocket residue (Varkaris, Fece de la Cruz, et al., 2024). The full set of inhibitors evaluated in the benchmark, with their categories and target context, is listed in Table 3.2.

Table 3.2 Ligand panel used in the p110α benchmark.

Ligand	Category	Role	Rationale
Alpelisib	PI3Kα-selective orthosteric inhibitor	Positive control	FDA-approved PI3Kα inhibitor; equipotent against WT and mutant (H1047R, E545K, E542K); core reference for ATP-site recognition (Fritsch et al., 2014).
Inavolisib	PI3Kα-selective orthosteric inhibitor	Positive control	>300-fold selective over other class I PI3K isoforms; uniquely combines ATP-competitive inhibition with mutant-selective p110α degradation (Hanan et al., 2022).
Taselisib	PI3Kα-selective orthosteric inhibitor	Positive control	β-sparing PI3Kα/δ inhibitor, >30-fold over PI3Kβ; co-crystal (Ndubaku et al., 2013).
Serabelisib	PI3Kα-selective orthosteric inhibitor	Positive control	>100-fold selectivity over other class I PI3K isoforms and mTOR; clinical data in PIK3CA-mutated patients (Juric et al., 2017).
RLY-2608 / Zovegalisib	PI3Kα-selective allosteric inhibitor	Positive / allosteric test	First-in-class pan-mutant-selective allosteric PI3Kα inhibitor; 12× selective for H1047R over WT; binds cryptic pocket outside ATP site (Varkaris, Fece de la Cruz, et al., 2024).
STX-478	PI3Kα-selective allosteric inhibitor	Positive / allosteric test	Allosteric PI3Kα inhibitor, 14× selective for H1047R over WT (Buckbinder et al., 2023a).
Buparlisib	Pan-PI3K inhibitor	Positive PI3K binder	Inhibits all four class I PI3K isoforms; ≥50-fold selectivity over other protein

			kinases (Maira et al., 2012).
Copanlisib	Pan-PI3K inhibitor	Positive PI3K binder	Pan-class I PI3K inhibitor with sub-nanomolar IC_{50} against PI3Kα and PI3Kδ (N. Liu et al., 2013).
Pictilisib	Pan-PI3K inhibitor	Positive PI3K binder	Potent class I PI3K inhibitor (Folkes et al., 2008).
ZSTK474	Pan-PI3K inhibitor	Positive PI3K binder	ATP-competitive pan-class I PI3K inhibitor (Kong & Yamori, 2007).
Duvelisib	PI3Kδ/γ-selective inhibitor	Negative control, Isoform-selective	Potent oral dual inhibitor of PI3Kδ and PI3Kγ; tests discrimination between PI3Kα and related PI3K isoforms (Winkler et al., 2013).
Idelalisib	PI3Kδ-selective inhibitor	Negative control, Isoform-selective	40 to 300-fold selective for p110δ over other class I PI3K enzymes (α/β/γ); tests whether PI3K-family binding is conflated with PI3Kα specificity (Lannutti et al., 2011).
Parsaclisib	PI3Kδ-selective inhibitor	Negative control, Isoform-selective	Next-generation PI3Kδ inhibitor with ≥19,000-fold selectivity over other class I PI3K isoforms; adds stringent isoform-level challenge (Shin et al., 2020).
Imatinib	BCR-Abl kinase inhibitor	Kinase decoy	ATP-competitive BCR-Abl inhibitor (non-receptor tyrosine kinase); tests whether tools confuse protein-kinase binders with lipid-kinase PI3Kα (Druker et al., 2001).
Erlotinib	EGFR kinase inhibitor	Kinase decoy	Reversible ATP-competitive EGFR inhibitor (receptor tyrosine kinase); tests whether tools conflate upstream RTK pathway partners with direct PI3Kα binding (Shepherd et al., 2005).
Lapatinib	EGFR/HER2 kinase inhibitor	Kinase decoy	Dual EGFR/HER2 ATP-competitive inhibitor; large kinase inhibitor, chemically plausible but biologically incorrect for PI3Kα (Wood et al., 2004).
Olaparib	PARP inhibitor	Non-kinase decoy	PARP inhibitor (nuclear DNA repair enzyme); unrelated to PI3Kα strongest negative control for pathway specificity (Fong et al., 2009).
Palbociclib	CDK4/6 inhibitor	Kinase decoy	Selective CDK4/6 inhibitor (cell-cycle kinase downstream of PI3K/AKT); tests whether tools attribute binding based on pathway proximity (Fry et al., 2004).
Tamoxifen	Estrogen receptor modulator	Non-kinase decoy	Selective estrogen receptor modulator (nuclear hormone receptor); tests whether purely hydrophobic/hormonal compounds are mis-scored as PI3Kα binders (Osborne, 1998).
Vemurafenib	BRAF inhibitor	Kinase decoy	Selective ATP-competitive inhibitor of

			mutant BRAF (serine/threonine kinase in MAPK pathway); not related to PI3Kα (Joseph et al., 2010).

The ligand panel was designed to create a graded specificity challenge and it's not a simple binder-versus-decoy task. PI3Kα-selective and pan-PI3K inhibitors define the positive binding space, while isoform-selective PI3Kδ/γ inhibitors test whether models can distinguish PI3Kα recognition from related PI3K-family binding. Kinase inhibitors and unrelated drugs provide additional negative controls with different levels of chemical and biological difficulty. The inclusion of RLY-2608 and STX-478 adds a separate allosteric challenge, because these ligands require recognition of cryptic or mutation-dependent binding-site states rather than conventional ATP-pocket binding.

3.2 Chemical Diversity of the Ligand Library

To evaluate whether the ligand panel is structurally diverse enough for a specificity benchmark, pairwise chemical similarities were quantified using the Tanimoto coefficient. Each ligand was encoded as a Morgan2 fingerprint (ECFP4), a binary bit vector of length 2048 that records the presence or absence of circular substructural fragments within a radius of two bonds from each heavy atom (Rogers & Hahn, 2010).

For two ligands A and B, let:

a = the number of substructural features present in ligand A (bits set to 1 in its Morgan2 fingerprint)

b = the number of substructural features present in ligand B

c = the number of substructural features present in both A and B (bits set to 1 in both fingerprints)

The Tanimoto coefficient is then:

$$T(A,B)=\frac{c}{a+b-c}$$

The denominator represents the union of features present in either molecule, while the numerator counts their intersection. The coefficient ranges from 0, when no substructures are shared, to 1, when both molecules contain an identical set of substructural features. With 20 ligands, there are $\binom{20}{2}$=190 unique pairwise comparisons in the upper triangle of the 20 × 20 similarity matrix.

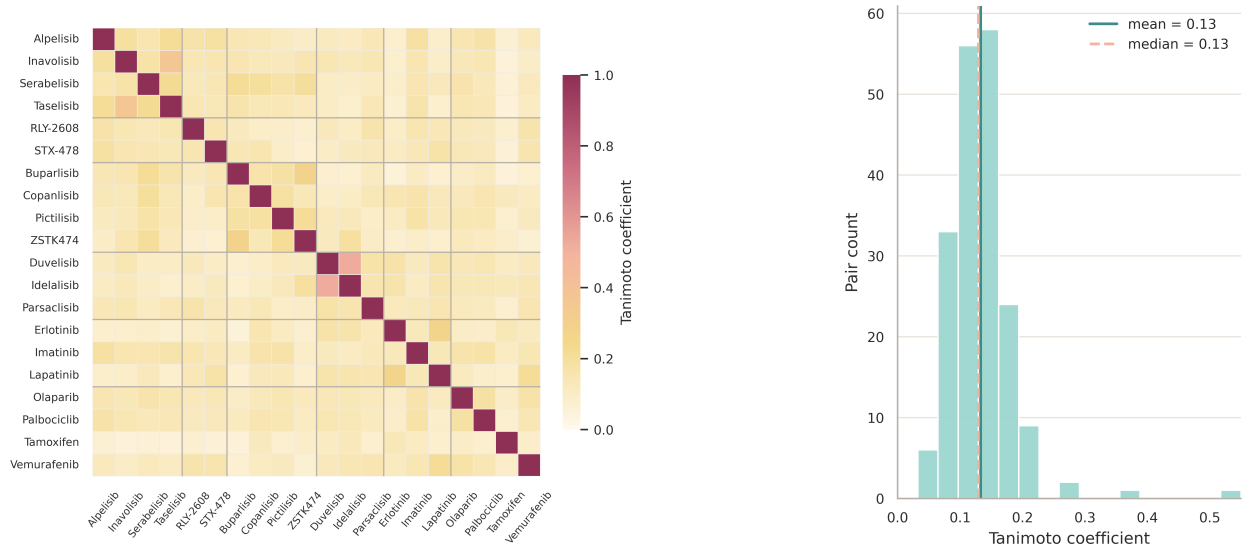


Figure 3.1 Pairwise Tanimoto similarity matrix and distribution of the ligand library

The ligand panel was first examined to determine whether the specificity task could be explained by gross chemical separation between positives and controls. Across the 20 ligands, 190 unique non-self pairwise Tanimoto comparisons were obtained. The overall distribution was low and narrow, with a mean coefficient of 0.133 and a Standard Deviation (SD) of 0.052 with a median of 0.129, consistent with the predominantly low off-diagonal similarities in the heatmap (Figure 3.1). The highest similarity was observed between Duvelisib and Idelalisib (Tanimoto = 0.526), and this was the only pair above 0.40. Comparisons between PI3Kα-positive ligands and negative-control ligands gave a mean similarity of 0.120, close to the full-library average. This indicates that the positive and negative groups are not separated simply by broad chemical dissimilarity, supporting the use of the panel for specificity analysis. In this context, successful separation by a tool is more likely to reflect its scoring behaviour across ligand-target contexts rather than a trivial distinction between chemically unrelated compound groups.

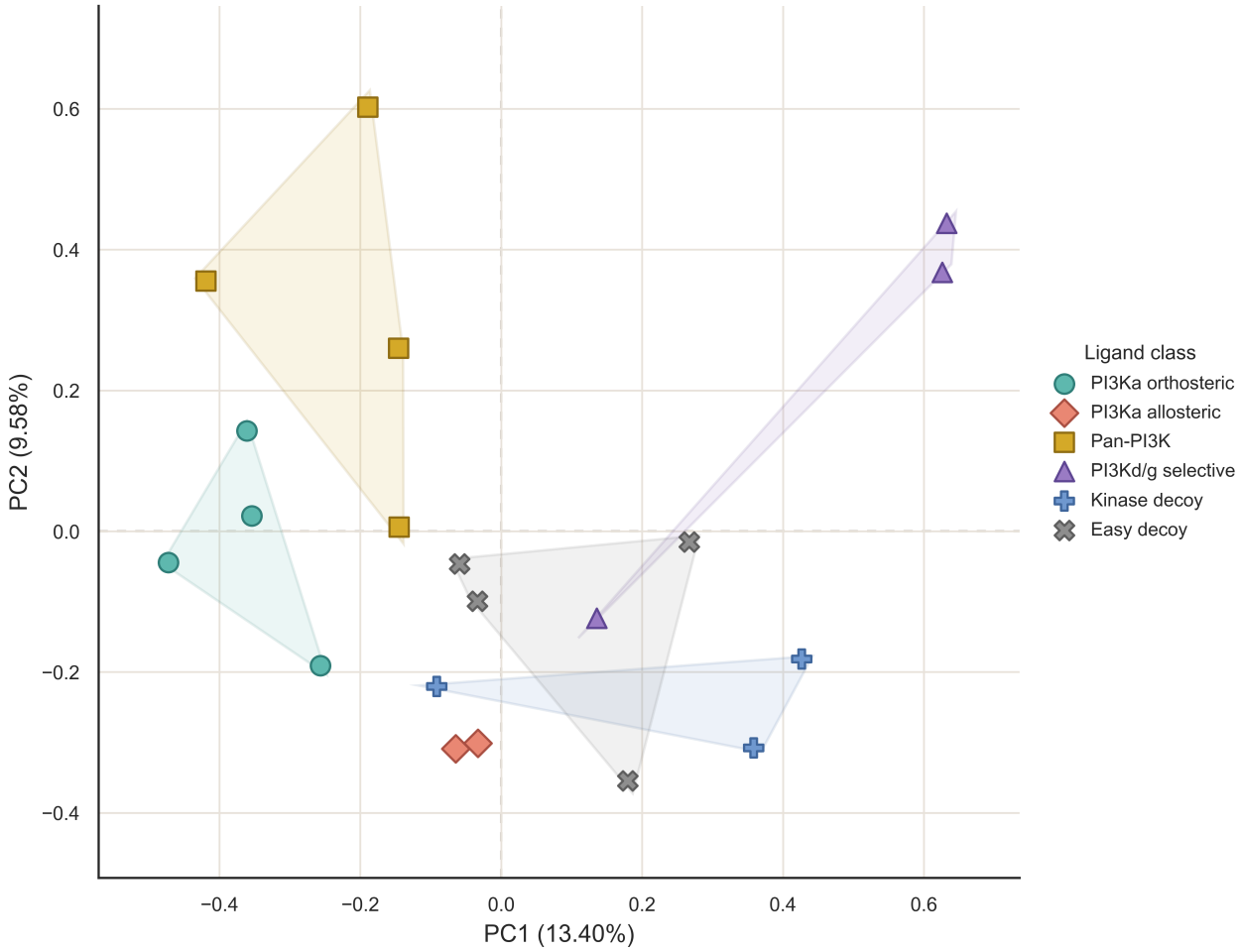


Figure 3.2 Principal Component Analysis (PCA) of ligands

Each ligand was represented by its Tanimoto similarity profile against all 20 ligands, and the first two principal components were plotted to show how the panel is distributed in chemical-similarity space (Figure 3.2). PC1 and PC2 explained 13.40% and 9.58% of the variance, respectively, indicating that the chemically, the ligands are pretty distinct.

3.3 Experimental Ground Truth

Experimental affinity evidence was curated separately from the full 300-combination score matrix. A combination was included in the quantitative affinity subset only when a direct biochemical IC_{50} or K_i value could be assigned to the ligand–variant pair.

Experimental reference data were assembled from three sources: BindingDB, manual extraction from literature for all 300 ligand-variant pairs.

Only pairs with a direct biochemical IC_{50} or K_i measurement assay were admitted as ground truth; cell-based functional readouts, and computational predictions were excluded.

IC_{50} values were converted to pIC_{50} and K_i values to pK_i to a uniform p-scale which is described in detail in section 3.4.4. The resulting curated biochemical subset is given in Table 3.3.

Table 3.3 Curated biochemical p-scale subset (n=19).

Ligand	Variant	Endpoint	Value (nM)	p-scale	Source
Alpelisib	E545K	IC_{50}	4.0	8.40	(Fritsch et al., 2014)
Alpelisib	H1047R	IC_{50}	4.8	8.32	(Fritsch et al., 2014)
Alpelisib	WT	IC_{50}	4.6	8.34	(Fritsch et al., 2014)
Buparlisib	WT	IC_{50}	52	7.28	(Maira et al., 2012)
Copanlisib	WT	IC_{50}	0.5	9.30	(N. Liu et al., 2013)
Inavolisib	E542K	K_i	0.016	10.80	(Hanan et al., 2022)
Inavolisib	E545K	K_i	0.014	10.85	(Hanan et al., 2022)
Inavolisib	H1047R	K_i	0.009	11.05	(Hanan et al., 2022)
Inavolisib	WT	K_i	0.038	10.42	(Hanan et al., 2022)
Pictilisib	WT	IC_{50}	3.0	8.52	(Folkes et al., 2008)
RLY-2608	H1047R	IC_{50}	4.0	8.40	(Varkaris, Fece de la Cruz, et al., 2024)
RLY-2608	WT	IC_{50}	48	7.32	(Varkaris, Fece de la Cruz, et al., 2024)
Serabelisib	WT	IC_{50}	15	7.82	(Jessen et al., 2011)
STX-478	E542K	IC_{50}	113	6.95	(Buckbinder et al., 2023a)
STX-478	E545K	IC_{50}	71	7.15	(Buckbinder et al., 2023a)
STX-478	H1047R	IC_{50}	9.4	8.03	(Buckbinder et al., 2023a)
STX-478	WT	IC_{50}	131	6.88	(Buckbinder et al., 2023a)
Taselisib	WT	K_i	0.29	9.54	(Ndubaku et al., 2013)
ZSTK474	WT	IC_{50}	16	7.80	(Kong & Yamori, 2007)

A second ground-truth view was constructed for mutation perturbation (Δp). For ligands with both a WT and mutant biochemical value measured on the same endpoint type, the experimental mutation delta was calculated as:

$$\Delta p = p(mutant) - p(WT)$$

A positive Δp indicates that the mutant binds the ligand more potently than WT.

The corresponding fold-change in raw potency was calculated from the absolute delta value:

$$\text{Fold-change} = 10^{|\Delta p|};\ |\Delta p| \geq 0.30,\ 10^{0.30} \approx 2$$

Thus, a Δp value of +1.0 represents a tenfold increase in potency relative to WT, while a Δp value of approximately ±0.30 corresponds to an approximately twofold potency shift. Deltas exceeding the two-fold threshold $|\Delta p| \geq 0.30$ were considered significant enough for tool evaluation, as smaller differences may fall within assay-level variability. Five of the nine calculated deltas met this criterion (Kramer et al., 2012).

Table 3.4 Experimental WT-mutant p-scale deltas (n=9).

Ligand	Variant	p(WT)	p(mut)	Δp	Fold-change	Considered
RLY-2608	H1047R	7.32	8.40	+1.08	12	Yes
STX-478	E542K	6.88	6.95	+0.06	1.2	No
STX-478	E545K	6.88	7.15	+0.27	1.8	No
STX-478	H1047R	6.88	8.03	+1.14	14	Yes
Alpelisib	E545K	8.34	8.40	+0.06	1.2	No
Alpelisib	H1047R	8.34	8.32	−0.02	1.0	No
Inavolisib	E542K	10.42	10.80	+0.38	2.4	Yes
Inavolisib	E545K	10.42	10.85	+0.43	2.7	Yes
Inavolisib	H1047R	10.42	11.05	+0.63	4.2	Yes

The nine paired mutation deltas cover three hotspot variants (E542K, E545K, H1047R); rare mutations, ATP-pocket resistance mutations, and distal mutations have no paired experimental reference and therefore excluded.

3.4 Computational Tool Families and Evaluation Settings

Co-folding models jointly predict protein–ligand complex structure and binding affinity from protein sequence and ligand SMILES, with no requirement for a pre-existing receptor structure. Structure-based docking and affinity models receive the AlphaFold-derived p110α receptor structure and a ligand SDF file as input, sampling binding poses and optionally predicting affinity scores. Affinity rescoring models operate on pre-generated three-dimensional protein–ligand complexes and produce binding affinity estimates; PLAPT is a sequence-only variant in this category that requires no three-dimensional pose. Boltz-2 and DiffDock-L poses were used as inputs to the three structure-dependent rescoring models, enabling analysis of whether affinity predictions remain stable across pose generators. Protenix was retained solely for structural quality control and was not included in the affinity or specificity scoring panel. Unless otherwise stated, all tools were run with default inference parameters as recommended in their original publications, and all computations were performed on NVIDIA A100 GPUs.

Ligand SMILES and 3D structures (SDF) were retrieved from PubChem and ChEMBL for all 20 compounds. The complete set of tools evaluated, with their families and run settings, is summarized in Table 3.5.

3.4.1 Biomolecular Co-folding Models

Co-folding models are relevant to this benchmark because they generate the receptor conformation and the ligand pose together from sequence and ligand chemistry, rather than docking into a fixed experimental structure. For a mutation-aware task this is attractive in principle: the oncogenic substitutions studied here, the helical-domain hotspots E542K and E545K and the kinase-domain hotspot H1047R, act by altering protein conformation and regulatory contacts, so a method that builds the pocket from the mutant sequence could in theory reflect those changes. The corresponding limitation is that a co-folding model can only reproduce conformations represented in its training data; a cryptic site that is rarely captured in public structures, such as the allosteric pocket engaged by mutant-selective inhibitors, may not be generated even when the mutant sequence is supplied.

Boltz-2

Boltz-2 (Passaro et al., 2025) is a biomolecular foundation model that jointly predicts protein–ligand complex structure and binding affinity within a single framework. Its diffusion module samples the three-dimensional complex from sequence, MSA, and ligand representations, while a coupled affinity head estimates a continuous IC_{50}-scale value from the final predicted complex state, making it the first deep learning model to approach the accuracy of physics-based free-energy perturbation methods on binding affinity benchmarks.

GitHub Repository**:** https://github.com/jwohlwend/boltz

Running Parameters**:** Default inference settings: recycling_steps = 3, sampling_steps = 200, diffusion_samples = 1, step_scale = 1.638, sampling_steps_affinity = 200, diffusion_samples_affinity = 5. Primary score: affinity_pred_value, reported on the $\log_{10}(IC_{50}\ \mu M)$ scale (lower raw values indicate stronger binding).

Runtime Environment: Run on NVIDIA A100 GPU.

Protenix

Protenix (Team et al., 2025) is an open-source all-atom AlphaFold3-style biomolecular structure predictor developed by ByteDance. It models proteins, ligands, nucleic acids, ions, and post-translational modifications within a single prediction framework, and supports high-throughput multi-sample inference with custom CUDA kernels for computational efficiency. In this benchmark, Protenix was used exclusively as a co-folding pose generator for structural quality control; its outputs were not included in the affinity or specificity scoring panel.

GitHub Repository**:** https://github.com/bytedance/protenix

Running Parameters: Default settings: N_cycle = 10, N_step = 200. Five samples were generated per pair; the highest-ranked sample by ranking_score was selected for structural analysis.

Runtime Environment: Run on NVIDIA A100 GPU.

3.4.2 Structure-Based Docking and Affinity Models

Structure-based docking methods place a ligand into a defined receptor pocket and score the resulting pose, and they are included here as the conventional reference against which the generative co-folding approach is compared. The mutant receptor inputs for this family were produced by in-silico residue substitution on the AlphaFold wild-type p110α backbone, so the docking search is performed against an essentially orthosteric, ATP-site geometry. This carries a structural limitation that must be stated at the outset: the cryptic allosteric pocket engaged by the mutant-selective inhibitors in the panel, RLY-2608 and STX-478, is not formed in the AlphaFold wild-type input, and a search box defined on that structure is centred on the orthosteric ATP site. A docking method can therefore only place a ligand where a pocket exists in the structure it is given; the allosteric site is physically absent from that input, so these methods are not expected to recover allosteric placement regardless of scoring quality.

FlowDock

FlowDock (Morehead & Cheng, 2025) formulates protein–ligand docking as geometric flow matching, learning a continuous vector field that transforms unbound apo structures into bound holo conformations for an arbitrary number of binding ligands. ESM2 protein language model embeddings (650M) condition the receptor representation alongside structural-template features, and both ranked binding poses and a continuous affinity estimate are returned per complex.

GitHub Repository**:** https://github.com/BioinfoMachineLearning/FlowDock

Running Parameters: Default settings: n_samples = 5, num_steps = 40, sampler = VDODE, sampler_eta = 1.0, use_template = true. Primary score: rank-1 affinity value.

Runtime Environment: Run on NVIDIA A100 40 GB GPU.

DiffDock-L

DiffDock-L (Corso et al., 2022) is a diffusion-based pose prediction model defined over the translational, rotational, and torsional degrees of freedom of the ligand. It is the large-scale generalised variant of DiffDock, trained with expanded data and

synthetic augmentation. A separate confidence model scores sampled poses; the confidence value is a unitless pose-quality estimate rather than a calibrated binding affinity.

GitHub Repository: https://github.com/gcorso/DiffDock

Running Parameters: Default settings: samples_per_complex = 40, inference_steps = 20, actual_steps = 19, no_final_step_noise = true. Primary score: rank-1 confidence value.

Runtime Environment: Run on NVIDIA A100 80 GB GPU.

GNINA

GNINA (McNutt et al., 2025) is an open-source molecular docking program that augments the AutoDock Vina search engine with convolutional neural network (CNN) scoring. Poses are sampled using Vina-style stochastic search and local optimisation, then rescored by a CNN that reads a three-dimensional voxel representation of the complex to produce a CNN affinity estimate and a pose-quality probability.

GitHub Repository: https://github.com/gnina/gnina

Running Parameters: Default settings: exhaustiveness = 8, num_modes = 9, CNN rescoring enabled, autobox_ligand set to the reference ligand SDF to define the search region. Primary score: cnn_affinity.

Runtime Environment: Run on NVIDIA A100 80 GB GPU.

AutoDock Vina

AutoDock Vina (Trott & Olson, 2010) is a widely used open-source molecular docking program that combines global stochastic search with local optimisation under an empirical scoring function. Receptor structures were prepared and converted to PDBQT format using Open Babel (v3.1.0); ligands were prepared using Meeko (v0.7.1). The docking search region was centred on the ATP-binding site as defined by superposition of the AlphaFold WT p110α model onto the alpelisib co-crystal structure (PDB 7PG6), yielding a grid centre of (−2.144, −8.122, 19.129) Å with a box size of 26×26×26 Å, sufficient to enclose the full orthosteric ATP pocket.

Software Version: AutoDock Vina 1.2.6; Meeko 0.7.1; Open Babel 3.1.0

GitHub Repository: https://github.com/ccsb-scripps/AutoDock-Vina

Running Parameters: center_x = −2.144, center_y = −8.122, center_z = 19.129 Å; size_x = size_y = size_z = 26 Å; exhaustiveness = 32; num_modes = 9. Primary score: top docking energy (kcal/mol; lower is stronger).

Runtime Environment: Run on CPU.

TankBind

TankBind (Lu et al., 2022) segments the receptor into candidate functional pockets identified by P2Rank and scores each pocket-ligand combination with a trigonometry-aware graph neural network that enforces geometric distance constraints. It also contains an affinity head trained on negative log-transformed PDBbind Kd/K_i/IC_{50} labels; in this rebuild, primary_score was therefore retained as p-scale rather than as a unitless complementarity value. Because the PDB-derived mutant inputs produce largely ligand-specific and sometimes variant-invariant scores, TankBind is interpreted as ligand-ranking evidence, not as reliable mutation-delta evidence.

GitHub Repository: https://github.com/luwei0917/TankBind

Running Parameters: Default settings. P2Rank (v2.3) was used for candidate pocket detection. Primary score: maximum pocket affinity prediction (score_pocket2; higher is stronger).

Runtime Environment: Run on NVIDIA A100 GPU.

3.4.3 Affinity Rescoring Models

Affinity rescoring models estimate binding affinity from pre-generated protein–ligand pose complexes from Boltz-2 and DiffDockL. PLAPT is a sequence-only affinity model in this category; because it requires no three-dimensional complex, it received the same inputs as the co-folding models (sequence and SMILES) but is categorised here as a structure-free affinity baseline. These models are included because a rescorer can sharpen the affinity estimate for a given complex, but it does not generate geometry of its own: its sensitivity to an oncogenic mutation is therefore bounded by the pose it is handed, and it inherits whatever the upstream generator captured or

missed, including the allosteric pocket that is absent from the AlphaFold wild-type input. A rescoring model can thus register a mutation-driven change in the orthosteric site where the supplied pose is well formed, but cannot recover a binding mode, such as the allosteric site engaged by RLY-2608 and STX-478, that the underlying pose does not contain.

CheapNet

CheapNet (Lim et al., 2024) applies differentiable pooling to compress atom-level protein–ligand complex graphs into hierarchical cluster representations, then uses bidirectional cross-attention between ligand and protein clusters to predict binding affinity. Published at ICLR 2025, it achieves state-of-the-art affinity prediction performance while maintaining low computational overhead relative to full pairwise attention models.

GitHub Repository: https://github.com/hyukjunlim/CheapNet

Running Parameters: Default settings. Three independent checkpoints were evaluated; the mean prediction (pred_mean) was used as the primary score. Applied independently to Boltz-2 and DiffDock-L pose sources.

Runtime Environment: Run on NVIDIA A100 40 GB GPU.

GatorAffinity

GatorAffinity (Wei et al., 2025) is a geometric deep learning affinity model initialised from ATOMICA weights pretrained on approximately two million biomolecular interaction complexes, then fine-tuned on experimental PDBbind-style structures augmented with over one million synthetic IC_{50}-labelled complexes from the SAIR database. It consumes three-dimensional pocket–ligand structures and produces a pK-scale affinity estimate.

GitHub Repository: https://github.com/AIDD-LiLab/GatorAffinity

Running Parameters: Default inference settings. Pocket PDB and ligand files were derived from Boltz-2 and DiffDock-L pose outputs respectively. Primary score: predicted pK value.

Runtime Environment: Run on NVIDIA A100 GPU.

DualBind

DualBind (M. Liu & Paliwal, 2024) is an SE(3)-invariant structure-based affinity model developed by NVIDIA that combines supervised mean-squared error training with an unsupervised denoising score matching (DSM) objective on perturbed ligand coordinates. The dual-loss design trains an energy-like scoring surface while anchoring predictions to labelled affinity data, so the local output requires an explicit sign-convention audit before it is interpreted as p-scale affinity.

GitHub Repository: https://github.com/NVIDIA-Digital-Bio/dualbind

Running Parameters: Default inference settings. Applied independently to Boltz-2 and DiffDock-L pose sources. Primary score: Raw predictions (internal −pK convention)

Runtime Environment: Run on NVIDIA A100 GPU.

Protein–Ligand Affinity Prediction using Pretrained Transformers (PLAPT)

PLAPT (Rose et al., 2024) encodes the protein sequence with ProtBERT and the canonical ligand SMILES with ChemBERTa, concatenates the resulting embeddings, and passes the joint representation through a branching neural network to predict binding affinity on the pIC_{50} scale. Because it requires no three-dimensional complex, any recovered signal reflects protein sequence context and ligand molecular identity rather than explicit binding geometry.

GitHub Repository: https://github.com/trrt-good/PLAPT

Running Parameters: Default settings: ProtBERT and ChemBERTa embeddings, ONNX inference, CUDA-accelerated where available. Primary score: predicted neg-log affinity mapped to p-scale (higher is stronger).

Runtime Environment: Run on NVIDIA A100 GPU.

Graph Interaction-based Graph Network (GIGN)

GIGN (Yang et al., 2023) is a structure-based binding affinity prediction model that constructs interaction graphs from protein–ligand complexes, representing atoms as nodes and encoding covalent, non-covalent, and inter-molecular interactions as edge features. A hierarchical message-passing scheme propagates information across intra-

protein, intra-ligand, and protein–ligand edges, and a graph-level readout produces the final affinity prediction. GIGN was applied to Boltz-2 pose and DiffDockL inputs.

GitHub Repository: https://github.com/guaguabujianle/GIGN

Running Parameters: Default inference settings from the published model checkpoint. Applied to Boltz-2 pose source. Primary score: predicted pIC_{50} value.

Runtime Environment: Run on NVIDIA A100 GPU.

Table 3.5 Summarizing all the tools used

Tool	Input	Type	Output and Scores	Interpretation
Boltz-2	sequence + SMILES	Co-folding with affinity head	Complex structure and $\log_{10}$ IC_{50} (μM), converted to p-scale	Higher is stronger.
Protenix	sequence + SMILES	Co-folding	Complex structure and pose only	No affinity score
FlowDock	structure + ligand	Geometric flow-matching docking with affinity head	Ranked complex structure and affinity (p-scale)	Higher is stronger.
DiffDock-L	structure + ligand	Diffusion docking model	Docked pose only	No affinity score.
GNINA 1.3	structure + ligand	Vina docking with CNN re-scoring	Docked pose and affinity (p-scale)	Higher is stronger.
AutoDock Vina	structure + ligand	Empirical docking and scoring	Docked pose and binding energy (kcal/mol) converted to p-scale	Higher is stronger.
TankBind	structure + ligand	Trigonometry-aware network predicts binding pose and affinity-like pocket score |	Docked pose and Affinity Score (p-scale)	Higher is stronger.
CheapNet	Boltz-2 / DiffDock-L pose	Hierarchical cross-attention rescoring	Affinity value on docked (p-scale)	Higher is stronger.
Gator Affinity	Boltz-2 / DiffDock-L pose	Geometric affinity regressor trained on synthetic data from Boltz-1x	Affinity value on docked structure (p-scale)	Higher is stronger.
DualBind	Boltz-2 / DiffDock-L pose	Supervised MSE + denoising score matching	ΔG affinity score on docked structures (p-scale)	Higher is stronger.
GIGN	Boltz-2 / DiffDock-L pose	Geometric interaction graph neural network	p-scale score on docked structure	Higher is stronger.

PLAPT	sequence + SMILES	BERT affinity model	p-scale score, negative-log affinity prediction	Higher is stronger.

3.4.4 Conversion to p-scale

All activity measurements (IC_{50}, K_i, Kd) were converted to p-scale using:

$$p = -\log_{10}\left(\frac{c}{\text{mol L}^{-1}}\right)$$

where c is the measured value in molar units and the division strips the unit to leave a dimensionless unit. For example, an IC_{50} of 10 nM gives $p = -\log_{10}(10^{-8}) = 8.0$. This convention follows ChEMBL (Bento et al., 2014) and is identical to the pK convention used in PDBbind (Wang et al., 2004) the primary training corpus for most tools evaluated here which output pK directly and therefore no need of conversion.

Boltz-2 reports $\log_{10}(IC_{50} / \mu M)$, giving $p = 6.0 - \log_{10}(IC_{50} / \mu M)$; AutoDock Vina reports ΔG (kcal mol^{-1}), giving $p = -\Delta G / 1.3633$ via $\Delta G = RT \ln Kd$ at 298 K ($RT \ln 10 = 1.3633$ kcal mol^{-1}); and DualBind stores predictions in the −pK convention, giving p = −(raw output).

Where only IC_{50} was available but K_i was required, the Cheng–Prusoff equation

$$K_i = \frac{IC_{50}}{1 + \frac{[S]}{K_m}}$$

was applied using the reported substrate concentration [S] and Michaelis constant K_m. The per-tool conversion formulae are summarized in Table 3.6.

Table 3.6 Conversion of raw tool outputs and experimental measurements to the common p-scale (negative base-10 logarithm in molar units)

Tool	Native output	Raw example	Conversion formula	p-scale value
Experimental IC_{50}	IC_{50} (nM)	4.3 nM	$-\log_{10}(4.3 \times 10^{-9})$	8.37
Experimental K_i	K_i (nM)	8.0 nM	$-\log_{10}(8.0 \times 10^{-9})$	8.10
Boltz-2	$\log_{10}(IC_{50}/\mu M)$	0.62	6.0 − 0.62	5.38

AutoDock Vina	ΔG (kcal mol^{-1})	−9.52	−(−9.52) / 1.3633	6.98
DualBind	−pK (raw output)	−7.32	−(−7.32)	7.32
TankBind	pK ($-\log_{10}$ M)	7.89	Conversion not needed	7.89
GatorAffinity	pK	8.21	Conversion not needed	8.21
CheapNet	pK (pred_mean)	7.45	Conversion not needed	7.45
GIGN	pIC_{50}	6.83	Conversion not needed	6.83
GNINA	pK (cnn_affinity)	7.12	Conversion not needed	7.12
PLAPT	pK (neg-log)	7.34	Conversion not needed	7.34

3.5 Evaluation axis

P-scale affinity, cross-docked specificity, delta p-scale affinity, and pose correctness each required a distinct evaluation strategy.

3.5.1 P-scale affinity and delta p-scale affinity

Affinity was ranked using Spearman's rank correlation coefficient. Spearman's ρ measures whether a tool preserves the experimental ordering of ligand–variant entries, rather than whether the predicted values match the experimental values exactly. This was appropriate because different tools may produce shifted or compressed score ranges while still correctly ranking stronger and weaker binders.

For each tool, predicted and experimental p-scale values were independently ranked. For the *i*-th entry, the rank difference was calculated as:

$$d_i = rank(predicted_i) - rank(experimental_i)$$

Spearman's coefficient was then calculated as:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

A value of $\rho = +1$ indicates that the tool ranks all entries in exactly the same order as the experimental data. A value of $\rho = -1$ indicates a perfectly reversed ranking, while a

value close to 0 indicates no monotonic relationship between predicted and experimental affinity order.

Because the curated p-scale subset is small (n = 19), a permutation test was used to assess whether the observed rank correlation was stronger than expected by chance. In each of 2,000 iterations, the experimental p-scale values were randomly shuffled, breaking any real association between prediction and experiment. Spearman's ρ was then recomputed for each shuffled dataset to generate a null distribution of correlations expected under random ranking

The permutation p-value was calculated as:

$$p=\frac{n_{extreme}+1}{2001}$$

Where $n_{extreme}$ is the number of permutations in which:

$|\rho_{permuted}|>=|\rho_{observed}|$

The added pseudo-count prevents zero p-values. A permutation p-value below 0.05 was interpreted as evidence that the observed ranking was unlikely to arise by chance.

3.5.2 Cross-docked specificity

Cross-docked specificity asked whether each tool assigned stronger oriented scores to positive controls than to negative controls. For each tool, raw outputs were converted to oriented scores before calculating discrimination metrics, ensuring that larger values consistently represented stronger predicted binding, or confidence, depending on the score type.

- AUROC was computed independently for each tool; scores from different tools were never placed in the same ranked list.
- AUROC is rank-based and does not require a common scale across tools, it measures only whether each tool ranks positive above negative controls.

The broad specificity task contained 202 ligand-variant rows: 52 broad positives and 150 negative controls, described in Table 3.7.

Table 3.7 Specificity-label used for AUROC/AP analysis (n=202).

Role	Count	Explanation
Negative control	150	Decoy ligands.
Positive control	19	IC_{50} or K_i data present.
Positive control	33	Weaker evidence like qualitative activity reports, active compounds without exact IC_{50} values. Ligand-variant pairs without a usable specificity label were excluded from the AUROC/AP analysis; these are listed in Table 3.8.

Table 3.8 Specificity-excluded ligands from AUROC/AP (n=98).

Ligand	Count	Variants excluded
RLY-2608	7	H1047L, C420R, N345K, Q546K, R88Q, G1049R, M1043I
STX-478	7	E726K, C420R, N345K, Q546K, R88Q, G1049R, M1043I
ZSTK474	14	E542K, E545K, H1047L, H1047R, E726K, Q859H, Q859K, W780R, C420R, N345K, Q546K, R88Q, G1049R, M1043I
Alpelisib	8	E542K, C420R, G1049R, H1047L, M1043I, N345K, Q546K, R88Q
Buparlisib	10	H1047L, E726K, Q859H, Q859K, C420R, N345K, Q546K, R88Q, G1049R, M1043I
Copanlisib	14	E542K, E545K, H1047L, H1047R, E726K, Q859H, Q859K, W780R, C420R, N345K, Q546K, R88Q, G1049R, M1043I
Inavolisib	7	H1047L, C420R, N345K, Q546K, R88Q, G1049R, M1043I
Pictilisib	10	E542K, E545K, H1047L, H1047R, C420R, N345K, Q546K, R88Q, G1049R, M1043I
Serabelisib	11	H1047L, E726K, Q859H, Q859K, W780R, C420R, N345K, Q546K, R88Q, G1049R, M1043I
Taselisib	10	E542K, E545K, H1047L, H1047R, C420R, N345K, Q546K, R88Q, G1049R, M1043I

The area under the receiver operating characteristic curve (AUROC) measures the probability that a randomly chosen positive receives a higher score than a randomly chosen negative. It is computed by sweeping through all possible classification thresholds and plotting the true positive rate against the false positive rate:

$$\text{AUROC} = \int_0^1 \text{TPR}\left(\text{FPR}^{-1}(x)\right) dx$$

AUROC ranges from 0.5, meaning random guessing, to 1.0, meaning perfect separation. It summarises rank separation across all thresholds but can be misleading when the positive class is small; a large negative background can produce a high AUROC even when precision at high recall is poor.

Average precision (AP), the area under the precision-recall curve, addresses this limitation. Precision is the fraction of positives among the top (k) scored items, while recall is the fraction of all positives captured at that cutoff. AP is calculated as:

$$\text{AP} = \sum_{k=1}^{n} P(k) \cdot 1\{\text{rank } k \text{ is positive}\}$$

where (P(k)) is precision at the cutoff that includes the top (k) predictions. AP ranges from 0 to 1 and rewards models that rank positives early. In both specificity tasks the positive class was smaller than the negative class, making AP the more diagnostically useful metric. A high AP means the tool consistently places known positive controls ahead of negative controls.

3.5.3 Pose correctness

Because several downstream models operate on predicted protein-ligand geometries, pose correctness was evaluated as an independent benchmark axis rather than inferred from affinity or specificity performance. Strict structural assessment was limited to ligand-variant cases with matched crystallographic protein-ligand complexes, allowing generated poses to be compared against experimental reference ligand positions.

The strict reference set contained six direct PIK3CA complex structures (Table 3.9). Orthosteric references were used to test conventional ATP-site placement, while allosteric references were used to test whether generated poses localized ligands to the cryptic allosteric site rather than defaulting to the orthosteric pocket. Only the selected ligand component was used as the strict reference ligand; additional ligands in the same structure were removed.

Table 3.9 Crystal structures used for strict structural pose assessment

PDB ID	Reference ligand	Variant	Binding site
7PG6	Alpelisib (1LT)	WT/reference	Orthosteric
8AM0	Inavolisib (MWF)	WT/reference	Orthosteric
8EXL	Taselisib (799)	WT/reference	Orthosteric
8TSD	RLY-2608 (XUZ)	WT/reference	Allosteric
8TDU	STX-478 (ZWE)	WT/reference	Allosteric
8TGD	STX-478 (ZWE)	H1047R	Allosteric

Generated poses from Boltz-2, Protenix, and DiffDock-L were evaluated against these references. Boltz-2 and Protenix were treated as co-folding or complex-generation pose sources, whereas DiffDock-L was treated as a docking pose generator using prepared protein structures. Pose-dependent rescoring tools were not treated as independent pose generators; their structural interpretation followed the Boltz-2 or DiffDock-L pose source used as input.

LRMSD was used as the primary geometric metric, following the local-pocket alignment protocol established by the CASF-2016 benchmark (Su et al., 2019). The predicted protein structure was aligned to the crystallographic reference using Cα atoms within the binding-site environment; the resulting transformation was then applied to the ligand, and heavy-atom displacement was computed via RDKit maximum-common-substructure mapping. This approach measures ligand placement relative to the pocket rather than global protein-frame differences. Values ≤ 2 Å were treated as successful pose recovery, consistent with the CASF-2016 threshold.

PoseBusters (Buttenschoen et al., 2023) is a comprehensive validation suite that evaluates chemical and physical plausibility of predicted ligand poses. It applies a series of modular checks covering ligand geometry (bond lengths, angles, ring puckers), protein-ligand steric clashes, internal steric clashes, intramolecular hydrogen bonding, stereochemistry, and RMSD-based pose agreement. The redocking preset was used, which applies 21 module functions comprising 46 boolean checks. The Python API was called as:

PoseBusters(config="redock", max_workers=0)x

with the generated ligand SDF as *mol_pred*, the matched reference ligand SDF as *mol_true*, and the extracted reference protein as *mol_cond*. Full per-check boolean reports were retained for each pose, allowing individual failure types, such as steric clash, RMSD threshold, or stereochemistry violations. Three checks related to crystallographic water or cofactor distances are not applicable to the complexes evaluated here and were excluded from interpretation.

4. Results and Discussion

4.1 Scope of the analysis

This chapter is organized around four analysis questions rather than a single overall ranking of tools. The same numerical output can look strong for different reasons. A tool may rank experimental affinities well, separate target-relevant cases from negative controls, respond correctly to mutation, or simply inherit a plausible pose from the structure generator that feeds it. Each of these behaviours is evaluated separately so that confidence, affinity, and pose correctness are not collapsed into a single claim.

Four distinct axes structure the analysis.

1) P-scale affinity (n = 19), which asks whether tool scores preserve the experimental p-scale ordering of ligand-variant pairs; performance is measured by Spearman rank correlation and meaningful-gap pairwise concordance.
2) Cross-docked specificity (n = 202), which asks whether tool scores separate positive controls from negative controls across the full labelled pair set; performance is measured by AUROC and average precision (AP).
3) Delta p-scale affinity (n = 9 paired WT–mutant comparisons), which asks whether tool scores move in the experimentally observed direction when only the variant is changed, with sign accuracy on the five significant Δp values ($|\Delta p| \geq 0.30$) as the primary indicator.
4) Pose correctness (n = 6 crystallographic references), which asks whether the generated ligand pose sits in the reference binding pocket; LRMSD, PoseBusters chemical validity checks

4.2 Results

4.2.1 P-scale affinity

This analysis tested whether tool scores preserved the experimental ordering of ligand-variant pairs on the curated biochemical subset (n = 19; 14 IC□ □-derived and 5

K_i-derived values). The endpoint mix and small size mean the subset is treated as a rank-ordering test rather than as a calibrated affinity benchmark.

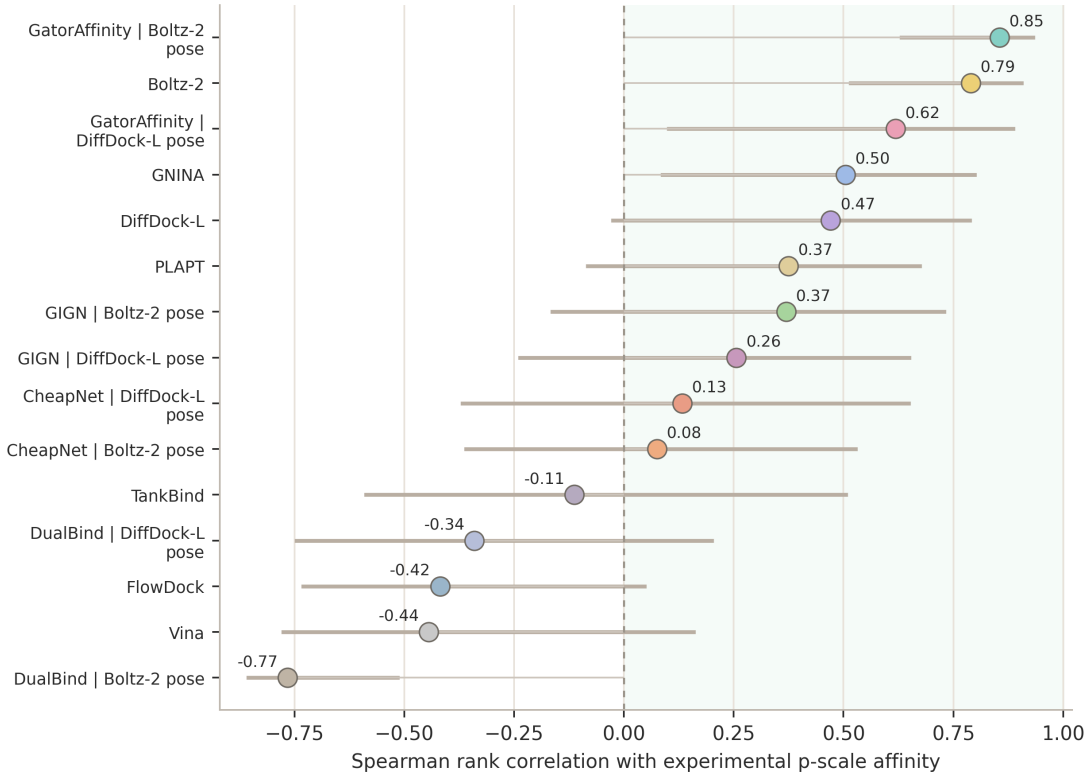


Figure 4.1. Spearman rank correlation (ρ) between tool scores and experimental p-scale affinity across the curated biochemical subset (n = 19 pairs). Each point is the median ρ; horizontal bars are bootstrap 95% confidence intervals (2,000 resamples of n = 19 with replacement, 2.5th– 97.5th percentiles). The shaded green region marks positive correlation (ρ > 0), indicating tools that recover the correct experimental rank ordering. Tools left of the dashed zero-line anti-correlate with measured affinity.

GatorAffinity rescoring on Boltz-2 poses achieved the highest rank recovery (ρ = 0.85, Figure 4.1), followed by Boltz-2 scoring its own co-folded complex (ρ = 0.80). Applying GatorAffinity to DiffDock-L poses yielded a weaker correlation (ρ = 0.62), showing that ligand pose affects rescoring performance. GNINA and DiffDock-L produced modest positive correlations (ρ = 0.50 and 0.47, respectively), although DiffDock-L remains a confidence output rather than an affinity predictor. Placing all tools on the p-scale did not guarantee useful rank recovery: TankBind remained near zero (ρ = −0.11), Vina anti-correlated (ρ = −0.44), and DualBind anti-correlated on both pose sources, most strongly on Boltz-2 poses (ρ = −0.77). Rank recovery is a weaker test than absolute potency prediction.

The p-scale analysis shows that p-scale output is not itself evidence of reliable affinity capture. GatorAffinity on Boltz-2 poses and Boltz-2 carried the clearest ligand-rank signal, whereas TankBind, Vina, and DualBind produced weak or inverted correlations on this target-specific panel.

4.2.2 Cross-docked specificity

This analysis asked whether each tool assigned stronger oriented scores to positive controls than to negative controls across the labelled cross-docked pair set (n = 202; 52 broad positives and 150 negative controls). The broad positive class spans all four evaluation axes, p-scale affinity, cross-docked specificity, delta p-scale affinity, and pose correctness, so the result is interpreted as benchmark-label separation rather than a single clean experimental task.

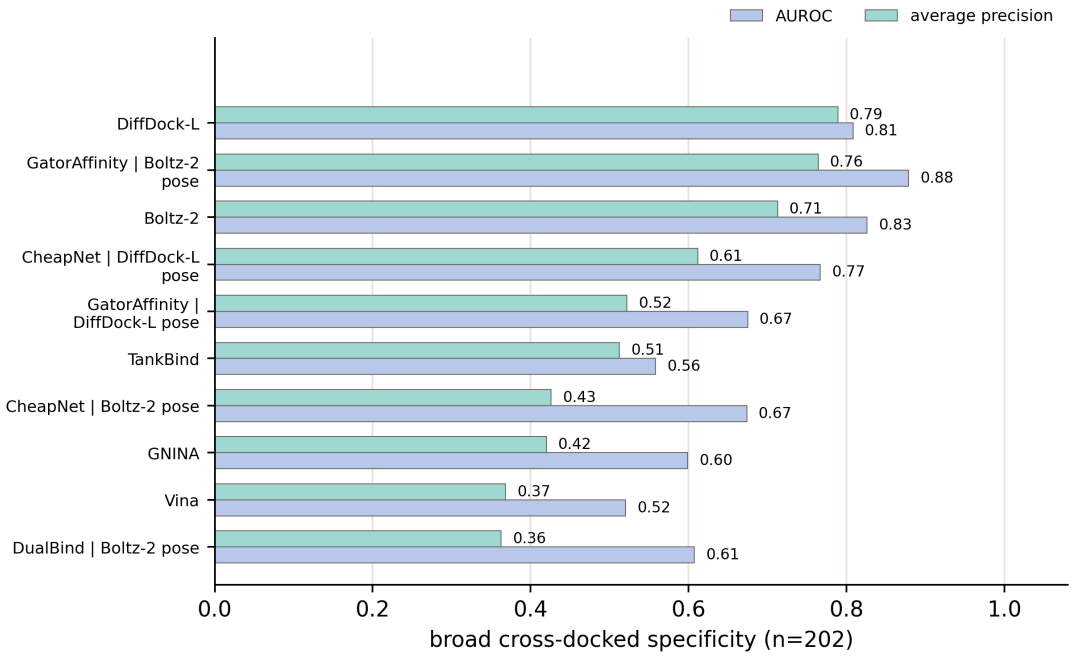


Figure 4.2. AUROC and average precision (AP) for the broad cross-docked specificity task (52 positives, 150 negative controls). Both metrics measure how well each tool separates known PIK3CA binders from non-binders. The dashed line marks random performance.

GatorAffinity on Boltz-2 poses achieved the highest AUROC (0.88, Figure 4.2), indicating the strongest overall separation between active and inactive compounds. DiffDock-L led on AP (0.79), meaning it pushed the most positives to the top of its ranked list. However, DiffDock-L outputs only a geometric docking confidence score rather than an affinity estimate, its AP advantage came from that score happening to assign low values to several negative controls, not from any calibrated potency

prediction. Boltz-2 alone was competitive on both metrics (AUROC = 0.83, AP = 0.71) without requiring a separate rescoring tool. Because the positive class is smaller than the negative background, AP is the more diagnostically useful of the two metrics here: low positive prevalence depresses precision at high recall, so strong AP values indicate that a tool consistently places the comparatively scarce positive controls ahead of the negative background rather than exploiting a large, easily separated negative pool.

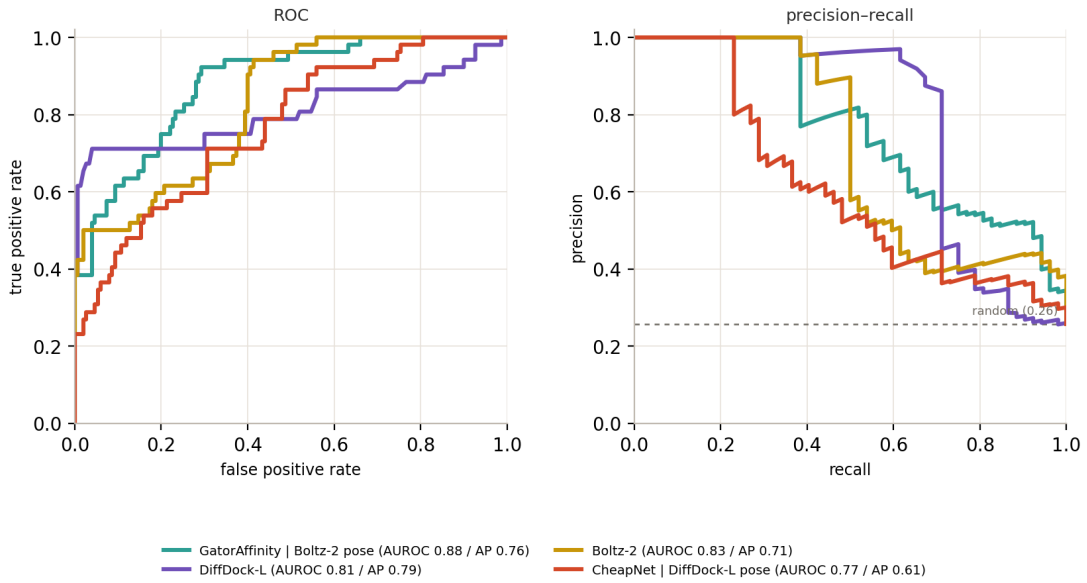


Figure 4.3. ROC and precision–recall curves for the four top tools on the broad cross-docked specificity task (52 positives, 150 negatives) for each tool. The dashed line in the PR panel marks the random performance, a tool with no signal would achieve precision equal to the positive-class prevalence (0.26) at every recall level. All four tools sit well above it, confirming meaningful discrimination. The PR curves appear weaker than the ROC curves because precision is sensitive to the 3:1 class imbalance (150 negatives to 52 positives): false positives accumulate as recall increases, pulling precision toward the prevalence baseline even when ranking is sound, whereas ROC normalizes within each class and is unaffected. The stepwise (staircase) shape of both curves reflects the modest number of positives, where each positive shift recalls by roughly 2%.

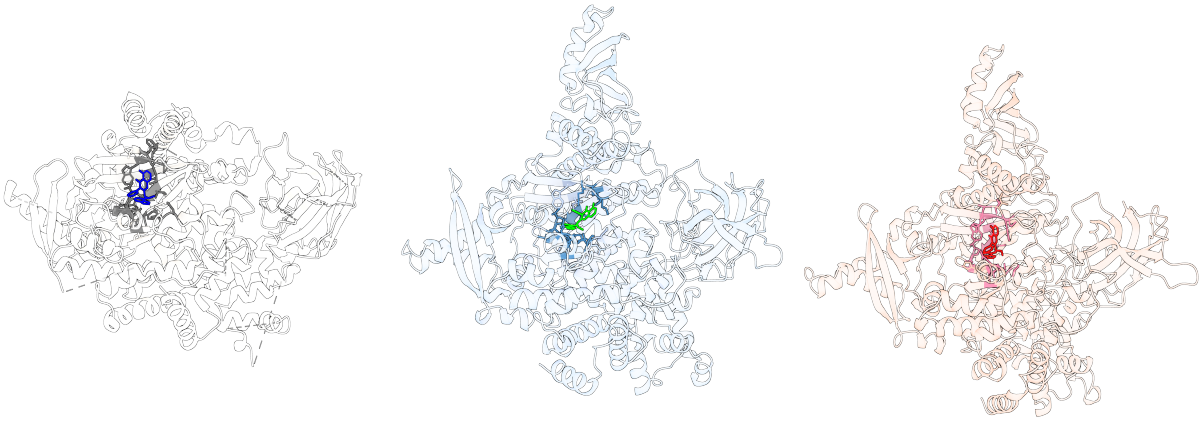


Figure 4.4. Idelalisib (PI3Kδ-selective, weakly active on PIK3CA) docked onto p110α. Left: PI3Kδ crystal pose (PDB 4XE0). Centre: DiffDock-L prediction (confidence −1.43, correctly low). Right: Boltz-2 prediction (confidence 0.92, predicted p-scale 6.59; inconsistent with its weak PI3Kα activity).

Idelalisib is weakly active on PIK3CA and serves as a negative control here. Both tools place it in the orthosteric pocket of p110α, but they diverge in what they report: DiffDock-L assigns a low docking confidence (−1.43), correctly flagging it as a poor binder, while Boltz-2 assigns high confidence (0.92) and a strong predicted p-scale of 6.59. This is why DiffDock-L leads on AP while Boltz-2 leads on p-scale ranking.

4.2.3 Delta p-scale affinity (Δp)

This analysis asked whether the tools captured the direction and approximate size of mutation-induced affinity changes when the ligand and assay endpoint were held constant. The test used the nine paired WT–mutant entries defined in Section 3.3 (Table 3.4), covering inavolisib, alpelisib, STX-478, and RLY-2608 against E542K, E545K, and H1047R. Five of the nine experimental deltas met the $|\Delta p| \geq 0.30$ threshold for a two-fold or greater shift. Performance was assessed by Spearman correlation between predicted and experimental deltas, directional sign accuracy on all nine pairs, and sign accuracy restricted to the five significant Δp values.

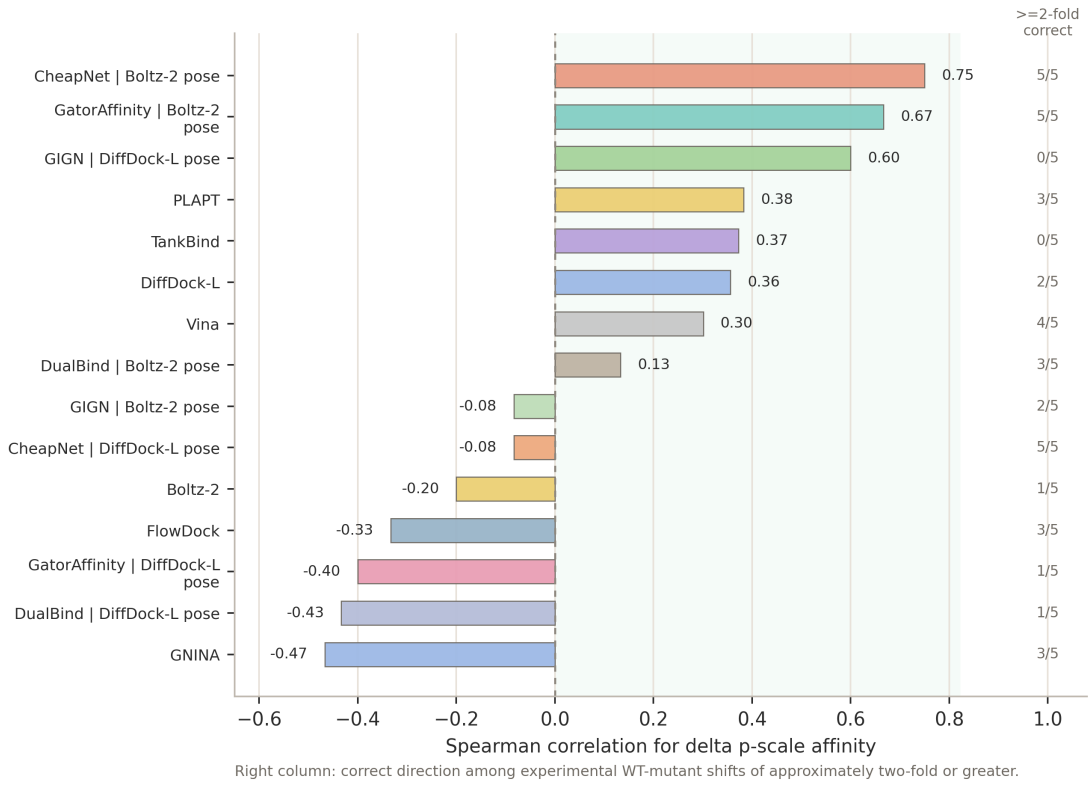


Figure 4.5. Spearman rank correlation between predicted and experimental WT–mutant Δp values across nine paired cases, for each tool and pose. The shaded green region marks positive correlation, tools to the right correctly track the direction and magnitude of mutation-driven potency shifts. The right numeric column (x/5) shows how many of the ($|\Delta p| \geq 0.30$, approximately two-fold or greater shift in potency) were signed in the correct direction.

CheapNet on Boltz-2 poses produced the strongest Δp signal ($\rho = 0.75$, Figure 4.5) and correctly signed all five significant Δp values. GatorAffinity on Boltz-2 poses ranked second ($\rho = 0.67$), also with 5/5 correct directions. TankBind ($\rho = 0.37$) and DualBind on Boltz-2 poses ($\rho = 0.13$) did not recover the mutation-driven shifts. Two observations stand out. First, the tools that led on Δp capture were not the same as those that led on absolute p-scale ranking; mutation response and ligand-affinity ranking are separate capabilities, so strength on one axis does not imply strength on the other. Second, pose source mattered: the leading rescoring tools used Boltz-2 poses, while the same models on DiffDock-L poses showed substantially weaker results.

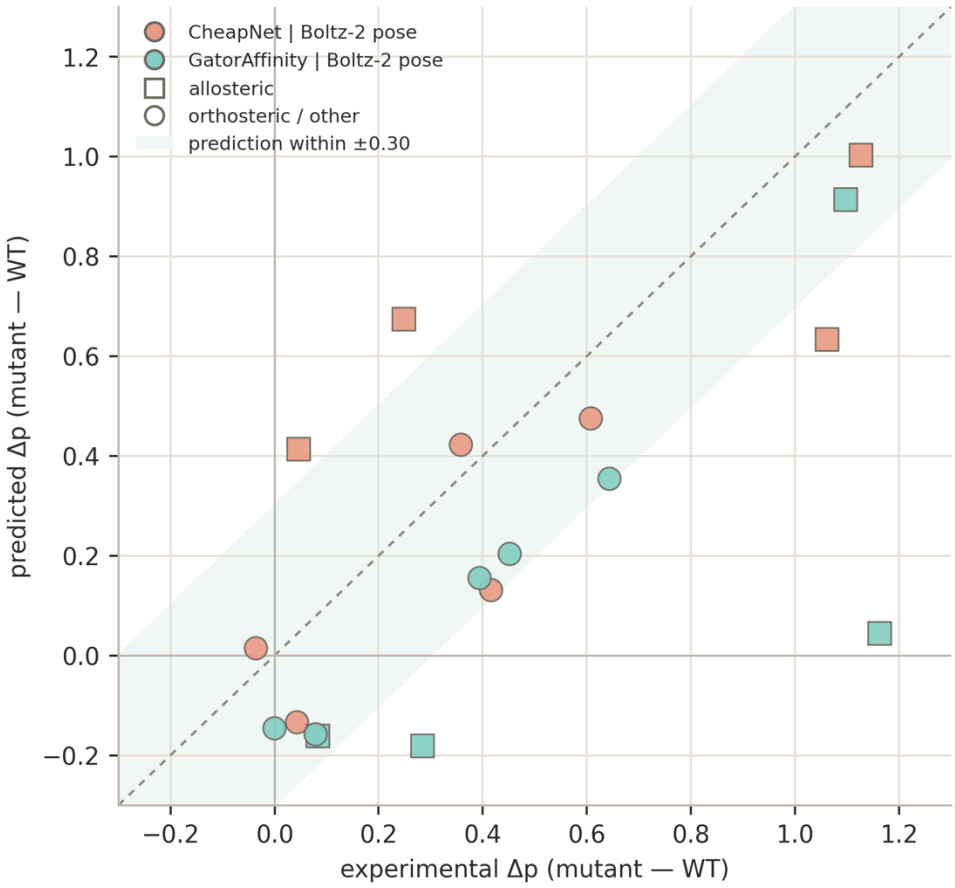


Figure 4.6. Predicted versus experimental Δp for CheapNet and GatorAffinity on Boltz-2 poses. The dashed line is perfect agreement (predicted = experimental); the shaded band marks $|\Delta p| \geq 0.30$, the two-fold shift threshold. Squares are allosteric cases (RLY-2608, STX-478); circles are orthosteric pocket ligands.

4.2.4 Pose correctness

Score-level performance can be inherited from a plausible underlying pose, so pose correctness was evaluated separately on the six crystallographic references in Section 3.5.3 (three orthosteric, three allosteric). LRMSD after local pocket alignment was used as the primary geometric metric, with LRMSD ≤ 2 Å treated as successful pose recovery. PoseBusters failure-mode breakdown were used as complementary measures.

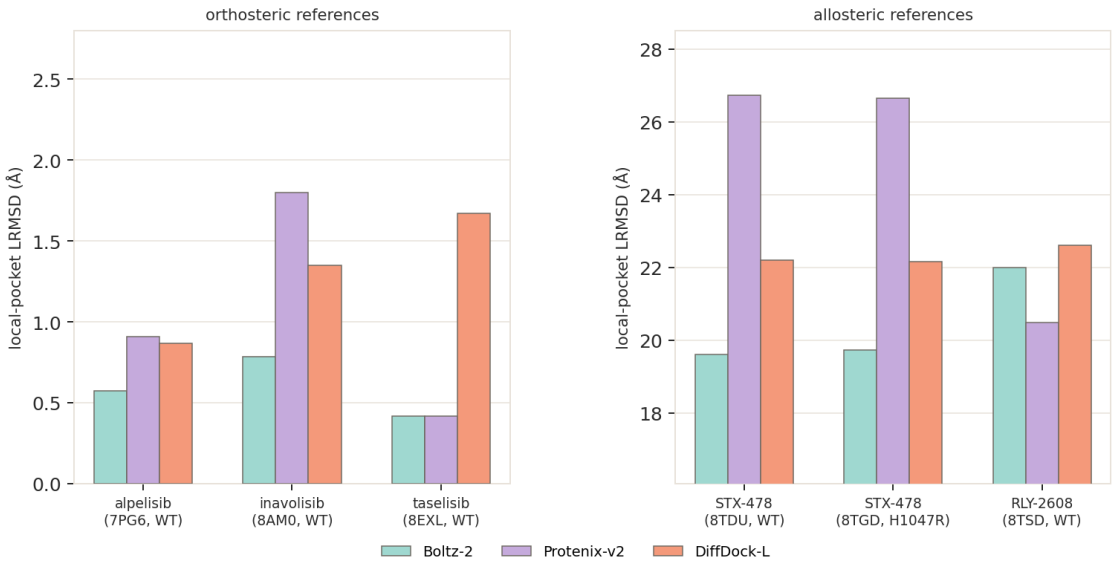


Figure 4.7. LRMSD for Boltz-2, Protenix, and DiffDock-L across six crystallographic references. Left: orthosteric references; right: allosteric references (note the different y-axis scales). Lower LRMSD indicates closer agreement with the crystal pose.

All three tools recovered the orthosteric crystal poses within 2 Å (Figure 4.7), Boltz-2 was the most accurate (0.58 Å for alpelisib), followed by DiffDock-L and Protenix. None recovered the allosteric pocket: every allosteric comparison returned LRMSD above 19 Å, consistent with wrong-site placement in the orthosteric ATP pocket rather than the allosteric site. PoseBusters were consistent with the LRMSD results, except for Protenix, which failed at least one PoseBusters check on 35.5% of its top poses (Boltz-2: 3.5%) when run without training-free guidance (--use_tfg_guidance), producing chemically invalid ligand geometries despite correct pocket placement and therefore was dropped from further evaluation.

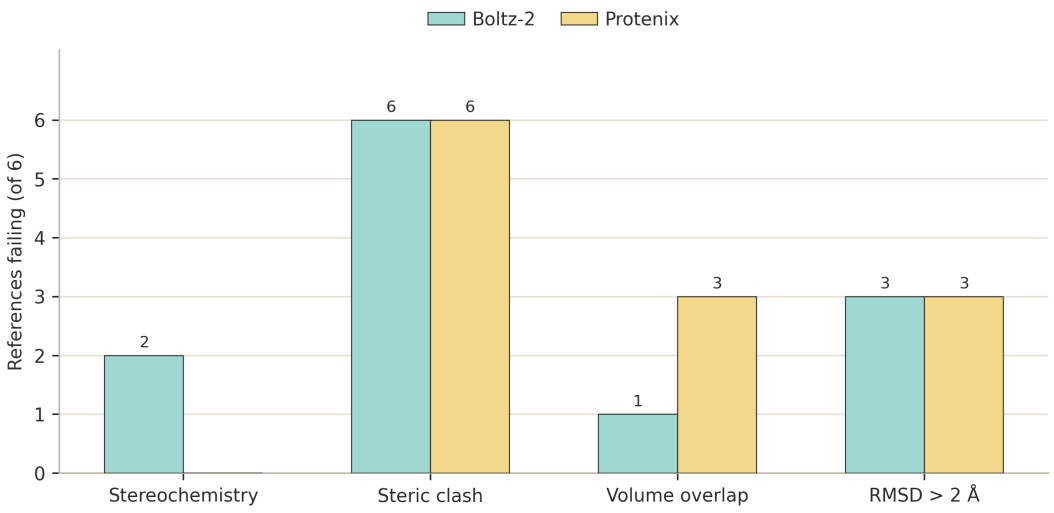


Figure 4.8. PoseBusters failure-mode breakdown across the six crystallographic references, for the two co-folding pose generators (Boltz-2 and Protenix). Bars show, per tool, the number of references (of 6) in which at least one PoseBusters check in the named category failed. The steric-clash category uses a very tight pairwise-distance threshold and is tripped on every reference for every cofolding tool, so it is reported as context rather than as a discriminating metric. Volume overlap, RMSD > 2 Å, and stereochemistry are the categories that separate the two tools.

Four PoseBusters categories were assessed: steric clash (atom pairs below a van-der-Waals threshold), volume overlap (ligand atoms inside protein density), RMSD > 2 Å (pose displaced from crystal reference), and stereochemistry (incorrect chirality or geometry). Steric clash failed on every reference for both tools, an artefact of the tight threshold, not a real clash. The other three separated them (Figure 4.8): Boltz-2 had one volume-overlap and two stereochemistry failures (inavolisib and STX-478 in 8TGD); Protenix had three volume-overlap failures and none for stereochemistry. Both failed RMSD on the same three allosteric references (wrong-site placement, Figure 4.7). Combined with its higher failure rate across the benchmark, PoseBusters supports dropping Protenix from downstream evaluation.

Boltz-2 was selected as the primary pose source on three grounds: lowest orthosteric LRMSD (0.58 Å vs. 0.91 Å), fewer volume-overlap failures (1 vs. 3 of 6), and stereochemistry failures limited to two references (inavolisib and STX-478 in 8TGD) rather than a systematic deficiency. Protenix avoided that chirality error but at the cost of worse LRMSD and more volume overlap. These chirality errors are a noted limitation and do not affect the affinity-ranking conclusions.

Overall, pose correctness split sharply by binding-site class: all three pose generators recovered the orthosteric pocket; none recovered the allosteric pocket on any reference or metric. Any score-level claim about RLY-2608 or STX-478 therefore rests on a pose that does not localise to the experimentally observed site.

4.2.5 Allosteric Pocket Recognition

The benchmark asks each tool to place a ligand into one of two structurally distinct PIK3CA pockets. The orthosteric ATP-site cleft sits between the N- and C-lobes of the kinase domain and is the binding region for the four PI3Kα-selective and the four pan-PI3K inhibitors in the panel. The allosteric pocket lies on the opposite face of the

kinase domain, is cryptic in the apo form, and is the binding region of RLY-2608 and STX-478.

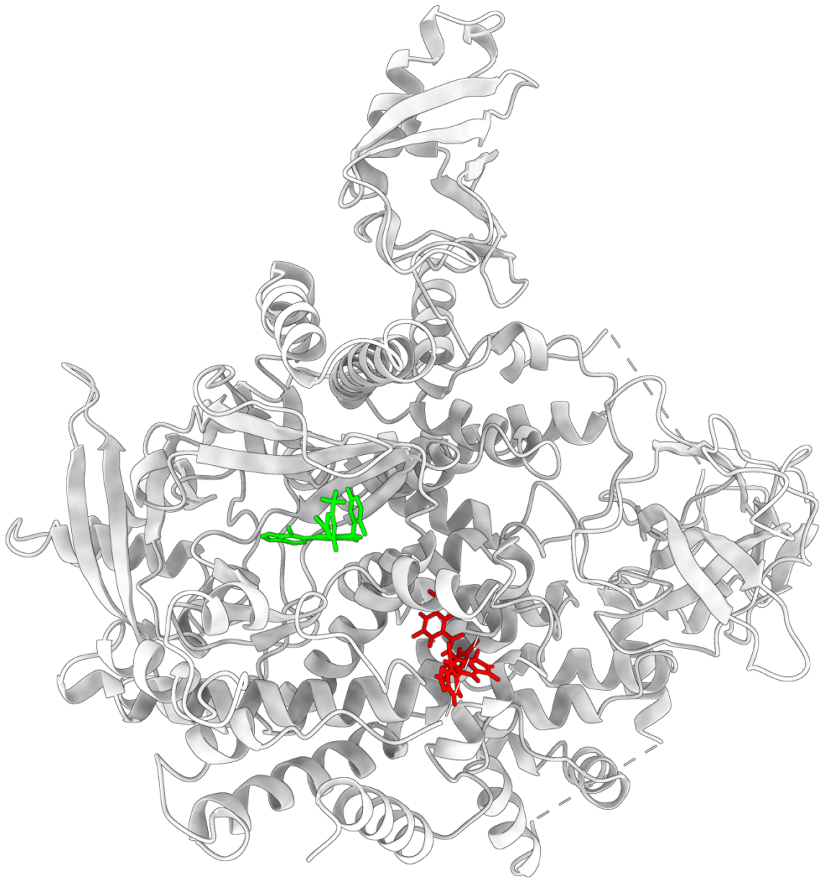


Figure 4.9. Pocket Blindness of Boltz-2 on the Allosteric Inhibitor RLY-2608. Boltz-2 predicted complex superimposed on the crystal structure (PDB: 8TSD). The crystal pose (red) places RLY-2608 in the allosteric site, whereas the Boltz-2 prediction (green) places it in the orthosteric ATP pocket with a confidence score of 0.91.

The orthosteric site is conserved across all variants and is well represented in kinase-inhibitor training data, making it familiar to both docking and co-folding models. The allosteric site, by contrast, is largely absent from ligand-bound PIK3CA structures and requires a conformational rearrangement not apparent from the apo backbone; most tools will have encountered few or no examples of it.

5. Conclusion and Future Scope

5.1 Summary of Findings

A 300-pair benchmark matrix was constructed by crossing 15 p110α variants with 20 ligands spanning approved PI3Kα inhibitors, clinical candidates, allosteric compounds, isoform-selective agents, kinase decoys, and structural negative controls. A high-confidence biochemical affinity subset (n = 19) was assembled from recombinant PI3Kα assay measurements in primary literature and BindingDB; nine WT–mutant delta pairs were derived from the same sources; and six crystallographic structures provided orthosteric and allosteric pose correctness ground truth.

The evaluated scoring and pose-generation panel separated across four independent axes. GatorAffinity on Boltz-2 poses produced the strongest p-scale Spearman correlation ($\rho = 0.855$). On cross-docked specificity, GatorAffinity led on AUROC (0.879) and DiffDock-L led on average precision (0.789). On mutation response, CheapNet on Boltz-2 poses produced the strongest Δp Spearman correlation ($\rho = 0.750$) and, like GatorAffinity ($\rho = 0.667$), correctly signed all five significant Δp values ($|\Delta p| \geq 0.30$). On pose correctness, Boltz-2, Protenix, and DiffDock-L all reached low LRMSD on orthosteric references, while the four pose generators placed the allosteric ligands in the orthosteric site across all three allosteric test cases.

5.2 Conclusions

The central question of this study is, when a mutation remodels the PIK3CA active site, does the model register that the same inhibitor now engages a different target? The Δp analysis tested this directly by asking whether predicted scores changed in the correct direction when the same inhibitor was compared between wild-type and mutant p110α. CheapNet and GatorAffinity on Boltz-2 poses captured the direction of all five significant Δp values, with moderate rank agreement for the size of these shifts. This is the most important, if preliminary, signal: the models were able to detect small mutation-dependent changes, from modest shifts such as inavolisib E542K to larger effects such as STX-478 H1047R, approaching at an experimental scale.

The broader affinity and specificity results support this conclusion GatorAffinity on Boltz-2 poses gave the strongest experimental p-scale rank, while GatorAffinity and

DiffDock-L performed best on broad specificity by AUROC and average precision. These results indicate that current tools can prioritise PI3Kα-relevant inhibitors and recover useful potency. Boltz-2, Protenix, and DiffDock-L recovered orthosteric crystal poses with low LRMSD, showing that current pose generators are strong when the binding site is conserved and well represented in the training data corpus, however, all allosteric references failed with LRMSD values above 19 Å. The RLY-2608 allosteric inhibitor also shows the danger of overconfident in these generative models even after wrong placement. The idelalisib binding to delta- isoform and not the alpha-isoform compliments it.

Further structural validation with PoseBusters showed that generated complexes still contain chemically important failure modes, including steric-clash flags, volume overlap, and stereochemical errors. This demonstrates that even when a model places a ligand in a plausible binding pocket or assigns a high confidence score, the resulting complex is not necessarily physically valid at the atomistic level. More broadly, these errors also show why model scaling alone is not sufficient. Larger training sets should improve protein–ligand prediction, and recent affinity models already suggest performance gains from large-scale structural pretraining, but the benefit depends on what is being scaled. If the data are dominated only by well-sampled protein families, it will only make models better at those common cases. Rare but important cases, such as mutant conformations, isoform-selective binding, inactive pairs, and allosteric states, need equally high-quality representation in training data.

5.3 Limitations

Several limitations bound the interpretation of the findings reported in this thesis. First, the p-scale subset is small and endpoint-mixed (n = 19; 14 IC□□-derived and 5 K_i-derived). The K_i-only subset has only five rows, which is insufficient for K_i-specific claims; it is used only as a caveat check. Bootstrap 95 % intervals for the Spearman ρ values span approximately ±0.30 for most tools on this subset.

Second, the Δp analysis uses nine paired WT–mutant cases, of which five meet the $|\Delta p| \geq 0.30$ significance threshold. A single mis-signed prediction changes the sign accuracy by 20 percentage points. Permutation p-values remain above 0.05 for most

tools even when sign accuracy appears high, so Δp results are reported as preliminary signals rather than as established performance estimates.

Third, the broad cross-docked specificity labels are heterogeneous. The 52 broad positives include biochemical numeric evidence, qualitative activity, resistance-context, and structural support. Tools that rank any PIK3CA-associated case highly will benefit from this breadth. The stricter biochemical-positive versus negative subset (n = 19 vs n = 150) consistently produces lower AUROC and AP values and is the more defensible specificity estimate for biochemical claims.

Fourth, all pose correctness conclusions rest on six crystallographic references (three orthosteric and three allosteric). The orthosteric success rate is consistent across tools, but the allosteric failure pattern, while uniform, rests on three references per pose source and cannot be quantified as a precise rate estimate. Failure on STX-478 and RLY-2608 may partly reflect the absence of public allosteric PIK3CA training data rather than a fundamental algorithmic limit.

Finally, the mutant input structures used for classical docking inputs were produced by in-silico residue substitution on the WT AlphaFold backbone. Backbone drift between WT and mutant inputs was small. Tools that respond to mutation on these inputs therefore respond mainly to local side-chain geometry and energy rather than to global conformational rearrangement. Whether these local signals generalise across the full oncogenic hotspot and resistance panel remains an open question that would require a larger paired experimental dataset to resolve.

5.4 Future Directions

Allosteric pose prediction remains the principal open problem. Across the pose generators evaluated here, allosteric LRMSD stayed above 19 Å in every case, consistent with the near-absence of PIK3CA allosteric co-crystal structures in public training data. The near-term priority is to generate such structures and use them as fine-tuning data for co-folding models or as structural templates for a pocket-opening docking step applied before any blind co-folding run.

A post-processing geometry-repair step using PLIP or force-field optimisation would be straightforward to implement and is a practical prerequisite for applying ABFE or RBFE calculations to the high-confidence biochemical subset, which would directly test whether the GatorAffinity correlation ceiling reflects a limit of pose quality or of the scoring function.

Finally, because the benchmark is organised around clinically observed variants rather than PIK3CA-specific chemistry, the 300-pair matrix design is transferable to other oncogenic kinase targets with comparable clinical variant and resistance-mutation data; EGFR T790M/C797S, BRAF V600E, and ALK resistance panels are natural candidates. Applying the same framework to one or two additional targets would test whether the performance profiles observed here generalise beyond the PIK3CA system to mutation-aware affinity prediction as a general task.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., … Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, *630*(8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w

André, F., Ciruelos, E., Rubovszky, G., Campone, M., Loibl, S., Rugo, H. S., Iwata, H., Conte, P., Mayer, I. A., Kaufman, B., Yamashita, T., Lu, Y.-S., Inoue, K., Takahashi, M., Pápai, Z., Longin, A.-S., Mills, D., Wilke, C., Hirawat, S., & Juric, D. (2019). Alpelisib for PIK3CA-Mutated, Hormone Receptor–Positive Advanced Breast Cancer. *New England Journal of Medicine*, *380*(20), 1929–1940. https://doi.org/10.1056/NEJMoa1813904

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krüger, F. A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., & Overington, J. P. (2014). The ChEMBL bioactivity database: An update. *Nucleic Acids Research*, *42*(D1), D1083–D1090. https://doi.org/10.1093/nar/gkt1031

Buckbinder, L., St Jean, D. J., Tieu, T., Ladd, B., Hilbert, B., Wang, W., Alltucker, J. T., Manimala, S., Kryukov, G. V., Brooijmans, N., Dowdell, G., Jonsson, P., Huff, M., Guzman-Perez, A., Jackson, E. L., Goncalves, M. D., & Stuart, D. D. (2023a). STX-478, a Mutant-Selective, Allosteric PI3Kα Inhibitor Spares Metabolic Dysfunction and Improves Therapeutic Response in PI3Kα-Mutant

Xenografts. *Cancer Discovery*, *13*(11), 2432–2447. https://doi.org/10.1158/2159-8290.CD-23-0396

Buckbinder, L., St Jean, D. J., Tieu, T., Ladd, B., Hilbert, B., Wang, W., Alltucker, J. T., Manimala, S., Kryukov, G. V., Brooijmans, N., Dowdell, G., Jonsson, P., Huff, M., Guzman-Perez, A., Jackson, E. L., Goncalves, M. D., & Stuart, D. D. (2023b). STX-478, a Mutant-Selective, Allosteric PI3Kα Inhibitor Spares Metabolic Dysfunction and Improves Therapeutic Response in PI3Kα-Mutant Xenografts. *Cancer Discovery*, *13*(11), 2432–2447. https://doi.org/10.1158/2159-8290.CD-23-0396

Buttenschoen, M., Morris, G. M., & Deane, C. M. (2023, August 10). *PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences*. arXiv.Org. https://doi.org/10.1039/D3SC04185A

Corso, G., Stärk, H., Jing, B., Barzilay, R., & Jaakkola, T. (2022, October 4). *DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking*. arXiv.Org. https://arxiv.org/abs/2210.01776v2

Dogruluk, T., Tsang, Y. H., Espitia, M., Chen, F., Chen, T., Chong, Z., Appadurai, V., Dogruluk, A., Eterovic, A. K., Bonnen, P. E., Creighton, C. J., Chen, K., Mills, G. B., & Scott, K. L. (2015). Identification of Variant-Specific Functions of PIK3CA by Rapid Phenotyping of Rare Mutations. *Cancer Research*, *75*(24), 5341–5354. https://doi.org/10.1158/0008-5472.CAN-15-1654

Druker, B. J., Talpaz, M., Resta, D. J., Peng, B., Buchdunger, E., Ford, J. M., Lydon, N. B., Kantarjian, H., Capdeville, R., Ohno-Jones, S., & Sawyers, C. L.

(2001). Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *The New England Journal of Medicine*, *344*(14), 1031–1037. https://doi.org/10.1056/NEJM200104053441401

Folkes, A. J., Ahmadi, K., Alderton, W. K., Alix, S., Baker, S. J., Box, G., Chuckowree, I. S., Clarke, P. A., Depledge, P., Eccles, S. A., Friedman, L. S., Hayes, A., Hancox, T. C., Kugendradas, A., Lensun, L., Moore, P., Olivero, A. G., Pang, J., Patel, S., … Shuttleworth, S. J. (2008). The Identification of 2-(1H-Indazol-4-yl)-6-(4-methanesulfonyl-piperazin-1-ylmethyl)-4-morpholin-4-yl-thieno[3,2-d]pyrimidine (GDC-0941) as a Potent, Selective, Orally Bioavailable Inhibitor of Class I PI3 Kinase for the Treatment of Cancer. *Journal of Medicinal Chemistry*, *51*(18), 5522–5532. https://doi.org/10.1021/jm800295d

Fong, P. C., Boss, D. S., Yap, T. A., Tutt, A., Wu, P., Mergui-Roelvink, M., Mortimer, P., Swaisland, H., Lau, A., O'Connor, M. J., Ashworth, A., Carmichael, J., Kaye, S. B., Schellens, J. H. M., & de Bono, J. S. (2009). Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *The New England Journal of Medicine*, *361*(2), 123–134. https://doi.org/10.1056/NEJMoa0900212

Fritsch, C., Huang, A., Chatenay-Rivauday, C., Schnell, C., Reddy, A., Liu, M., Kauffmann, A., Guthy, D., Erdmann, D., De Pover, A., Furet, P., Gao, H., Ferretti, S., Wang, Y., Trappe, J., Brachmann, S. M., Maira, S.-M., Wilson, C., Boehm, M., … Sellers, W. R. (2014). Characterization of the Novel and Specific PI3Kα Inhibitor NVP-BYL719 and Development of the Patient

Stratification Strategy for Clinical Trials. *Molecular Cancer Therapeutics*, *13*(5), 1117–1129. https://doi.org/10.1158/1535-7163.MCT-13-0865

Fry, D. W., Harvey, P. J., Keller, P. R., Elliott, W. L., Meade, M., Trachet, E., Albassam, M., Zheng, X., Leopold, W. R., Pryer, N. K., & Toogood, P. L. (2004). Specific inhibition of cyclin-dependent kinase 4/6 by PD 0332991 and associated antitumor activity in human tumor xenografts. *Molecular Cancer Therapeutics*, *3*(11), 1427–1438. https://doi.org/10.1158/1535-7163.1427.3.11

Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2016). BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, *44*(D1), D1045–D1053. https://doi.org/10.1093/nar/gkv1072

Gymnopoulos, M., Elsliger, M.-A., & Vogt, P. K. (2007). Rare cancer-specific mutations in PIK3CA show gain of function. *Proceedings of the National Academy of Sciences*, *104*(13), 5569–5574. https://doi.org/10.1073/pnas.0701005104

Hanan, E. J., Braun, M.-G., Heald, R. A., MacLeod, C., Chan, C., Clausen, S., Edgar, K. A., Eigenbrot, C., Elliott, R., Endres, N., Friedman, L. S., Gogol, E., Gu, X.-H., Thibodeau, R. H., Jackson, P. S., Kiefer, J. R., Knight, J. D., Nannini, M., Narukulla, R., … Staben, S. T. (2022). Discovery of GDC-0077 (Inavolisib), a Highly Selective Inhibitor and Degrader of Mutant PI3Kα. *Journal of Medicinal Chemistry*, *65*(24), 16589–16621. https://doi.org/10.1021/acs.jmedchem.2c01422

Hopkins, B. D., Pauli, C., Du, X., Wang, D. G., Li, X., Wu, D., Amadiume, S. C., Goncalves, M. D., Hodakoski, C., Lundquist, M. R., Bareja, R., Ma, Y., Harris, E. M., Sboner, A., Beltran, H., Rubin, M. A., Mukherjee, S., & Cantley, L. C. (2018). Suppression of insulin feedback enhances the efficacy of PI3K inhibitors. *Nature*, *560*(7719), 499–503. https://doi.org/10.1038/s41586-018-0343-4

Huang, C.-H., Mandelker, D., Schmidt-Kittler, O., Samuels, Y., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., Gabelli, S. B., & Amzel, L. M. (2007). The structure of a human p110alpha/p85alpha complex elucidates the effects of oncogenic PI3Kalpha mutations. *Science*, *318*(5857), 1744–1748. https://doi.org/10.1126/science.1150799

Jessen, K., Kessler, L., Kucharski, J., Guo, X., Staunton, J., Janes, M., Elia, M., Banerjee, U., Lan, L., Wang, S., Stewart, J., Luzader, A., Darjania, L., Li, L., Chan, K., Martin, M., Ren, P., Rommel, C., & Liu, Y. (2011). Abstract A171: A potent and selective PI3K inhibitor, INK1117, targets human cancers harboring oncogenic PIK3CA mutations. *Molecular Cancer Therapeutics*, *10*(11_Supplement), A171. https://doi.org/10.1158/1535-7163.TARG-11-A171

Jiang, Y., Li, X., Zhang, Y., Han, J., Xu, Y., Pandit, A., Zhang, Z., Wang, M., Wang, M., Shen, M., Yang, G., Choi, Y., Li, W.-J., Fu, T., Wu, F., & Liu, J. (2026). *PoseX: AI Defeats Physics Approaches on Protein-Ligand Cross Docking* (arXiv:2505.01700). arXiv. https://doi.org/10.48550/arXiv.2505.01700

Joseph, E. W., Pratilas, C. A., Poulikakos, P. I., Tadi, M., Wang, W., Taylor, B. S., Halilovic, E., Persaud, Y., Xing, F., Viale, A., Tsai, J., Chapman, P. B.,

Bollag, G., Solit, D. B., & Rosen, N. (2010). The RAF inhibitor PLX4032 inhibits ERK signaling and tumor cell proliferation in a V600E BRAF-selective manner. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(33), 14903–14908. https://doi.org/10.1073/pnas.1008990107

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Juric, D., de Bono, J. S., LoRusso, P. M., Nemunaitis, J., Heath, E. I., Kwak, E. L., Macarulla Mercadé, T., Geuna, E., Jose de Miguel-Luken, M., Patel, C., Kuida, K., Sankoh, S., Westin, E. H., Zohren, F., Shou, Y., & Tabernero, J. (2017). A First-in-Human, Phase I, Dose-Escalation Study of TAK-117, a Selective PI3Kα Isoform Inhibitor, in Patients with Advanced Solid Malignancies. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *23*(17), 5015–5023. https://doi.org/10.1158/1078-0432.CCR-16-2888

Knight, Z. A., Gonzalez, B., Feldman, M. E., Zunder, E. R., Goldenberg, D. D., Williams, O., Loewith, R., Stokoe, D., Balla, A., Toth, B., Balla, T., Weiss, W. A., Williams, R. L., & Shokat, K. M. (2006). A Pharmacological Map of the PI3-K Family Defines a Role for p110α in Insulin Signaling. *Cell*, *125*(4), 733–747. https://doi.org/10.1016/j.cell.2006.03.035

Kong, D., & Yamori, T. (2007). ZSTK474 is an ATP-competitive inhibitor of class I phosphatidylinositol 3 kinase isoforms. *Cancer Science*, *98*(10), 1638–1642. https://doi.org/10.1111/j.1349-7006.2007.00580.x

Kramer, C., Kalliokoski, T., Gedeck, P., & Vulpetti, A. (2012). The Experimental Uncertainty of Heterogeneous Public Ki Data. *Journal of Medicinal Chemistry*, *55*(11), 5165–5173. https://doi.org/10.1021/jm300131x

Lannutti, B. J., Meadows, S. A., Herman, S. E. M., Kashishian, A., Steiner, B., Johnson, A. J., Byrd, J. C., Tyner, J. W., Loriaux, M. M., Deininger, M., Druker, B. J., Puri, K. D., Ulrich, R. G., & Giese, N. A. (2011). CAL-101, a p110delta selective phosphatidylinositol-3-kinase inhibitor for the treatment of B-cell malignancies, inhibits PI3K signaling and cellular viability. *Blood*, *117*(2), 591–594. https://doi.org/10.1182/blood-2010-03-275305

Lim, H., Kim, S., & Lee, S. (2024, October 4). *CheapNet: Cross-attention on Hierarchical representations for Efficient protein-ligand binding Affinity Prediction*. The Thirteenth International Conference on Learning Representations. https://openreview.net/forum?id=A1HhtITVEi

Liu, M., & Paliwal, S. G. (2024, June 11). *DualBind: A Dual-Loss Framework for Protein-Ligand Binding Affinity Prediction*. arXiv.Org. https://arxiv.org/abs/2406.07770v1

Liu, N., Rowley, B. R., Bull, C. O., Schneider, C., Haegebarth, A., Schatz, C. A., Fracasso, P. R., Wilkie, D. P., Hentemann, M., Wilhelm, S. M., Scott, W. J., Mumberg, D., & Ziegelbauer, K. (2013). BAY 80-6946 is a highly selective intravenous PI3K inhibitor with potent p110α and p110δ activities in tumor

cell lines and xenograft models. *Molecular Cancer Therapeutics*, *12*(11), 2319–2330. https://doi.org/10.1158/1535-7163.MCT-12-0993-T

Lu, W., Wu, Q., Zhang, J., Rao, J., Li, C., & Zheng, S. (2022). *TANKBind: Trigonometry-Aware Neural NetworKs for Drug-Protein Binding Structure Prediction* (p. 2022.06.06.495043). bioRxiv. https://doi.org/10.1101/2022.06.06.495043

Maira, S.-M., Pecchi, S., Huang, A., Burger, M., Knapp, M., Sterker, D., Schnell, C., Guthy, D., Nagel, T., Wiesmann, M., Brachmann, S., Fritsch, C., Dorsch, M., Chène, P., Shoemaker, K., De Pover, A., Menezes, D., Martiny-Baron, G., Fabbro, D., … Voliva, C. F. (2012). Identification and characterization of NVP-BKM120, an orally available pan-class I PI3-kinase inhibitor. *Molecular Cancer Therapeutics*, *11*(2), 317–328. https://doi.org/10.1158/1535-7163.MCT-11-0474

McNutt, A. T., Li, Y., Meli, R., Aggarwal, R., & Koes, D. R. (2025). GNINA 1.3: The next increment in molecular docking with deep learning. *Journal of Cheminformatics*, *17*, 28. https://doi.org/10.1186/s13321-025-00973-x

Menteş, M., Karakuzulu, B. B., Uçar, G. B., & Yandım, C. (2022). Comparative molecular dynamics analyses on PIK3CA hotspot mutations with PI3Kα specific inhibitors and ATP. *Computational Biology and Chemistry*, *99*, 107726. https://doi.org/10.1016/j.compbiolchem.2022.107726

Morehead, A., & Cheng, J. (2025). *FlowDock: Geometric Flow Matching for Generative Protein-Ligand Docking and Affinity Prediction* (arXiv:2412.10966). arXiv. https://doi.org/10.48550/arXiv.2412.10966

Ndubaku, C. O., Heffron, T. P., Staben, S. T., Baumgardner, M., Blaquiere, N., Bradley, E., Bull, R., Do, S., Dotson, J., Dudley, D., Edgar, K. A., Friedman, L. S., Goldsmith, R., Heald, R. A., Kolesnikov, A., Lee, L., Lewis, C., Nannini, M., Nonomiya, J., … Olivero, A. G. (2013). Discovery of 2-{3-[2-(1-Isopropyl-3-methyl-1H-1,2–4-triazol-5-yl)-5,6-dihydrobenzo[f]imidazo[1,2-d][1,4]oxazepin-9-yl]-1H-pyrazol-1-yl}-2-methylpropanamide (GDC-0032): A β-Sparing Phosphoinositide 3-Kinase Inhibitor with High Unbound Exposure and Robust in Vivo Antitumor Activity. *Journal of Medicinal Chemistry*, *56*(11), 4597–4610. https://doi.org/10.1021/jm4003632

Oda, K., Okada, J., Timmerman, L., Rodriguez-Viciana, P., Stokoe, D., Shoji, K., Taketani, Y., Kuramoto, H., Knight, Z. A., Shokat, K. M., & McCormick, F. (2008). PIK3CA cooperates with other phosphatidylinositol 3'-kinase pathway mutations to effect oncogenic transformation. *Cancer Research*, *68*(19), 8127–8136. https://doi.org/10.1158/0008-5472.CAN-08-0755

Osborne, C. K. (1998). Tamoxifen in the treatment of breast cancer. *The New England Journal of Medicine*, *339*(22), 1609–1618. https://doi.org/10.1056/NEJM199811263392207

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., Kwabi-Addo, D., Beaini, D., Jaakkola, T., & Barzilay, R. (2025). *Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction* (p. 2025.06.14.659707). bioRxiv. https://doi.org/10.1101/2025.06.14.659707

Pires, D. E. V., Blundell, T. L., & Ascher, D. B. (2015a). Platinum: A database of experimentally measured effects of mutations on structurally defined protein-

ligand complexes. *Nucleic Acids Research*, *43*(Database issue), D387-391. https://doi.org/10.1093/nar/gku966

Pires, D. E. V., Blundell, T. L., & Ascher, D. B. (2015b). Platinum: A database of experimentally measured effects of mutations on structurally defined protein–ligand complexes. *Nucleic Acids Research*, *43*(Database issue), D387–D391. https://doi.org/10.1093/nar/gku966

Rogers, D., & Hahn, M. (2010). Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, *50*(5), 742–754. https://doi.org/10.1021/ci100050t

Rose, T., Monti, N., Anand, N., & Shen, T. (2024). *PLAPT: Protein-Ligand Binding Affinity Prediction Using Pretrained Transformers* (p. 2024.02.08.575577). bioRxiv. https://doi.org/10.1101/2024.02.08.575577

Samuels, Y., Wang, Z., Bardelli, A., Silliman, N., Ptak, J., Szabo, S., Yan, H., Gazdar, A., Powell, S. M., Riggins, G. J., Willson, J. K. V., Markowitz, S., Kinzler, K. W., Vogelstein, B., & Velculescu, V. E. (2004). High frequency of mutations of the PIK3CA gene in human cancers. *Science*, *304*(5670), 554. https://doi.org/10.1126/science.1096502

Schindler, C. E. M., Baumann, H., Blum, A., Böse, D., Buchstaller, H.-P., Burgdorf, L., Cappel, D., Chekler, E., Czodrowski, P., Dorsch, D., Eguida, M. K. I., Follows, B., Fuchß, T., Grädler, U., Gunera, J., Johnson, T., Jorand Lebrun, C., Karra, S., Klein, M., … Kuhn, D. (2020). Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. *Journal*

of Chemical Information and Modeling, *60*(11), 5457–5474. https://doi.org/10.1021/acs.jcim.0c00900

Shepherd, F. A., Rodrigues Pereira, J., Ciuleanu, T., Tan, E. H., Hirsh, V., Thongprasert, S., Campos, D., Maoleekoonpiroj, S., Smylie, M., Martins, R., van Kooten, M., Dediu, M., Findlay, B., Tu, D., Johnston, D., Bezjak, A., Clark, G., Santabárbara, P., Seymour, L., & National Cancer Institute of Canada Clinical Trials Group. (2005). Erlotinib in previously treated non-small-cell lung cancer. *The New England Journal of Medicine*, *353*(2), 123–132. https://doi.org/10.1056/NEJMoa050753

Shin, N., Stubbs, M., Koblish, H., Yue, E. W., Soloviev, M., Douty, B., Wang, K. H., Wang, Q., Gao, M., Feldman, P., Yang, G., Hall, L., Hansbury, M., O'Connor, S., Leffet, L., Collins, R., Katiyar, K., He, X., Waeltz, P., … Scherle, P. (2020). Parsaclisib Is a Next-Generation Phosphoinositide 3-Kinase δ Inhibitor with Reduced Hepatotoxicity and Potent Antitumor and Immunomodulatory Activities in Models of B-Cell Malignancy. *The Journal of Pharmacology and Experimental Therapeutics*, *374*(1), 211–222. https://doi.org/10.1124/jpet.120.265538

Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., & Wang, R. (2019). Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling*, *59*(2), 895–913. https://doi.org/10.1021/acs.jcim.8b00545

Team, B. A. A., Chen, X., Zhang, Y., Lu, C., Ma, W., Guan, J., Gong, C., Yang, J., Zhang, H., Zhang, K., Wu, S., Zhou, K., Yang, Y., Liu, Z., Wang, L., Shi, B., Shi, S., & Xiao, W. (2025). *Protenix—Advancing Structure Prediction*

Through a Comprehensive AlphaFold3 Reproduction (p. 2025.01.08.631967). bioRxiv. https://doi.org/10.1101/2025.01.08.631967

Trott, O., & Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of Computational Chemistry*, *31*(2), 455–461. https://doi.org/10.1002/jcc.21334

Vanhaesebroeck, B., Guillermet-Guibert, J., Graupera, M., & Bilanges, B. (2010). The emerging mechanisms of isoform-specific PI3K signalling. *Nature Reviews. Molecular Cell Biology*, *11*(5), 329–341. https://doi.org/10.1038/nrm2882

Varkaris, A., Fece de la Cruz, F., Martin, E. E., Norden, B. L., Chevalier, N., Kehlmann, A. M., Leshchiner, I., Barnes, H., Ehnstrom, S., Stavridi, A.-M., Yuan, X., Kim, J. S., Ellis, H., Papatheodoridi, A., Gunaydin, H., Danysh, B. P., Parida, L., Sanidas, I., Ji, Y., … Juric, D. (2024). Allosteric PI3Kα Inhibition Overcomes On-target Resistance to Orthosteric Inhibitors Mediated by Secondary PIK3CA Mutations. *Cancer Discovery*, *14*(2), 227–239. https://doi.org/10.1158/2159-8290.CD-23-0704

Varkaris, A., Pazolli, E., Gunaydin, H., Wang, Q., Pierce, L., Boezio, A. A., Bulku, A., DiPietro, L., Fridrich, C., Frost, A., Giordanetto, F., Hamilton, E. P., Harris, K., Holliday, M., Hunter, T. L., Iskandar, A., Ji, Y., Larivée, A., LaRochelle, J. R., … Kipp, D. R. (2024). Discovery and Clinical Proof-of-Concept of RLY-2608, a First-in-Class Mutant-Selective Allosteric PI3Kα Inhibitor That Decouples Antitumor Activity from Hyperinsulinemia. *Cancer Discovery*, *14*(2), 240–257. https://doi.org/10.1158/2159-8290.CD-23-0944

Wan, S., Zhang, X., Xue, X., & Coveney, P. V. (2026, March 2). *On the Reliability of AI Methods in Drug Discovery: Evaluation of Boltz-2 for Structure and Binding Affinity Prediction*. arXiv.Org. https://arxiv.org/abs/2603.05532v1

Wang, R., Fang, X., Lu, Y., & Wang, S. (2004). The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *Journal of Medicinal Chemistry*, *47*(12), 2977–2980. https://doi.org/10.1021/jm030580l

Wei, J., Zhang, Y., Ramdhan, P. A., Huang, Z., Seabra, G., Jiang, Z., Li, C., & Li, Y. (2025). *GatorAffinity: Boosting Protein-Ligand Binding Affinity Prediction with Large-Scale Synthetic Structural Data* (p. 2025.09.29.679384). bioRxiv. https://doi.org/10.1101/2025.09.29.679384

Winkler, D. G., Faia, K. L., DiNitto, J. P., Ali, J. A., White, K. F., Brophy, E. E., Pink, M. M., Proctor, J. L., Lussier, J., Martin, C. M., Hoyt, J. G., Tillotson, B., Murphy, E. L., Lim, A. R., Thomas, B. D., Macdougall, J. R., Ren, P., Liu, Y., Li, L.-S., … Kutok, J. L. (2013). PI3K-δ and PI3K-γ inhibition by IPI-145 abrogates immune responses and suppresses activity in autoimmune and inflammatory disease models. *Chemistry & Biology*, *20*(11), 1364–1374. https://doi.org/10.1016/j.chembiol.2013.09.017

Wood, E. R., Truesdale, A. T., McDonald, O. B., Yuan, D., Hassell, A., Dickerson, S. H., Ellis, B., Pennisi, C., Horne, E., Lackey, K., Alligood, K. J., Rusnak, D. W., Gilmer, T. M., & Shewchuk, L. (2004). A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib): Relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells.

Cancer Research, *64*(18), 6652–6659. https://doi.org/10.1158/0008-5472.CAN-04-1168

Wu, M.-H., Xie, Z., Ji, S., & Zhi, D. (2025). *Towards Precision Protein-Ligand Affinity Prediction Benchmark: A Complete and Modification-Aware DAVIS Dataset* (arXiv:2512.00708). arXiv. https://doi.org/10.48550/arXiv.2512.00708

Yang, Z., Zhong, W., Lv, Q., Dong, T., & Yu-Chian Chen, C. (2023). Geometric Interaction Graph Neural Network for Predicting Protein–Ligand Binding Affinities from 3D Structures (GIGN). *The Journal of Physical Chemistry Letters*, *14*(8), 2020–2033. https://doi.org/10.1021/acs.jpclett.2c03906