EVALUATING PHYSICAL INTUITION IN LARGE LANGUAGE MODELS FOR STRUCTURE-GUIDED MUTATION DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Structure-guided protein design requires interpreting molecular geometry and chemistry to anticipate mutation effects. Existing LLM evaluations emphasize sequence-based prediction or tool-integrated workflows, leaving unclear whether models themselves can reason directly from three-dimensional structures. Here, we systematically evaluate this capacity, termed physical intuition, through a partner-contrastive benchmark that holds the protein and substitution fixed across different binding partners. Models inspect local coordinates with or without Python, without dedicated mutation-effect predictors. Although overall accuracy remains modest, most evaluated LLMs recover more complete partner contrasts when affinity improvements are larger, and successful judgments also occur without Python. Structural interventions and evidence audits support selective dependence on mutation-relevant geometry and chemistry, with auditable observations largely grounded in the supplied coordinates. Case studies further show that models can identify relevant interactions and propose mutant-state hypotheses, but struggle to anticipate structural rearrangements and weigh their energetic consequences. Together, these experiments provide, to our knowledge, the first systematic evidence that general-purpose LLMs can use three-dimensional structural evidence to reason about partner-dependent mutation effects, without relying on specialized predictors to supply those judgments.

1 INTRODUCTION

Structure-guided protein design requires relating a proposed mutation to its molecular environment. Experts inspect geometry and chemistry to identify interactions a substitution may alter and assess whether its surroundings can accommodate the change (Kortemme et al., 2004; Kortemme & Baker, 2002). The same substitution can strengthen binding to one partner while weakening binding to another, because packing, electrostatics, hydrogen bonding and conformational accommodation differ across interfaces (Kortemme et al., 2004; Havranek & Harbury, 2003). This context dependence is central to specificity engineering: a useful judgment must connect structural evidence to partner-specific consequences, rather than treat a substitution as intrinsically beneficial or detrimental.

General-purpose large language models (LLMs) are increasingly evaluated in tool-integrated protein-engineering and biomedical workflows (Ghafarollahi & Buehler, 2024; Swanson et al., 2025; Huang et al., 2026; Tan et al., 2026a). These systems combine LLM planning and interpretation with specialized modelling and prediction. Their success does not isolate whether the LLM can infer mutation consequences from structural evidence. The relevant distinction is not tool use itself, but whether computation supplies evidence for the LLM to interpret or the target prediction on its behalf.

We ask whether, and under what conditions, an LLM can make partner-dependent mutation judgments without a dedicated predictor of the target effect. We use *physical intuition* to denote qualitative judgment based on structural observations and relevant interaction principles, without assuming human-like internal reasoning or exact energetic calculation. Overall accuracy alone may obscure where such judgments succeed. We therefore compare prediction success across experimentally measured affinity-improvement magnitudes, and separately examine dependence on mutation-relevant structural information.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our benchmark holds the protein and substitution fixed across partners, using SKEMPI 2.0 and literature-supported systems (Jankauskaitė et al., 2019). Each complex is presented independently, without revealing group membership; every partner-specific outcome must be correct for a group to succeed. In Dataset 1, two disjoint strata—larger and smaller affinity improvements—provide an internal comparison with the same opposite-effect label composition. Models inspect coordinates with or without access to Python and basic computational libraries, but cannot use dedicated energetic or mutation-effect predictors.

The experiments address three linked questions: where correct partner contrasts are recovered, what structural information judgments depend on, and how well explanations recover experimentally supported mechanisms. Three of four Python-assisted LLMs and both physical predictors resolve more groups in the larger-improvement stratum. LLM group accuracies fall in a similar range to those of the evaluated physical predictors, but with distinct precision–recall profiles. On the shared larger-improvement subset, GLM-5.3 Flash also matches the tool-integrated workflow Biomni while recovering partially different groups. Structural interventions reveal selective dependence on mutation-site and partner context, while rotations preserve aggregate performance. Claim audits and case analyses then distinguish agreement with observed coordinates from correct effect prediction and mechanistic understanding. Together, these experiments provide systematic evidence that general-purpose LLMs can themselves select and interpret three-dimensional structural evidence to reason about partner-dependent mutation effects, without relying on specialized predictors to supply those judgments.

Related work

Specialized protein models. Protein engineering uses physical energy functions for mutation-effect estimation (Schymkowitz et al., 2005; Barlow et al., 2018), sequence design conditioned on backbone or atomic context (Dauparas et al., 2022; Hsu et al., 2022; Dauparas et al., 2025), and variant-effect prediction from sequence and evolutionary information (Meier et al., 2021; Frazer et al., 2021; Notin et al., 2022). ProteinGym benchmarks fitness prediction across experimental assays (Notin et al., 2023). These methods supply specialized design or prediction capabilities.

Tool-integrated workflows. Agentic systems combine LLM planning with molecular modelling and biomedical tools (Ghafarollahi & Buehler, 2024; Swanson et al., 2025; Shen et al., 2024; Huang et al., 2026); related systems automate machine-learning workflows and expose protein-analysis tools (Liu et al., 2025; Xu et al., 2026; Teneggi et al., 2026). Rank-and-Reason (VenusRAR) uses specialist-model rankings and computed biophysical context (Tan et al., 2026a). Such workflows evaluate the combined system, not the LLM's ability to infer effects without a dedicated predictor.

LLM-based protein-engineering benchmarks. VenusAgentEval, BioDesignBench and ABLE evaluate tool-enabled engineering through planning, tool orchestration, candidate generation and validation (Tan et al., 2026b; Kim & Romero, 2026; Cai et al., 2026). PG-LLM evaluates tool-free variant ranking from sequences and assay descriptions without structural input (Arora et al., 2026). Our study instead uses fixed-substitution partner contrasts, effect-magnitude stratification, structural interventions and evidence audits to examine context-dependent structural judgment.

2 A benchmark for partner-dependent mutation effects

The benchmark contains 60 non-overlapping mutation groups and 126 mutation–partner observations (Fig. 1). Each group holds the focal protein, position and substitution fixed across partners. Fifty-six groups contain opposite affinity effects; four contain only affinity-enhancing observations. Dataset 1 derives from SKEMPI 2.0, and Dataset 2 adds literature-supported systems. Appendix A details curation and processing.

2.1 Benchmark design and curation

Dataset 1 retains single substitutions at binding interfaces with measured wild-type (WT) and mutant dissociation constants from SKEMPI 2.0 (Jankauskaitė et al., 2019). Effects are defined as $r = \log_{10}(K_D^{\mathrm{mut}}/K_D^{\mathrm{WT}})$, with negative values indicating enhancement. Repeated measurements are represented by their median log ratio. Selecting exactly two partners with opposite nonzero effects,

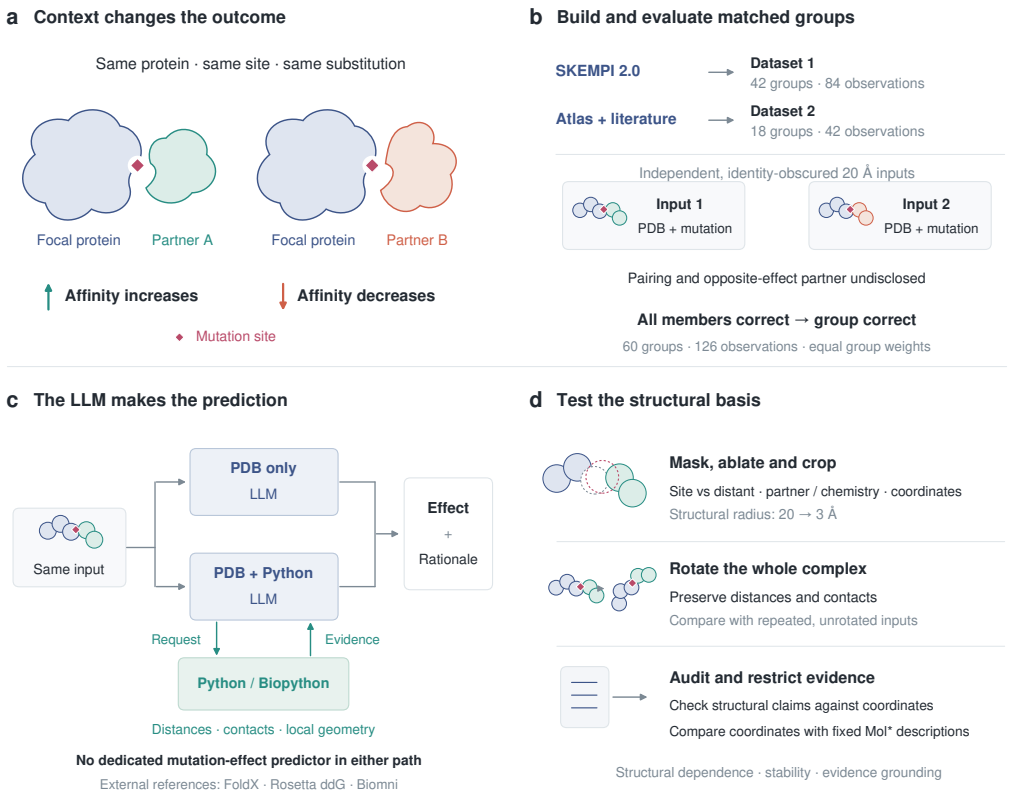


Figure 1: **Evaluating partner-dependent structural judgment.** a, A fixed substitution has opposite affinity effects across partners. b, Two curation routes yield 60 groups and 126 observations: 56 opposite-effect groups and four all-enhancing groups. Complexes are evaluated independently, with group membership undisclosed. c, Models inspect coordinates with or without Python; dedicated mutation-effect predictors are excluded. Each condition independently yields a prediction and rationale. d, Structural interventions, rotations, claim audits and fixed descriptions probe the basis of the judgments. Panels depict experimental design, not measured performance.

without an additional magnitude threshold, yielded 42 groups and 84 observations across 17 WT sequences. All mutation sites lie within 5 Å of partner-protein heavy atoms.

Dataset 2 was assembled through a PDB-wide search (Berman et al., 2000) and manual verification of structures and literature-supported effects. It draws primarily on experimental data published after 2019. After excluding Dataset 1 overlap, it contains 18 groups and 42 observations: 14 opposite-effect groups and four all-enhancing groups. The latter assess recovery of beneficial mutations across partners, rather than effect reversal. Their inclusion was motivated by observed negative-prediction tendencies and the affinity-improvement objective. Dataset 2 includes groups with more than two observations and one protein–small-molecule group; it does not estimate the natural prevalence of beneficial mutations.

2.2 Structural inputs and inference conditions

Each complex was presented independently as a mutation specification and coordinates within 20 Å of the mutation site. Molecular identifiers and group membership were withheld; renumbering and record permutation reduced direct identity cues without guaranteeing that identity was unrecoverable (Appendix A.5). An identity-recovery check with DeepSeek V4.1 Flash reported 16/84 successes for processed PDBs versus 76/84 for reference inputs (Appendix A.6, including reporting qualifications).

DeepSeek V4.1 Flash is abbreviated as DeepSeek Flash. Models received identical coordinate records in two conditions: PDB only, without external tools or code execution, and PDB + Python, allowing

basic libraries such as Biopython (Cock et al., 2009) for structural inspection. Dedicated energetic and mutation-effect predictors were excluded; the LLM selected and interpreted evidence to predict the effect direction. Model API identifiers and inference settings are listed in Appendix C.

2.3 Evaluation and effect-magnitude comparison

The primary endpoint is group accuracy, with equal weight per group and every member required to be correct. LLM method comparisons use majority-vote consensus from three runs. Observation-level precision and recall describe enhancement predictions; task accuracy provides an additional descriptive measure.

For each Dataset 1 pair, its enhancing member defines $F = K_D^{\mathrm{WT}}/K_D^{\mathrm{mut}} > 1$. We compare 21 larger-improvement groups ($F \geq 2$) with 21 smaller-improvement groups ($1 < F < 2$). Both strata contain one enhancing and one weakening observation per group.

FoldX, Rosetta ddG and Biomni serve as external comparisons (Schymkowitz et al., 2005; Barlow et al., 2018; Huang et al., 2026). Physical predictors use full structures and exclude the Dataset 2 protein–small-molecule group, leaving 17 groups and 40 observations.

3 Where partner-dependent predictions succeed

3.1 Prediction success is concentrated in larger-improvement contrasts

Overall group accuracy pools contrasts with different effect magnitudes. The two disjoint Dataset 1 strata instead compare larger and smaller improvements under the same requirement: both partner-specific directions must be correct. Three of four Python-assisted LLMs and both physical predictors achieved higher group accuracy in the larger-improvement stratum (Fig. 2; Table 1). GLM-5.3 Flash resolved 7 of 21 groups (33.33%), compared with 1 of 21 (4.76%) in the smaller-improvement stratum. The corresponding counts were six versus zero for GLM-5.3 and five versus one for DeepSeek Flash. FoldX and Rosetta ddG resolved seven versus zero and eight versus two groups, respectively. Qwen3.8 Flash was the exception, resolving two groups in each stratum.

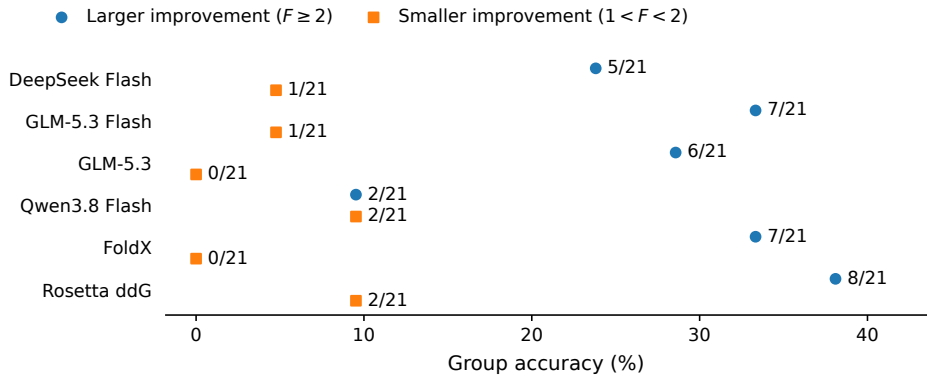


Figure 2: **Complete partner contrasts are recovered more often in the larger-improvement stratum.** Each point shows group accuracy and is annotated with correct groups out of 21. The two strata partition Dataset 1 and each retain one enhancing and one weakening observation per group. LLMs use PDB + Python majority-vote consensus from three runs. Qwen3.8 Flash shows no difference.

These differences concern complete opposite-effect pairs, not just identification of the enhancing member. The smaller-improvement stratum documents substantially lower accuracy for most evaluated methods and delimits where their observed successes extend. The higher accuracy for larger improvements across most evaluated LLMs and both physical predictors is consistent with the expectation that more pronounced mutation effects are easier to predict than subtler changes.

Table 1: Group accuracy on the full datasets and disjoint affinity-improvement strata.

Method	Dataset 1			Dataset 2
	All 42 groups	Larger improvement $F \geq 2$: 21 groups	Smaller improvement $1 < F < 2$: 21 groups	All 18 groups
LLMs with Python				
DeepSeek Flash	14.29 (6/42)	23.81 (5/21)	4.76 (1/21)	11.11 (2/18)
GLM-5.3 Flash	19.05 (8/42)	33.33 (7/21)	4.76 (1/21)	16.67 (3/18)
GLM-5.3	14.29 (6/42)	28.57 (6/21)	0.00 (0/21)	11.11 (2/18)
Qwen3.8 Flash	9.52 (4/42)	9.52 (2/21)	9.52 (2/21)	0.00 (0/18)
Physical predictors				
FoldX	16.67 (7/42)	33.33 (7/21)	0.00 (0/21)	11.76 (2/17)[†]
Rosetta ddG	23.81 (10/42)	38.10 (8/21)	9.52 (2/21)	5.88 (1/17)[†]
Agentic workflow				
Biomni	—	33.33 (7/21)	—	5.56 (1/18)

Values are percentages (correct/total groups); every member must be correct. LLMs use PDB + Python consensus. F is the enhancing member's affinity improvement; no extra threshold applies to the weakening member. The strata partition Dataset 1 and reuse its predictions. Dashes indicate unavailable results.
[†] Physical predictors exclude the Dataset 2 protein–small-molecule group (17 groups, 40 observations). FoldX subset results were reported as FoldX5.



Figure 3: **Prediction patterns and the contribution of Python.** a, Enhancement precision and recall pooled across Datasets 1 and 2: 126 observations for LLMs with Python (circles), and 124 for physical predictors (diamonds), excluding one protein–small-molecule group. The nested $\geq 2\times$ subset is not counted again. b, Consensus group accuracy with and without Python on Dataset 1, its nested larger-improvement subset, and Dataset 2. The direct comparison between disjoint effect strata is shown separately in Fig. 2.

3.2 Similar accuracy to physical predictors and Biomni, with distinct prediction patterns

LLM group accuracies are close to those of the evaluated physical predictors, and GLM-5.3 Flash also matches the tool-integrated workflow Biomni on the shared larger-improvement subset (Table 1). On Dataset 1, GLM-5.3 Flash resolved eight groups, compared with seven for FoldX and ten for Rosetta ddG. LLMs and physical predictors nevertheless show distinct precision–recall profiles (Fig. 3a): several LLMs make fewer but more precise enhancement predictions. DeepSeek Flash achieved 91.67% precision at 16.18% recall across both datasets, whereas the physical predictors recovered more enhancing observations.

On the matched 21-group larger-improvement subset, GLM-5.3 Flash and Biomni each resolved seven groups (33.33%). Five correct groups were shared, while each method resolved two missed by the other. Thus, direct LLM judgment matched Biomni's group accuracy on this subset while recovering partially different cases. Overall accuracies remain modest, but the comparisons reveal distinct prediction patterns at similar performance levels.

3.3 Successful partner judgments also occur without Python

On Dataset 1, GLM-5.3 Flash resolved the same number of groups with and without Python; DeepSeek Flash and Qwen3.8 Flash each resolved two fewer without it (Fig. 3b). Python provided additional gains on the larger-improvement subset, but effects on Dataset 2 varied. Without Python, DeepSeek Flash and Qwen3.8 Flash succeeded only on the same all-enhancing group, whereas GLM-5.3 Flash resolved two opposite-effect groups. Thus, code execution is not necessary for every successful partner-dependent judgment, but the results do not establish equivalence between tool conditions. Appendix A.4 separates Dataset 2 successes by group composition.

3.4 Successful judgments span multiple protein systems

With Python, DeepSeek Flash, GLM-5.3 Flash and Qwen3.8 Flash resolved Dataset 1 groups across four, five and four WT protein systems, respectively, collectively covering seven of the 17 systems (Appendix D, Fig. S5). Successes also span multiple systems without Python. Coverage is uneven: BPTI accounts for half of the successes of GLM-5.3 Flash with Python and of Rosetta ddG. These observations establish within-benchmark breadth, not generalization to unseen systems.

4 Structural evidence underlying mutation judgments

The effect-magnitude comparison identifies where correct partner contrasts are more often recovered, but not which structural features support the judgments. We therefore test selective dependence on mutation-relevant geometry and partner chemistry. These interventions complement the stratified results; they do not directly measure whether larger-improvement groups have more salient structural features.

4.1 Judgments depend selectively on local structural context

Mutation-site masking changed predictions more often than distant side-chain masking. With Python, a mean 42.06% of Dataset 1 groups changed across DeepSeek Flash, GLM-5.3 Flash and Qwen3.8 Flash after mutation-site masking, versus 10.32% after distant masking (Fig. 4a). A group changed when at least one member differed from its intact-input consensus. All three models showed the larger response to mutation-site masking; without Python, such masking changed a mean 19.05% of groups. This selective sensitivity concerns information dependence, not accuracy loss: removing relevant evidence can reasonably change a judgment.

Removing the partner chain lowered mean group accuracy across these three models from 13.22% to 7.78%, with decreases also observed in the larger-improvement subset (Fig. 4b). In a separate experiment, retaining the partner backbone while masking its side chains and residue identities reduced the mean from 14.55% to 7.14%. Each ablation uses its corresponding intact-input control. These patterns support a contribution of partner information to the judgments.

Removing all atomic coordinates while retaining residue identities and other non-geometric information reduced mean group accuracy from 14.55% to 1.59%. Mean enhancement recall fell from 17.20% to 2.38%, while specificity remained high. Geometry removal thus chiefly reduced recovery of enhancements, leaving negative predictions dominant; it also changed prediction tendencies.

For DeepSeek Flash, we varied crop radius from 20 Å to 15, 10, 5 and 3 Å, with three runs per condition (Fig. 4c). Full-dataset mean group accuracy varied non-monotonically between 10.32% and 13.49% over 5–20 Å, but fell to 3.97% at 3 Å, versus 12.70% at 20 Å. Corresponding larger-improvement-subset values were 1.59% and 20.63%. The tested 5–20 Å neighbourhoods retained much of the observed performance, whereas 3 Å did not.

4.2 Tested coordinate rotations preserve aggregate performance

We applied three independent joint rigid-body rotations to each complex, preserving relative geometry, and compared predictions with three repetitions on unchanged coordinates. Across the three models, 88.49% of tasks received identical predictions within the rotated condition, versus 84.92% under unrotated repetition; consistency decreased for none (Fig. 4d). Mean group accuracy was 15.08%

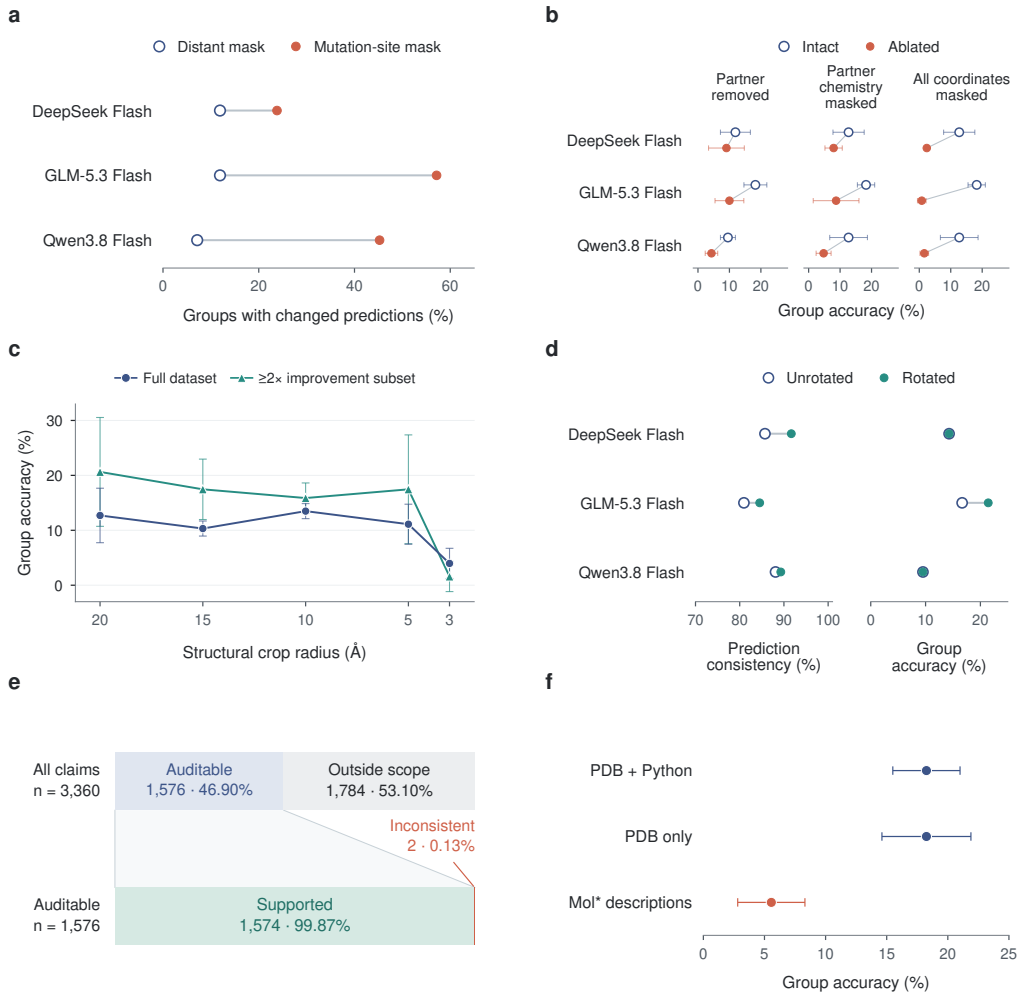


Figure 4: **Structural dependence and use of molecular evidence.** All results use Dataset 1. a, Groups with changed consensus predictions after distant or mutation-site masking, relative to intact input, with Python. b, Group accuracy after partner removal, partner-chemistry masking with backbone retained, or coordinate removal; each has its own intact control. c, DeepSeek Flash accuracy across crop radii on the full dataset and nested $\geq 2\times$ subset. d, Consistency and accuracy under joint rigid-body rotations versus unchanged-coordinate repetitions; consistency requires identical task predictions in all three runs of a condition. e, Claim audit; the lower bar expands auditable claims. Other claims remain unassessed. f, GLM-5.3 Flash with PDB plus Python, PDB alone or fixed Mol* descriptions. Conditions use three runs, except mutation-site masking and partner-removed inputs for DeepSeek, GLM and Qwen3.8 (five); all partner-removal controls use three. Points and error bars in b, c and f are means $\pm$ sample standard deviations. Comparisons are descriptive.

after rotation and 13.49% under repetition; larger-improvement-subset values were 23.81% and 17.46%. The tested rotations did not reduce within-condition consistency or aggregate accuracy. This is compatible with use of relative geometry, not a proof of exact rotational invariance.

4.3 Auditable structural claims largely agree with coordinates

We audited 3,360 structural claims from 1,344 Dataset 1 predictions, including contacts, distances and residue-level spatial relationships (Fig. 4e). Mol* structural analysis supported 1,574 claims and found two inconsistent with the coordinates. The other 1,784 concerned properties outside the predefined scope, including solvent accessibility and qualitative pocket geometry. Thus, 46.90%

of claims were assessable, and 99.87% of that subset were supported. This conditional agreement supports the grounding of checkable observations, not the validity of unassessed claims or the causal role of a cited interaction in an affinity change. The audit procedure is detailed in Appendix E.

4.4 Fixed descriptions incompletely substitute for coordinate access

We replaced coordinates with Mol*-derived descriptions of predefined contacts, distances and local-environment features for the same neighbourhoods (Sehnal et al., 2021). GLM-5.3 Flash achieved 5.56% $\pm$ 2.75% group accuracy with descriptions, versus 18.25% $\pm$ 3.64% with PDB alone and 18.25% $\pm$ 2.75% with PDB plus Python (three-run means $\pm$ sample standard deviations; Fig. 4f). The tested descriptions did not reproduce coordinate-access performance. This is consistent with a benefit from context-specific evidence selection, but differences in coverage and presentation also remain possible explanations.

These experiments test structural dependence and agreement with supplied coordinates. Whether an explanation recovers the experimentally supported mutation mechanism requires separate evidence, which we examine next.

5 Mechanistic insights from case studies

These cases reveal a progressive picture of LLM-based mutation-effect reasoning, from recognizing partner-dependent determinants to evaluating mutant-state mechanisms and mechanistic validity.

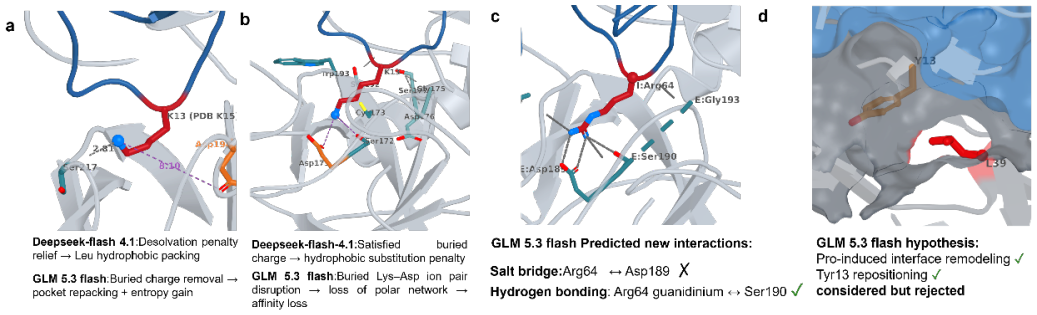


Figure 5: Mechanistic analysis of LLM mutation-effect reasoning. (a,b) DeepSeek Flash and GLM-5.3 Flash capture partner-dependent mutation effects by considering distinct interaction environments (PDB 3T62 and 3M7Q). (c) GLM-5.3 Flash proposes mutation-induced interaction gains, with partially validated and unsupported contacts (PDB 7JOD). (d) GLM-5.3 Flash generates a mutant-state hypothesis consistent with the literature but rejects it due to WT-state structural anchoring (PDB 7SCU).

Partner-dependent interpretation.

We first examined whether LLMs can recognize that mutation effects are determined by the local binding environment rather than by the mutation itself. The ShPI-1 K13L case provides a direct test because the same substitution produces opposite effects against different protease partners (García-Fernández et al., 2015). GLM-5.3 Flash correctly predicted that K13L improves α-chymotrypsin inhibition but weakens trypsin inhibition. For α-chymotrypsin, the model identified the unfavorable energetic consequence of a buried lysine lacking sufficient electrostatic compensation and proposed that leucine could provide more favorable hydrophobic packing (Fig. 5a). For trypsin, GLM-5.3 Flash similarly identified K13 as a deeply buried interface anchor stabilized by electrostatic interactions and hydrogen-bonding networks; therefore, K13L would disrupt this polar stabilization network (Fig. 5b). Thus, LLMs can capture partner-dependent structural determinants when the energetic consequences are strongly encoded by the local interaction environment.

Mutant-state interaction prediction.

We next examined whether LLMs can infer mutation-induced interaction changes from WT structures. The BbKI F64R cases were analyzed in two distinct complexes involving bovine trypsin and human

kallikrein 4 (KLK4) (Li et al., 2021). The results show that LLMs can propose relevant structural changes after mutation, including the formation of Arg64-mediated interactions, water displacement, and local interaction-network remodeling. Although several predicted Arg64-mediated contacts were partially supported by experimentally determined mutant structures, some specific interactions were overestimated (Fig. 5c). Notably, DeepSeek Flash and GLM-5.3 Flash proposed similar mutant-state hypotheses, but reached different final predictions because they assigned different weights to the energetic contributions inferred from structural geometry. These results indicate that LLMs can generate plausible mutant-state hypotheses, but remain limited in translating local interaction changes into overall binding free-energy effects. More broadly, uncertainty arises from two major challenges: extrapolating unobserved mutant conformations from known WT structures and converting structural changes into energetic consequences for decision-making.

Difficulty in evaluating indirect structural effects.

The EVA-A Y44A and L39P cases (Devkota et al., 2023) revealed another limitation: in these cases, affinity changes were mediated by indirect structural effects rather than by direct interactions formed by the mutated residues, and therefore the evaluated LLMs failed to correctly predict these cases (Fig. 5d). Detailed analysis of the reasoning trajectories showed that the major difficulty was not the inability to propose plausible mechanisms or mechanisms consistent with the literature, but rather an asymmetric weighting between WT structural evidence and mutant-state hypotheses. Notably, these models did generate hypotheses consistent with the experimentally observed mechanisms, including ligand relaxation and interface repositioning. However, because the models could only observe static experimentally determined structures, these hypotheses were rapidly self-rejected during the reasoning process due to the dominance of static structural evidence and WT-based interactions.

Correct prediction with incorrect reasoning.

Finally, the BLIP E73W case demonstrates that correct prediction of the affinity-effect direction does not necessarily indicate mechanistic understanding (Rivera et al., 2025). Although GLM-5.3 Flash correctly predicted the affinity enhancement and identified several valid wild-type (WT) structural features, the proposed mechanistic explanation was not consistent with the mechanism described in the literature. The model attributed the affinity improvement mainly to the removal of the unfavorable energetic penalty associated with the buried Glu residue, water displacement, and aromatic repacking. However, the mutant structure revealed a distinct interaction mode involving packing between Trp73 and Tyr105 and stabilization of an alternative partner conformation. This case shares a similar failure mode with the EVA-A cases (Y44A and L39P). When LLMs lack relevant prior knowledge or the ability to reason beyond static geometric information, they face challenges in more complex mutation-effect interpretation problems.

Together, these cases suggest that LLMs possess the ability to extract chemically meaningful structural determinants from observed protein geometries and generate experimentally testable mutation hypotheses. However, when performing mutation-effect reasoning from WT structures, their major limitation lies in extrapolating beyond directly observed structural information, including mutant-state adaptation, conformational changes, and the energetic consequences of local interaction remodeling.

6 Conclusion

We provide systematic evidence that general-purpose LLMs can use structural evidence for partner-dependent mutation reasoning without dedicated predictors. Success is concentrated in larger-improvement contrasts; performance is close to that of physical predictors and, on the shared subset, Biomni, but prediction patterns differ. Structural interventions and audits support the use of relevant evidence, while mutant-state rearrangements and energetic interpretation remain key challenges.

These findings distinguish coordinate-grounded observations from validated mutation mechanisms: a correct affinity-direction prediction does not establish that its explanation captures the underlying energetic change. For structure-guided mutation design, the results suggest a role for LLMs in formulating partner-specific hypotheses for subsequent validation. Progress will require better evaluation of alternative mutant conformations and their energetic consequences, alongside broader tests across protein systems. Such evaluations should jointly assess predictive accuracy, structural grounding and mechanistic validity.

REFERENCES

Rohit Arora, Leo Tianlai Chen, Melissa Du, Debora S. Marks, and George M. Church. PG-LLM: Benchmarking General-Purpose Language Models for Protein Variant Ranking. *bioRxiv*, 2026. doi: 10.64898/2026.07.27.741045. Preprint.

Kyle A. Barlow, Shane Ó Conchúir, Samuel Thompson, Pooja Suresh, James E. Lucas, Markus Heinonen, and Tanja Kortemme. Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein–Protein Binding Affinity upon Mutation. *The Journal of Physical Chemistry B*, 122(21): 5389–5399, 2018. doi: 10.1021/acs.jpcb.7b11367.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235.

Bryce Cai, Geetha Jeyapragasan, Samira Nedungadi, Jake Yukich, and Seth Donoughe. Agentic BAIM-LLM Evaluation (ABLE): Benchmarking LLM Use of Protein Design Tools. *arXiv preprint arXiv:2609.05818*, 2026. doi: 10.48550/arXiv.2609.05818.

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009. doi: 10.1093/bioinformatics/btp163.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187.

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

Shankar Raj Devkota, Pramod Aryal, Rina Pokhrel, Wanting Jiao, Andrew Perry, Santosh Panjikar, Richard J. Payne, Matthew C. J. Wilce, Ram Prasad Bhusal, and Martin J. Stone. Engineering broad-spectrum inhibitors of inflammatory chemokines from subclass A3 tick evasins. *Nature Communications*, 14(1):4204, 2023. doi: 10.1038/s41467-023-39879-3.

Jason E. Donald, Daniel W. Kulp, and William F. DeGrado. Salt bridges: Geometrically specific, designable interactions. *Proteins: Structure, Function, and Bioinformatics*, 79(3):898–915, 2011. doi: 10.1002/prot.22927.

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K. Min, Kelly Brock, Yarin Gal, and Debora S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021. doi: 10.1038/s41586-021-04043-8.

Rossana García-Fernández, Markus Perbandt, Dirk Rehders, Patrick Ziegelmüller, Nicolas Piganeau, Ulrich Hahn, Christian Betzel, María de los Ángeles Chávez, and Lars Redecke. Three-dimensional Structure of a Kunitz-type Inhibitor in Complex with an Elastase-like Enzyme. *Journal of Biological Chemistry*, 290(22):14154–14165, 2015. doi: 10.1074/jbc.m115.647586.

Alireza Ghafarollahi and Markus J. Buehler. ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 3(7): 1389–1409, 2024. doi: 10.1039/d4dd00013g.

James J. Havranek and Pehr B. Harbury. Automated design of specificity in molecular recognition. *Nature Structural Biology*, 10(1):45–52, 2003. doi: 10.1038/nsb877.

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

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Ryan Li, Yusuf Roohani, Lin Qiu, Shiyi Cao, Gavin Li, Junze Zhang, Di Yin, Rick Wierenga, Deniz Kavi, Sherry Liu, Tianwei She, Shruti Marwaha, Jennefer N. Carter, Xin Zhou, Matthew T. Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael P. Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Autonomous biomedical research with an artificial intelligence agent. *Science*, 393 (6813):eadz4351, 2026. doi: 10.1126/science.adz4351.

Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019. doi: 10.1093/bioinformatics/bty635.

Jeonghyeon Kim and Philip Romero. Benchmarking and behavioral characterization of LLM agents for protein design. *bioRxiv*, 2026. doi: 10.64898/2026.05.06.723381. Preprint.

Tanja Kortemme and David Baker. A simple physical model for binding energy hot spots in protein–protein complexes. *Proceedings of the National Academy of Sciences*, 99(22):14116–14121, 2002. doi: 10.1073/pnas.202485799.

Tanja Kortemme, Lukasz A Joachimiak, Alex N Bullock, Aaron D Schuler, Barry L Stoddard, and David Baker. Computational redesign of protein-protein interaction specificity. *Nature Structural & Molecular Biology*, 11(4):371–379, 2004. doi: 10.1038/nsmb749.

Mi Li, Jaroslav Srp, Michael Mareš, Alexander Wlodawer, and Alla Gustchina. Structural studies of complexes of kallikrein 4 with wild-type and mutated forms of the Kunitz-type inhibitor BbKI. *Acta Crystallographica Section D Structural Biology*, 77(8):1084–1098, 2021. doi: 10.1107/s2059798321006483.

Yungeng Liu, Zan Chen, Yuguang Wang, and Yiqing Shen. AutoProteinEngine: A Large Language Model Driven Agent Framework for Multimodal AutoML in Protein Engineering. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 422–430. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-industry.36/.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29287–29303. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f51338d736f95dd42427296047067694-Abstract.html.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N. Gomez, Debora Marks, and Yarin Gal. Tranception: Protein Fitness Prediction with Autoregressive Transformers and Inference-time Retrieval. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 16990–17017. PMLR, 2022. URL https://proceedings.mlr.press/v162/notin22a.html.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379. Curran Associates, Inc., 2023. doi: 10.52202/075280-2810.

Paola Rivera, Shuo Lu, Dignite Ngango, Banumathi Sankaran, Bidadi Venkataram Prasad, and Timothy Palzkill. A β-lactamase inhibitory protein mutant displays high potency and a broad inhibition profile due to an altered binding mode with β-lactamases. *Journal of Biological Chemistry*, 301(12):110850, 2025. doi: 10.1016/j.jbc.2025.110850.

J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano. The FoldX web server: an online force field. *Nucleic Acids Research*, 33(Web Server):W382–W388, 2005. doi: 10.1093/nar/gki387.

David Sehnal, Sebastian Bittrich, Mandar Deshpande, Radka Svobodová, Karel Berka, Václav Bazgier, Sameer Velankar, Stephen K Burley, Jaroslav Koča, and Alexander S Rose. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, 49(W1):W431–W437, 2021. doi: 10.1093/nar/gkab314.

Yiqing Shen, Zan Chen, Michail Mamalakis, Yungeng Liu, Tianbin Li, Yanzhou Su, Junjun He, Pietro Liò, and Yu Guang Wang. TourSynbio: A Multi-Modal Large Model and Agent Framework to Bridge Text and Protein Sequences for Protein Engineering. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2382–2389. IEEE, 2024. doi: 10.1109/bibm62325.2024.10822695.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723, 2025. doi: 10.1038/s41586-025-09442-9.

Yang Tan, Yuyuan Xi, Can Wu, Bozitao Zhong, Mingchen Li, Guisheng Fan, Jiankang Zhu, Yafeng Liang, Nanqing Dong, and Liang Hong. Rank-and-Reason: Multi-Agent Collaboration Accelerates Zero-Shot Protein Mutation Prediction. *arXiv preprint arXiv:2602.00197*, 2026a. doi: 10.48550/arXiv.2602.00197.

Yang Tan, Lingrong Zhang, Mingchen Li, Yuanxi Yu, Bozitao Zhong, Bingxin Zhou, Nanqing Dong, and Liang Hong. Self-evolving AI agents for protein discovery and directed evolution. *arXiv preprint arXiv:2603.27303*, 2026b. doi: 10.48550/arXiv.2603.27303.

Jacopo Teneggi, S. M. Bargeen A. Turzo, Tanya Marwah, Alberto Bietti, P. Douglas Renfrew, Vikram Khipple Mulligan, and Siavash Golkar. Protein Design with Agent Rosetta: A Case Study for Specialized Scientific Agents. *arXiv preprint arXiv:2603.15952*, 2026. doi: 10.48550/arXiv.2603.15952.

Xiaopeng Xu, Chenjie Feng, Chao Zha, Wenjia He, Maolin He, Bin Xiao, and Xin Gao. ProteinMCP: An agentic AI framework for autonomous protein engineering. *Protein Science*, 35(4):e70547, 2026. doi: 10.1002/pro.70547.

AI USE STATEMENT

We used generative AI tools to assist with translation, language editing, literature search, and software coding, including implementation support. AI-assisted translation and editing were used to improve the clarity, grammar, and readability of the manuscript. AI tools were also used to help identify relevant literature and to assist with writing and editing code. All AI-assisted text was reviewed and revised by the authors, relevant literature was independently checked against the original sources, and AI-assisted code was reviewed and tested for correctness.

We did not use generative AI tools to generate synthetic data, formulate or refine research hypotheses, design the research methodology or experiments, develop theoretical or conceptual claims, formulate or prove mathematical claims, perform qualitative or thematic data analysis, or interpret experimental results. The authors take full responsibility for the final content of the paper, including all text, citations, code, claims, and results produced with the assistance of generative AI.

A Additional benchmark curation and input-processing details

A.1 Dataset 1: filtering, identity mapping and aggregation

Initial deduplication reduced 7,085 SKEMPI records to 6,794 measurement records before quality filtering and aggregation of repeated experimental measurements. We required finite, positive WT and mutant dissociation constants and excluded censored affinities. Original and cleaned single-substitution annotations had to agree, and the mutated chain had to resolve unambiguously to one side of the complex and a unique full-length WT sequence. WT residue identity was checked before and after sequence mapping. Mutation groups were defined by the exact WT sequence and mapped substitution.

Partner identity was the sorted, deduplicated set of complete partner-chain sequences, independent of protein names or PDB numbering. Full-length UniProt WT sequences were used when all partner chains mapped unambiguously; otherwise, the entire partner side was represented by complete PDB polymer sequences. This definition does not distinguish repeated copies of an identical chain. Repeated measurements for the same complex, WT sequence, substitution and partner were consolidated using the median of individual log ratios, rather than the log ratio of separately aggregated WT and mutant affinities. Experimental conditions and publications were retained for auditing but did not partition this aggregation.

Interface filtering used the first PDB model and the highest-occupancy alternate location for each atom name. The mutated WT residue had to lie within 5 Å of a non-water heavy atom on a partner chain. Groups with more than two eligible partner identities were excluded from Dataset 1 rather than expanded into pairwise comparisons. If multiple eligible complex records represented a partner, the record supported by the most original measurements was selected; ties were broken lexicographically by complex identifier. Selected partner effects had to have strictly opposite signs; zero effects were excluded without imposing an additional effect-magnitude threshold. The resulting 42 groups spanned 17 WT sequences and 32 PDB entries. All 84 final observations passed a second 5 Å interface check using only partner-protein heavy atoms and were supported by 122 unique source measurements.

Table S1: SKEMPI curation stages.

Stage	Retained records or groups
Raw SKEMPI records	7,085
After initial deduplication	6,794
After record-level quality filtering	4,025
After repeated-measurement aggregation	3,396 evidence records
After interface filtering	2,445 evidence records
Groups with exactly two partner identities	259 groups
Selected groups with opposite effects	42 groups / 84 observations

A.2 Effect-magnitude stratification

The enhancing member of each Dataset 1 pair defines $F = K_D^{\mathrm{WT}}/K_D^{\mathrm{mut}} > 1$. The larger-improvement stratum contains 21 groups with $F \geq 2$; the smaller-improvement stratum contains 21 with $1 < F < 2$. Both retain one enhancing and one weakening observation per group. The smaller-improvement stratum is an internal comparison for subtler experimental effects, not a discarded set. No additional magnitude threshold applies to the weakening member, so neither stratum requires both members to have effects of a specified magnitude.

The comparison is descriptive and uses two bins within the same benchmark. It does not establish a continuous effect–accuracy relationship, matching of protein systems or mutation types between strata, or independently measured differences in structural salience. No statistical-significance or prediction-bias-adjustment claim is inferred from the stratum counts alone.

A.3 DATASET 2: CANDIDATE GENERATION AND LITERATURE CURATION

The PDB-wide atlas began with 259,129 identifiers and linked focal proteins to distinct partners using biological assemblies, sequence mappings and direct-contact evidence. PDBe biological-assembly and molecular-entity annotations were combined with UniProt chain mappings. Co-assembly associations required direct-contact support from PDBe PISA or ligand binding-site annotations. Focal proteins required confirmed non-self contacts in at least two distinct PDB entries with at least two distinct raw partner identities. This generated 33,321 structural candidates without imposing partner-class, resolution or minimum interface-size thresholds at this stage.

Candidates were manually assessed for a fixed protein and substitution, identifiable partner complexes and literature-supported affinity effects. Structural contact alone was insufficient for inclusion. Despite extensive structural and literature searches, experimentally characterized multi-partner mutation groups were scarce. Fourteen groups contained opposite-sign effects. We additionally included four groups in which all measured effects were affinity-enhancing, motivated by the models' observed tendency towards negative predictions and the practical goal of identifying beneficial mutations. These groups assess recovery of improvements across partners rather than opposite-effect discrimination. Groups containing only affinity-weakening observations were not included because they did not address this affinity-improvement objective. This selection reflects the evaluation objective, not the natural prevalence of beneficial mutations.

The 14 opposite-effect groups comprised ten groups with one enhancing and one weakening observation, two with two enhancing and one weakening observation, and two with two of each. Each of the four all-enhancing groups contained two observations. The final 18 groups comprised 42 observations (26 enhancing and 16 weakening), did not overlap Dataset 1, and included groups with more than two observations and one protein–small-molecule group. Excluding the latter left 17 groups and 40 observations (25 enhancing and 15 weakening) for physical-predictor comparisons. All groups were scored using the same requirement that every member be predicted correctly.

A.4 DATASET 2 SUCCESSES BY GROUP COMPOSITION

The four all-enhancing groups each contain two positive observations. Only `group_00401fc010988ad75ce4` was correctly resolved, by DeepSeek Flash and Qwen3.8 Flash with PDB-only input. No other evaluated method correctly resolved this group, and no method correctly resolved any of the other three all-enhancing groups. Correctly predicting only one member does not count as a correct group.

Accordingly, all correct Dataset 2 groups for the Python-assisted LLMs came from the 14 opposite-effect groups: DeepSeek Flash resolved two, GLM-5.3 Flash three, GLM-5.3 two, and Qwen3.8 Flash none. Biomni resolved one opposite-effect group. FoldX and Rosetta ddG resolved two and one, respectively, among the 13 opposite-effect groups available to physical predictors after excluding the protein–small-molecule group. Without Python, DeepSeek Flash and Qwen3.8 Flash each resolved one all-enhancing group and no opposite-effect groups, whereas GLM-5.3 Flash resolved two opposite-effect groups and no all-enhancing groups. These counts partition the existing predictions; no additional model runs or changes to the reported overall accuracies are involved.

A.5 STRUCTURAL INPUT PROCESSING

Inputs contained the mutation specification and a local structural neighbourhood extending 20 Å from the mutation site. Residue and atom identities were retained, while metadata, original identifiers and source annotations were withheld. Residue renumbering and record permutation reduced straightforward reconstruction of molecular identity without guaranteeing that identity was unrecoverable. Each complex was presented independently, without disclosing its mutation group or the effects measured for other partners. The same processed coordinate records were used with and without Python; Python and basic structural libraries could inspect distances, contacts and other structural relationships but did not supply dedicated energetic or mutation-effect predictions.

A.6 PROTEIN-IDENTITY RECOVERY UNDER INPUT PREPROCESSING

Task and inputs. This separate task asks the model to identify each protein chain's family and specific protein name, not to predict mutation effects. For the processed-PDB condition, each attempt receives only /input/complex.pdb, /input/task.json and /input/instruction.txt. The task JSON contains only a task identifier and `task_type=protein_identity`; mutation fields, source names and answer labels are withheld. The loader does not further rewrite the supplied PDB bytes. It identifies protein chains using amino-acid Cα records; non-protein components may provide context but are not identity targets. Reference labels, source-identity mappings and aliases remain on the host scoring side.

Model and analysis budget. The protocol specifies `deepseek-v4.1-flash` through an OpenAI-compatible interface, temperature 0.0, thinking enabled and `reasoning_effort=max`. Up to ten tool-assisted analysis rounds are allowed, followed by one tool-disabled JSON submission if no valid answer has been submitted. A valid earlier submission terminates the attempt. These are analysis rounds, not ten independent repetitions. The documented command uses the local Python runtime with eight workers and separate task messages and working directories.

Permitted analysis and scoring. The model may list and read files and run deterministic Python analyses of coordinates, contacts, composition, approximate secondary structure, disulfides and interfaces. Instructions prohibit internet or database access, literature or sequence/structure searches, other models and external prediction or simulation systems. Each chain must have one family hypothesis, one protein-name hypothesis, confidence and structural evidence; uncertain identities may be `null`. Host-side scoring accepts canonical labels or registered aliases after normalizing case, spaces and punctuation, without fuzzy substring matching. Family and protein-name accuracies are defined separately over chains with reference labels for each field. Null or wrong answers, incomplete chain coverage and invalid outputs are not counted as correct; unlabelled fields are excluded from the corresponding denominator. Repeated attempts are scored separately.

Reported aggregate results. The supplied summary gives 16/84 identity-recovery successes for processed, shuffled PDBs and 76/84 for the reference input (Table S2), a difference of 71.43 percentage points. These are identity-recovery counts, not mutation-effect group accuracies.

Table S2: Reported aggregate identity-recovery results for DeepSeek Flash.

Input condition	Reported correct/total	Rate (%)
Processed PDB (shuffled records)	16/84	19.05
Reference input[†]	76/84	90.48

[†]**Draft reporting note—author confirmation required.** The reference was described as "PDB or sequence"; its exact representation is not specified. The supplied counts also do not specify whether the 84 scored items are chains, focal proteins or whole tasks, or whether the field is family or specific protein name. These aggregates are therefore not assigned to either chain-level metric above. The detailed protocol describes the processed-PDB condition, not a separately specified reference-input protocol.

Interpretation. Subject to a common scoring definition across conditions, the contrast supports reduced identity recoverability after preprocessing, not elimination of memorization or proof of exclusively structural reasoning. It does not isolate the contributions of cropping, renumbering and record permutation. Family recognition and exact protein identification remain distinct.

B DETAILED CASE ANALYSES

B.1 SCOPE AND TRACEABILITY

The examples below compare model-generated explanations with structural observations and independent experimental reports. They distinguish an observation in the supplied starting structure, a hypothesis about an unobserved substituted state, and the final predicted effect direction. Agreement with a crystal structure supports particular contacts or conformations; it does not by itself determine their free-energy contributions or establish that the reported rationale is a faithful account of the model's internal computation. These selected cases do not estimate the frequency of any failure mode.

Table S3 identifies the structures and task records available for the audit. The DeepSeek case records refer to DeepSeek V4.1 Flash, abbreviated here as DeepSeek Flash; these examples are not used for additional aggregate model comparisons. Quoted outputs retain their original residue identifiers and wording. Public residue numbering is used in the accompanying discussion. For ShPI-1, literature Lys13 corresponds to chain D Lys15 in 3T62 and chain B Lys13 in 3M7Q. In the EVA-A L39P structure 8FK8, chain A author residues Pro41, Tyr46 and His30 correspond to literature Pro39, Tyr44 and His28. No task identifier was supplied for the ShPI-1–trypsin example, and no identifier is inferred here.

Table S3: Structural and task mapping for the case analyses. Starting and comparison structures refer to the direction of the prediction task; they need not correspond to natural WT and mutant, respectively.

Case	Start	Compare	Task identifier
ShPI-1 K13L, chymotrypsin	3T62	—	`task_6e6af7a717859e578544`
ShPI-1 K13L, trypsin	3M7Q	—	Not supplied
BbKI F64R, KLK4	7JQV	7JOD	`task_2530a32c128ad94f8dfc`
BbKI F64R, trypsin	7JR1	6DWH	`task_023db67447af92c088be`
EVA-A Y44A, CCL7	7SCU	8FK6	`task_85c20dee617ba09a64f1`
EVA-A L39P, CCL7	7SCU	8FK8	`task_01d69c7d396248df43b2`
BLIP E73W, CTX-M-15	7S5S	9Q0B	`task_707a59bbf9958665ae9b`

B.2 SHPI-1 K13L: INTERPRETING DIFFERENT PARTNER ENVIRONMENTS

K13L decreases the inhibition constant for bovine α-chymotrypsin from 14.8 ± 1.5 to 4.5 ± 0.5 nM, an approximately 3.3-fold improvement, but increases it for bovine cationic trypsin from 2.7 ± 0.3 to 320 ± 35 nM, an approximately 118-fold loss (García-Fernández et al., 2015). These are K_i measurements rather than direct K_D measurements. The comparison concerns a fixed chemical substitution in distinct binding environments.

For chymotrypsin, GLM-5.3 Flash emphasized burial of the Lys side chain and proposed that replacement with Leu would improve hydrophobic packing. The reasoning excerpts identify both an input-state observation and proposed energetic consequences:

GLM-5.3 Flash: WT observation.

> "Lys15 side chain is completely buried inside the chain R pocket on complex formation, including the terminal ammonium."

GLM-5.3 Flash: proposed charge effect.

> "Replacing the fully buried, countercharge-free ammonium with leucine removes the large desolvation/charge-burial penalty that the WT complex pays on binding."

GLM-5.3 Flash: proposed packing.

> "Leu's isobutyl group packs favorably against the hydrophobic pocket wall (MET192 CE/SD, VAL213 CG1, TRP215, CYS220 SG)."

The published interpretation favors increased P1 hydrophobicity in the chymotrypsin pocket (García-Fernández et al., 2015). Nevertheless, the coordinate view shows a short contact between Lys15 NZ and the Ser217 backbone oxygen (2.81 Å); the reported nearest Asp194 carboxylate contact is much farther away (8.10 Å). Thus, the Lys group lacks a nearby acidic side-chain partner but is not completely devoid of polar compensation. The absence of an ion pair alone does not quantify the net charge-burial penalty. The favorable explanation is supported at the level of pocket preference, rather than as a complete energetic decomposition.

For trypsin, the model outputs instead emphasized recognition of the basic Lys side chain by the Asp171/Ser172 region, corresponding to the canonical Asp189/Ser190 pocket numbering. DeepSeek Flash and GLM-5.3 Flash both predicted that Leu would fail to reproduce this polar network. The paired interpretations therefore respond to partner context rather than applying a uniformly favorable or unfavorable rule to K13L. Statements about a precise cavity volume or energetic penalty remain model hypotheses unless independently quantified.

B.3 BbKI F64R: SHARED HYPOTHESES AND DIFFERENT FINAL JUDGMENTS

The BbKI tasks start from Phe64 and ask about replacement with Arg64. The input structures 7JQV (KLK4) and 7JR1 (trypsin) are deposited as R64F variants of natural BbKI, whereas the comparison structures 7JOD and 6DWH contain Arg64. Accordingly, this is an F64R reversion relative to the supplied Phe64 background, not a mutation away from natural WT BbKI. The source study reports weaker inhibition after replacing the natural P1 Arg with hydrophobic residues (Li et al., 2021); the task direction reverses that comparison. Both cases discussed here have favorable F64R effects and are analyzed for mechanistic interpretation, not as an opposite-effect partner pair.

For KLK4, GLM-5.3 Flash predicted improvement and proposed Arg-mediated polar contacts and water displacement. DeepSeek Flash identified a similar opportunity near Asp189 but predicted decreased binding, assigning greater importance to lost Phe packing and an already satisfied water network. The disagreement therefore occurred despite a partly shared interaction hypothesis.

GLM-5.3 Flash: KLK4 contact hypothesis.

> "Guanidinium donates H-bonds to the SER190 backbone carbonyl (and potentially PHE215/VAL227 carbonyls) and displaces the under-coordinated buried water E490, releasing ordered water to bulk."

GLM-5.3 Flash: KLK4 ion-pair hypothesis.

> "Arg guanidinium forms a direct bidentate salt bridge with the water-satisfied ASP189 carboxylate of chain E, creating a new inter-chain ion pair absent in WT."

DeepSeek Flash: KLK4 contact hypothesis.

> "An extended Arg64 side chain points its guanidinium along the former para axis towards the buried ASP189 carboxylate (OD1/OD2 at 7.7-8.4 Å from CB, only 15-20 degrees off that axis, essentially where HOH E490 sits) and could form a new buried salt bridge/H-bond pair, displacing that water."

GLM-5.3 Flash: trypsin contact and energetic hypothesis.

> "An idealized-Arg torsion scan (248,832 combinations) shows the guanidinium can geometrically reach the adjacent buried Asp 189 (best rotamer: NH1···OD2 3.11 Å → inter-chain salt bridge) while donating extra H-bonds to chain-A backbone carbonyls (Gly 219 O at 2.73 Å, Gly 216 O at 2.96 Å) and displacing two ordered slot waters (HOH 424/405) — worth roughly -2 to -3 kcal at best."

The supplied structural audit reports an Arg64 contact with Ser190 but does not recover the proposed Phe215 and Val227 contacts in the selected KLK4 comparison. In the trypsin comparison, the reported contacts include Gly219 and Ser190 but not Gly216. These are atom- and conformation-specific comparisons: a contact missing from the observed structure is not necessarily impossible in every accessible conformation.

The Arg64–Asp189 interaction also requires careful interpretation. Hydrogen bonding and salt bridging are not mutually exclusive descriptions of an Arg–Asp contact (Donald et al., 2011). The KLK4 study reports alternate Arg64 conformations, including different direct and water-mediated interactions (Li et al., 2021). We therefore do not classify the model as wrong merely because a visualization labels a contact as a hydrogen bond. A stronger claim about a direct bidentate ion pair must be assessed against the participating atoms, distances and alternate conformations. Likewise, a numerical energetic estimate in a model excerpt is not an independent free-energy measurement; the quoted torsion-scan count is a reported model statement rather than verification of the underlying execution trace.

Across the BbKI examples, the case records contain favorable and unfavorable conclusions that recognize similar potential Arg-mediated contacts but weight their compensation for lost packing differently. They illustrate a distinction between identifying a plausible interaction opportunity and reliably predicting the net effect, without establishing how frequently this distinction explains errors.

B.4 EVA-A: FAVORABLE REARRANGEMENT HYPOTHESES THAT WERE REJECTED

For EVA-A L39P, the DeepSeek Flash and GLM-5.3 Flash case outputs predicted no improvement from the WT EVA-A–CCL7 structure 7SCU. They described buried Leu39 packing near Tyr57, His28 and Ala53, limited direct contact with CCL7 Tyr13, and a Leu39 backbone-amide contact with the Pro58 carbonyl. Pro substitution was therefore interpreted as loss of packing and a hydrogen-bond donor, with an additional rearrangement cost if the WT backbone were retained.

However, the models also proposed favorable alternatives. The following excerpts separate candidate generation from the unfavorable conclusion:

DeepSeek Flash: L39P competing hypotheses.

> "If the pocket is tight around the Tyr, replacing Leu with Pro (which would leave a small void near CD1's position) could either: (a) destabilize binding slightly (lost VdW + possibly allow water in), (b) allow the Tyr to relax into the void, possibly making better contacts (unlikely to improve much)."

GLM-5.3 Flash: L39P repositioning hypothesis.

> "(c) Pro-induced kink could reposition the 7734 strand relative to R to make better contacts? Pure speculation, no evidence in WT geometry: the R contact is one loose TYR."

L39P unfavorable conclusion quoted in the case analysis.

> "Overall: the mutation is clearly not improving binding; it reduces it — primarily by destabilizing the interface-adjacent sheet (lost N-H donor at a nearly buried unsatisfiable carbonyl + mandatory clash-relieving rearrangement coupled to the interface) and by removing the side-chain volume that packs the sealed pocket against the partner (R6694 TYR), with no offsetting gain."

Y44A favorable alternative quoted in the case analysis.

> "H2 (yes, stronger binding): The WT pocket may be suboptimal: Y7371 (partner) is squeezed against Y4560 (contacts at 3.93 Å CB...CE2 and 4.26 CG...CE2); removing Y4560 opens the pocket, allowing Y7371 to relax into a better-packed, lower-strain conformation, potentially improving the π/CH contact complementarity and letting the disulfide loop (C8006-C9616-Y7371-R3042) adopt a more favorable conformation."

Y44A unfavorable conclusion quoted in the case analysis.

> "Given the structural analysis strongly indicates a favorable WT contact network that is lost by Y$\rightarrow$A (with no obvious strain relief or clash), "no" is the best answer."

In the L39P structure 8FK8, CCL7 Tyr13 occupies the CC+1 pocket more deeply and forms favorable aromatic interactions with EVA-A Tyr44 and His28 (Devkota et al., 2023). The model's alternative of allowing the partner Tyr to repack is therefore directionally consistent with the observed change, even though the excerpt is not a complete prediction of that structure. The error in these outputs

includes selecting against a potentially useful hypothesis, rather than simply failing to generate any rearrangement mechanism.

Y44A provides a related case. The four model outputs discussed in the case records predicted no improvement, emphasizing Tyr44 burial, packing and local stabilization. Yet the quoted favorable alternative considered opening the pocket to permit better accommodation of CCL7 Tyr13. The original structural interpretation is more specific: the mutation avoids the Tyr44 side-chain repositioning required for WT EVA-A binding, while retaining an otherwise similar complex (Devkota et al., 2023). Thus, the generated alternative and the published explanation are related but not identical. The model ultimately preferred the loss-of-contact explanation because it did not observe an obvious clash in the bound WT structure.

These cases are consistent with stronger weighting of directly observable starting-state interactions than of favorable but unobserved relaxation. We use WT-state anchoring as a description of this pattern in the selected outputs, not as an established internal mechanism or a claim about the models' training distributions. Neither static structures nor the explanation text alone determine the complete conformational free-energy balance.

B.5 BLIP E73W: CORRECT DIRECTION AND INCOMPLETE CONTACT PREDICTION

For CTX-M-15, the published inhibition constant decreases from 2.70 ± 0.25 nM for WT BLIP to 0.33 ± 0.01 nM for E73W under the reported nitrocefin assay, corresponding to approximately 8.2-fold stronger inhibition (Rivera et al., 2025). GLM-5.3 Flash predicted improvement from the WT complex 7S5S. Its explanation combined a burial observation with specific hypotheses about charge compensation, water and aromatic packing:

GLM-5.3 Flash: WT observation.

> "Glu73 side chain is essentially fully buried in the complex ... Side-chain SASA 6.7 A^2 in the complex vs 32.9 A^2 for chain B alone."

GLM-5.3 Flash: aromatic-contact hypothesis.

> "After mutation the indole is expected to intercalate into the aromatic pocket, stacking against Trp112/Tyr143/Tyr53 and contacting the partner chain near Asn104/Pro107."

GLM-5.3 Flash: charge and water hypothesis.

> "Substitution removes a buried carboxylate that is only partially H-bond satisfied ... and expels the small ordered water cluster."

The experimental E73W complex 9Q0B shows a prominent contact of Trp73 with CTX-M-15 Tyr105. The source study links the altered binding mode to hydrophobic interactions and an apo-like partner conformation (Rivera et al., 2025). This differs from the excerpt's emphasis on packing against Trp112, Tyr143 and Tyr53. The model therefore anticipated a broad hydrophobic contribution but did not fully recover the observed contact pattern and conformational interpretation.

The evidence does not require every proposed contribution, such as water release, to be absent. It establishes that a correct effect label and a chemically plausible explanation are insufficient to demonstrate the specific mechanism. This complements the EVA-A cases: the latter contain potentially useful hypotheses despite incorrect final labels, whereas BLIP has a correct label despite incomplete mechanistic agreement.

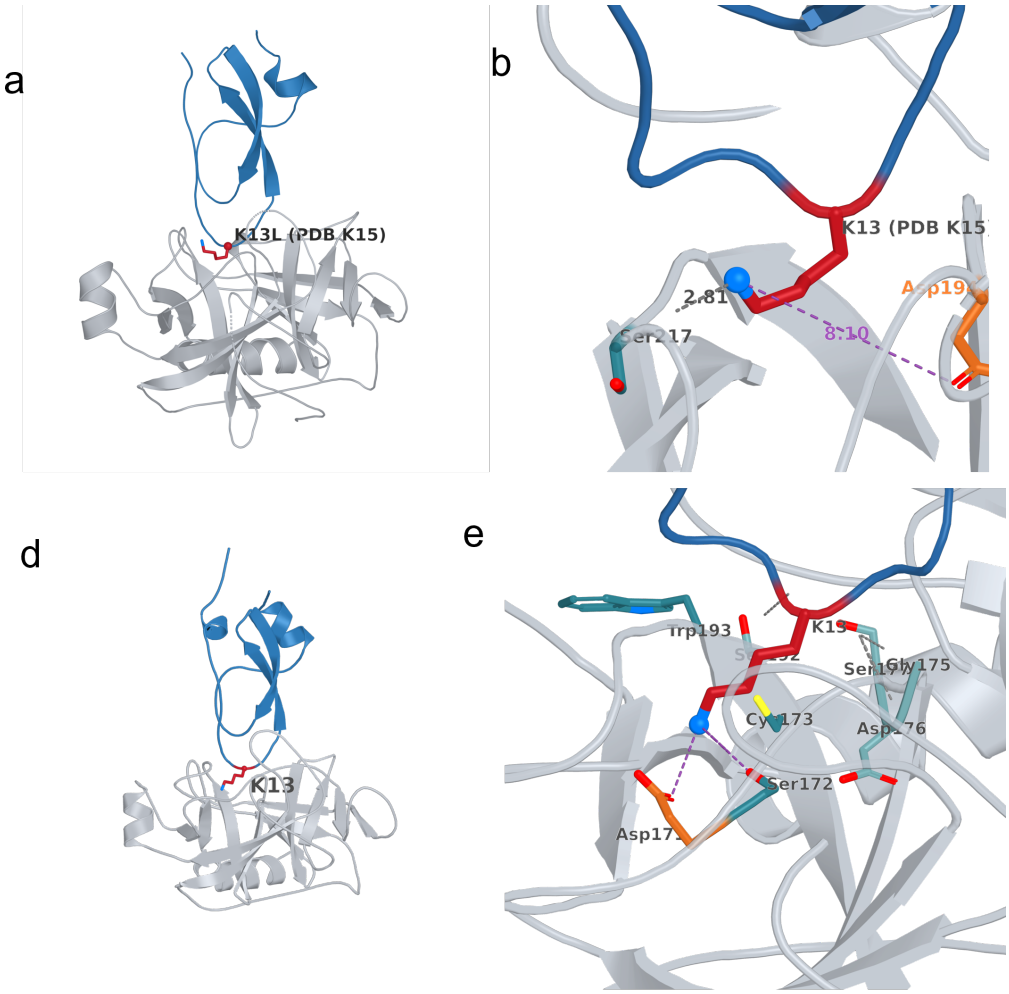


Figure S1: **Additional views of ShPI-1 K13L.** a,b, Chymotrypsin complex 3T62 and the local environment of literature Lys13 (PDB chain D Lys15), including partial polar compensation by Ser217. d,e, Trypsin complex 3M7Q and the local Lys13 recognition network. Views and original panel letters are reproduced from the case-analysis material; the highlighted Lys residues are the observed starting residues, not modeled Leu conformations.

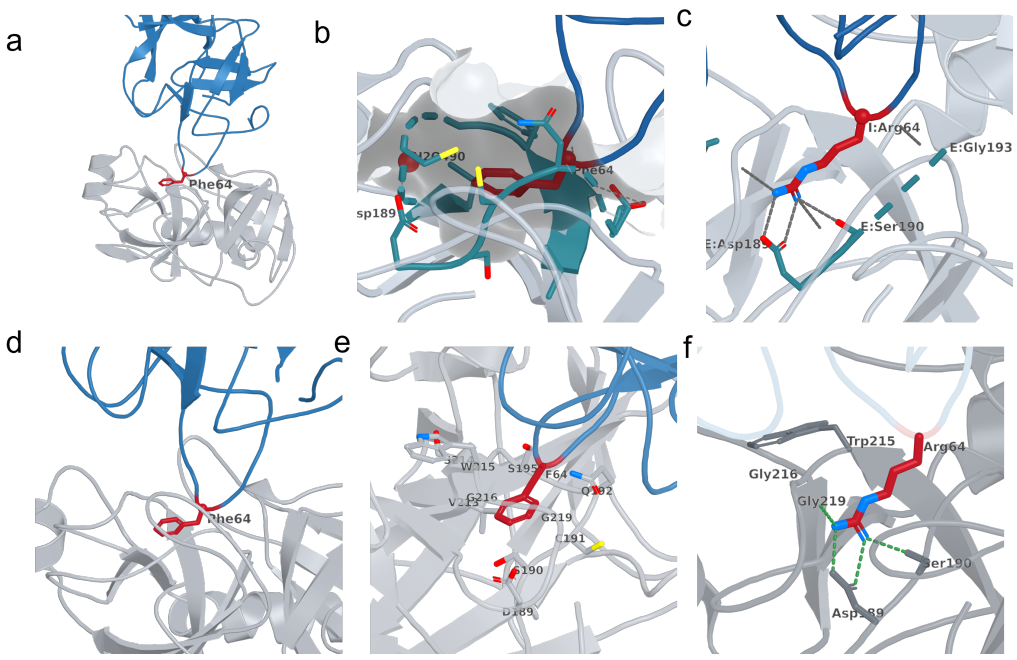


Figure S2: **Additional views of BbKI F64R.** a–c, Phe64 starting complex with KLK4 (7JQV), its local pocket, and the Arg64 comparison (7JOD). d–f, Phe64 starting complex with trypsin (7JR1), its local pocket, and the Arg64 comparison (6DWH). Starting and comparison designate task direction; the Phe64 proteins are R64F variants of natural BbKI. Displayed polar-contact lines do not distinguish hydrogen bonds from salt bridges or resolve all alternate conformations.

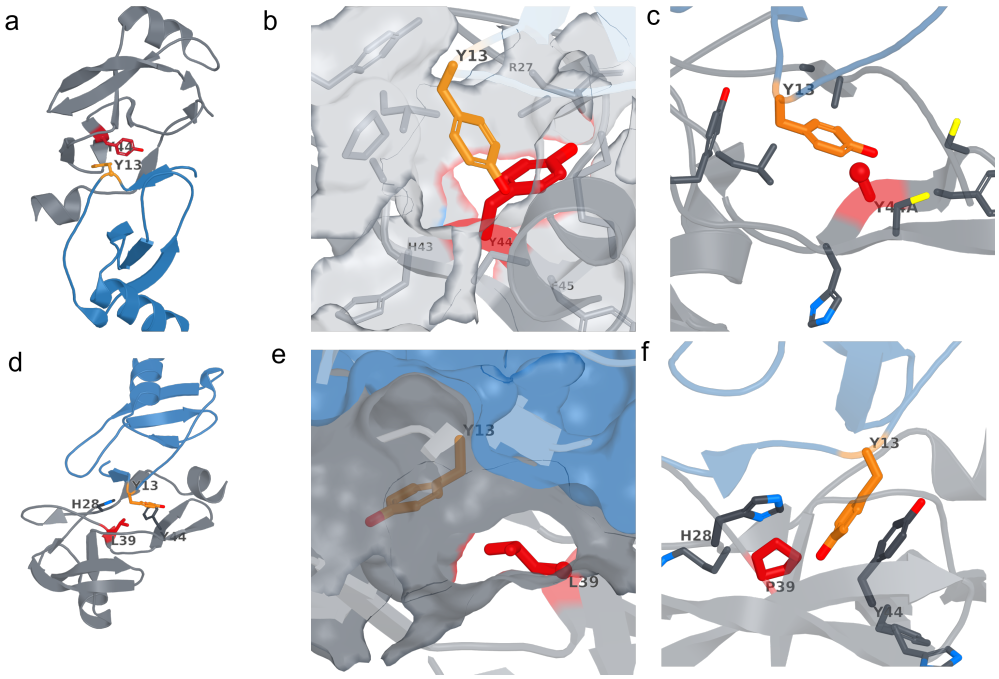


Figure S3: **Additional views of EVA-A Y44A and L39P.** a–c, WT EVA-A–CCL7 (7SCU), the Tyr44 pocket and the Y44A comparison (8FK6). d–f, WT interface, the Leu39 region and the L39P comparison (8FK8). Labels use literature residue numbering. The views illustrate structural changes discussed in the source study; they do not represent structures generated by the LLMs.

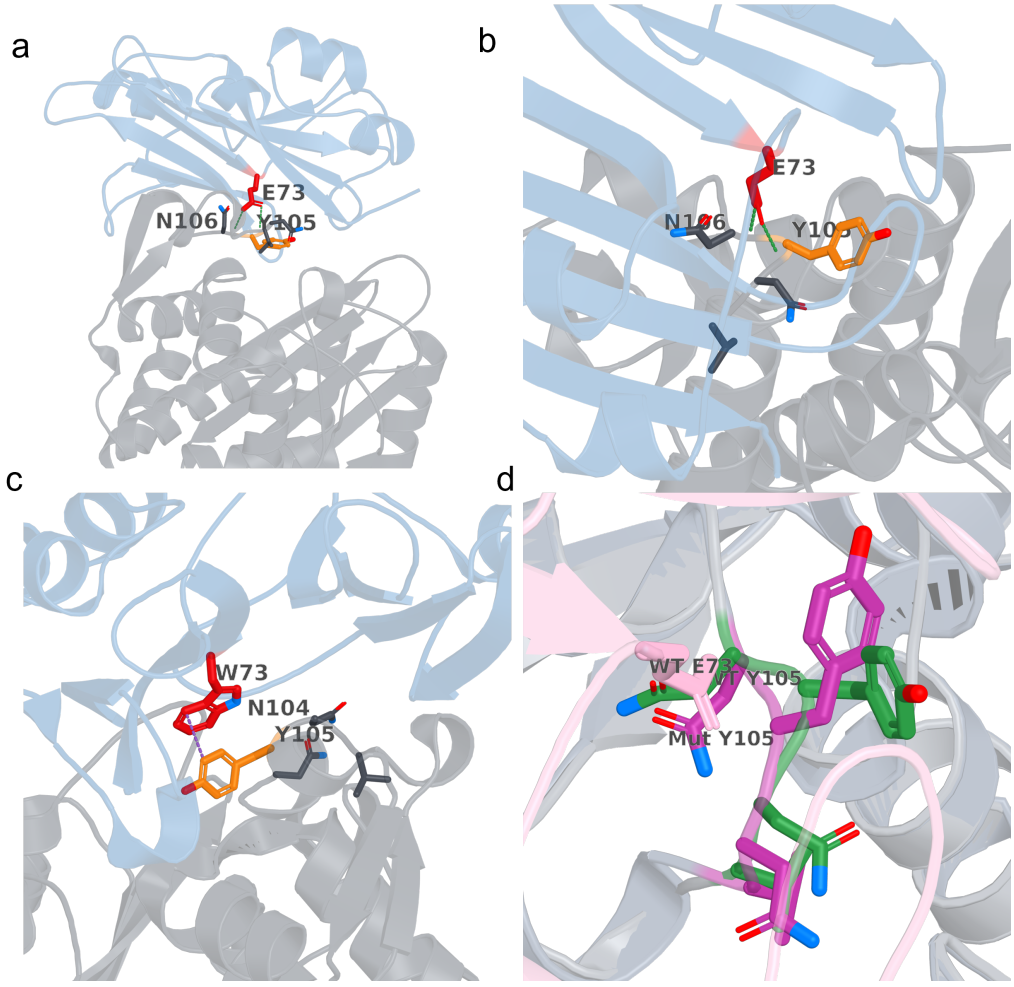


Figure S4: **Additional views of BLIP E73W.** a,b, WT BLIP–CTX-M-15 complex 7S5S and the Glu73/partner-loop environment. c, Experimental E73W complex 9Q0B. d, Structural overlay illustrating the change around BLIP residue 73 and CTX-M-15 Tyr105. These observed structures provide a comparison for the model's proposed contacts; they do not independently quantify charge-burial or water-release energies.

C MODEL INFERENCE CONFIGURATIONS

The configurations below record the API identifiers and request settings for the four LLMs. All entries use the `openai_compatible` provider interface, `timeout_seconds=600` and `max_retries=3`. The retry setting concerns API requests, not the three runs used for majority-vote consensus. A machine-readable copy is included in config/model_inference_config.json.

C.1 MODEL IDENTIFIERS AND GENERATION SETTINGS

Table S4 distinguishes manuscript names from the model identifiers sent to the API. In particular, the configuration entry `deepseek-v4.1-flash` sends `deepseek-v4-flash` as its `model` value; both identifiers are retained as configured.

Table S4: Configured model identifiers and generation settings.

Manuscript name	API `model`	Temperature	Token limit
DeepSeek Flash	`deepseek-v4-flash`	0.0	86,384
GLM-5.3 Flash	`glm-5.3-flash`	1.0	131,072
GLM-5.3	`glm-5.3`	1.0	131,072
Qwen3.8 Flash	`qwen3.8-flash`	1.0	131,072

The token-limit field is `max_tokens` for DeepSeek and both GLM models, and `max_completion_tokens` for Qwen3.8 Flash. These are configured request limits, not measured token usage.

C.2 REASONING AND ADDITIONAL REQUEST FIELDS

The fields in Table S5 are passed through `extra_body`. Dot notation denotes nested JSON fields; a dash means the field is absent from the supplied configuration, not that it is disabled. The two GLM models share these settings.

Table S5: Configured fields within `extra_body`.

Field	DeepSeek Flash	Both GLM models	Qwen3.8 Flash
`thinking.type`	`enabled`	`enabled`	—
`thinking.clear_thinking`	—	`false`	—
`top_p`	—	0.95	—
`enable_thinking`	—	—	`true`
`reasoning_effort`	`max`	`max`	`xhigh`
`preserve_thinking`	—	—	`true`

C.3 ENDPOINTS AND CREDENTIAL-VARIABLE NAMES

The configured `base_url` values and `api_key_env` names are listed below. Only environment-variable names are recorded; no API credentials are included.

DeepSeek Flash. https://api.deepseek.com
`api_key_env`: `DEEPSEEK_API_KEY`.

GLM-5.3 Flash and GLM-5.3. https://open.bigmodel.cn/api/paas/v4
`api_key_env`: `ZAI_API_KEY`.

Qwen3.8 Flash. https://ws-0ipxolve4xo489fe.cn-beijing.maas.aliyuncs.com/compatible-mode/v1
`api_key_env`: `DASHSCOPE_API_KEY`.

D PROTEIN-SYSTEM DISTRIBUTION OF SUCCESSFUL JUDGMENTS

Figure S5 gives the system-level counts underlying Section 3.4. All LLM rows are Flash models. The Python-assisted DeepSeek, GLM and Qwen3.8 rows cover four, five and four systems, respectively, and collectively seven of the 17 Dataset 1 systems. Counts are reproduced from the supplied system summary; they do not identify exact individual-group overlap or add independent model runs.

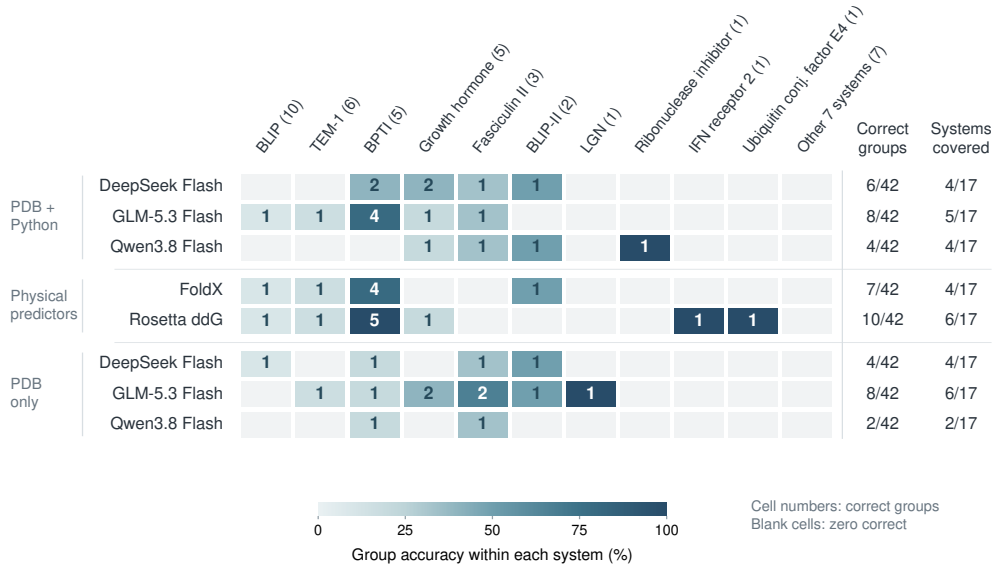


Figure S5: **Successful judgments across protein systems.** Cells show correct-group counts; shading denotes within-system group accuracy on Dataset 1. Blanks indicate zero. Column-label parentheses give group counts; the seven pooled systems have no successes. Right columns report correct groups and systems covered. All LLM rows are Flash models. System-level counts do not establish individual-group overlap.

E STRUCTURAL-CLAIM EXTRACTION AND AUDIT PROCEDURE

This appendix details the structural-claim audit in Section 4.3. Mol5 refers here to the Mol* 5.11.0 interaction analysis.

E.1 CLAIM EXTRACTION AND AUDIT SCOPE

We scanned the structural-evidence text in each `prediction.json` file and split independently verifiable relationships within a sentence into atom- or residue-level claims. For each claim, we recorded the mutated residue, interaction type, whether the relationship was asserted or negated, interaction target, chain, residue number, reported distance, inter-chain or intra-chain scope, and backbone or side-chain scope. A single evidence entry could yield multiple claims; evidence-entry counts and claim counts therefore need not be equal.

The recognized interaction types were hydrogen bonds, weak hydrogen bonds, ionic interactions or salt bridges, water bridges, hydrophobic contacts, π–π stacking, cation–π interactions, halogen bonds and metal coordination. Solvent-accessible surface area (SASA), cavity volume, burial, B factors, flexibility and hypothesized post-mutation rearrangements were not directly mapped to Mol5 interaction claims and were outside the audit scope.

E.2 STRUCTURE ANALYSIS AND CLAIM MATCHING

Input PDB structures were analyzed with Mol* 5.11.0, with all interaction providers enabled and the default Mol5 distance and geometry criteria retained. Claims were matched to Mol5 records by task, PDB, mutation site, interaction type, target chain and residue. Where the model reported a distance, a numerical tolerance of 0.15 Å was applied only when the distance definitions were consistent.

E.3 MANUAL REVIEW

All 12 records automatically flagged as conflicts were subsequently inspected manually. Particular attention was given to records involving negation, multiple residues, multiple distances, water bridges, salt-bridge distances, and ambiguity about backbone versus side-chain scope.

F STRUCTURE-GUIDED MUTATION BINDING PREDICTION

Prompt configuration. All mutation-binding tasks were evaluated using a standardized prompt designed to frame each example as a prospective structure-based drug-design decision rather than a benchmark-label classification problem. The model received only the anonymized 20 Å-cropped PDB coordinates, chain assignments, and mutation metadata; residue indices could be randomly permuted to reduce the possibility of recovering protein identity from numbering. External databases, literature, experimental affinity data, protein-identity lookup, pretrained predictive systems, docking or molecular-simulation software, and other external modeling tools were explicitly prohibited. The model was permitted to use general structural and biophysical knowledge and to execute self-written deterministic Python analyses on the supplied coordinates. The prompt further required thermodynamic reasoning about the mutation-induced change in binding free energy, explicit consideration of plausible mutant-state conformations and local repacking, and evaluation of both affinity-improving and affinity-reducing mechanisms before committing to a final binary prediction. Responses were returned in the requested JSON format with a concise structure-based rationale.

F.1 FULL EVALUATION PROMPT

Treat this as a prospective real-world drug-design decision rather than a benchmark classification task: reason as an experienced medicinal chemistry and structure-based drug design expert.

F.2 STRUCTURAL MUTATION BINDING PREDICTION

For each complex independently, predict whether the specified mutation in chain M improves binding affinity relative to the wild-type complex.

- "yes" = improves binding affinity
- "no" = reduces binding affinity

Use only:

- the supplied anonymized PDB coordinates;
- chain assignments;
- mutation metadata;
- general knowledge of protein structure, chemistry, biophysics, and molecular interactions.

The current PDB structure is a cropped region (20Å) around the mutation site rather than the complete structure. To prevent data leakage or cheating, the residue numbering within each chain may have been randomly permuted.

Do not infer the possible identity of the protein from its local structure, and do not use such an inferred identity to search for the protein itself or for existing studies related to it.

All residue indices in the response must follow the residue numbering in the original PDB file.

You may use broad expert reasoning and are not limited to any predefined checklist. You may consider any structurally or chemically relevant factor, including plausible mutant conformations and local rearrangements.

You may write and execute your own Python code for deterministic analysis directly from the supplied coordinates, such as distances, contacts, SASA, burial, packing, sterics, dihedrals, interface geometry, or any other useful calculation.

F.2.1 AVAILABLE PYTHON ENVIRONMENT

The current Python environment already provides the following packages for deterministic structural analysis:

- Python 3.11

- NumPy
- SciPy
- pandas
- Biopython
- standard Python libraries

You may use functions provided by these installed packages when useful, including numerical calculations, linear algebra, spatial searches, distance calculations, coordinate manipulation, PDB parsing, and other deterministic operations.

The execution environment is fixed.

F.2.2 STRICTLY PROHIBITED

Do not use or call any external:

- database, internet source, literature, or experimental affinity data;
- protein/PDB identity information;
- AI model, LLM, ML/DL model, protein language model, or external agent;
- affinity or $\Delta\Delta G$ predictor;
- structure/mutation prediction tool;
- rotamer or side-chain modeling software;
- docking, MD, Rosetta, force-field, energy-minimization, or free-energy software;
- other pre-existing predictive/modeling system.

Do not attempt to identify the original structure or biological system.

The rule is simple:

You may reason yourself and write your own deterministic code, but you may not outsource the analysis to external models, AI tools, databases, predictors, or modeling software.

You may qualitatively infer plausible mutant-side-chain conformations from the wild-type structure and amino-acid chemistry. You do not need to assume that the mutant retains the exact wild-type conformation.

Return the answer in the requested JSON format with a concise structure-based reason.

F.2.3 THERMODYNAMIC REASONING REQUIREMENT

Predict the sign of the mutation-induced change in binding free energy, rather than judging whether the wild-type bound structure contains favorable contacts. A favorable-looking WT hydrogen bond, salt bridge, or buried polar interaction is not automatically evidence against mutation, because it may compensate for desolvation, conformational strain, buried charge, or entropy costs. Consider both the bound and approximate unbound-state consequences of the substitution.

F.2.4 MUTANT-STATE REQUIREMENT

Actively consider plausible alternative mutant side-chain conformations and local repacking. Do not anchor the mutant to the wild-type terminal-atom position or assume that preservation of the WT contact network is optimal.

Before making the final yes/no prediction, explicitly consider both competing hypotheses:

1. how the mutation could improve binding affinity, and
2. how it could reduce binding affinity.

Do not commit to either direction until both possibilities have been evaluated.

Compare the strongest plausible thermodynamic mechanism on each side, including effects that may not be obvious from the WT bound contacts, and choose the outcome that is more likely overall.

Table S6: Structural and task mapping for the case analyses. Starting and comparison structures refer to the direction of the prediction task; they need not correspond to natural WT and mutant, respectively.

Case	**Start**	**Compare**	**Task identifier**
ShPI-1 K13L, chymotrypsin	3T62	—	task_6e6af7a717859e578544
ShPI-1 K13L, trypsin	3M7Q	—	task_f5a99eeddcc998910be6
BbKI F64R, KLK4	7JQV	7JOD	task_2530a32c128ad94f8dfc
BbKI F64R, trypsin	7JR1	6DWH	task_023db67447af92c088be
EVA-A Y44A, CCL7	7SCU	8FK6	task_85c20dee617ba09a64f1
EVA-A L39P, CCL7	7SCU	8FK8	task_01d69c7d396248df43b2
BLIP E73W, CTX-M-15	7S5S	9Q0B	task_707a59bbf9958665ae9b

G Dataset Composition and Structural Complex Curation

G.1 Dataset 1: SKEMPI 2.0-derived benchmark

Dataset 1 composition and screening. Dataset 1 was screened from SKEMPI 2.0. Mutation cases were grouped by the same full wild-type protein sequence and mapped single-amino-acid substitution, and each retained case contained two distinct partner full-WT sequences. The benchmark retained cases with opposite binding effects across the two partners and with the mutation site located at the protein-protein interface in both selected complex structures under the 5 Å interface criterion used during dataset construction. The resulting Dataset 1 contains 17 distinct mutated-WT protein systems, 42 mutation cases, 34 WT-partner relationships, and 84 mutation-partner observations. Across all systems, the partners correspond to 33 unique full-WT sequence hashes.

Table S7: Composition of Dataset 1 by mutated-WT protein system.

Mutated WT	**Mutation cases**	**Unique partners**	**Partner observations**
BLIP	10	2	20
TEM-1 beta-lactamase	6	2	12
BPTI	5	2	10
Human growth hormone	5	2	10
Fasciculin II	3	2	6
BLIP-II	2	2	4
BoNT/A1	1	2*	2
Colicin E9 DNase	1	2	2
G-protein-signaling modulator 2 (LGN)	1	2	2
GM-CSF	1	2	2
IgG1-kappa D1.3 Fv	1	2	2
Interferon alpha/beta receptor 2	1	2	2
Membrane-type serine protease 1	1	2	2
PTP delta	1	2	2
Raf-RBD	1	2	2
Ribonuclease inhibitor	1	2	2
Ubiquitin conjugation factor E4	1	2	2
Total	42	34 WT-partner relationships	84

Table note. “Mutation cases” denotes unique mapped single-amino-acid substitutions within each mutated-WT protein system. “Unique partners” is counted within each mutated-WT system; the total therefore represents 34 WT-partner relationships. When partner sequences are globally deduplicated across all systems by full-WT sequence hash, Dataset 1 contains 33 unique partner sequences. *For BoNT/A1, both partners are named AR2 mab in the source table, but their full-WT sequence hashes differ; they are therefore counted as two partners under the benchmark partner-identity definition.

G.2 LLM-generated structural evidence in Dataset 1

We further analyzed the structural-evidence statements generated by the LLMs for Dataset 1. Among 1,574 interaction statements supported by Mol5 structural analysis, hydrogen bonds were the most frequently cited basis for prediction, with 1,326 statements (84.24%), followed by ionic interactions or salt bridges, with 166 statements (10.55%). Together, these two polar or electrostatic interaction categories accounted for 94.79% of the supported statements, indicating that the models strongly favored hydrogen-bond and ionic-interaction evidence when presenting structural rationales. In comparison, hydrophobic or van der Waals contacts accounted for 2.86%, water bridges for 1.08%, pi-pi stacking for 0.95%, and cation-pi interactions for 0.32%. Importantly, this distribution reflects the models’ preferences for citing different interaction types in their structural evidence, rather than the actual abundance of these interaction types across all input complexes.

Table S8: Interaction types among LLM-generated structural evidence statements in Dataset 1 supported by Mol5 structural analysis.

Interaction type	Count	Proportion
Hydrogen bonds	1,326	84.24%
Ionic interactions / salt bridges	166	10.55%
Hydrophobic / van der Waals contacts	45	2.86%
Water bridges	17	1.08%
Pi-pi stacking	15	0.95%
Cation-pi interactions	5	0.32%
Total	1,574	100.00%

Table note. Percentages summarize the interaction types cited in the model-generated structural evidence; they do not represent the physical prevalence of these interaction types in the complete set of input complexes.

G.3 Dataset 2: Literature-curated, structure-supported collection

Structural complex collection and curation. For Dataset 2, we first identified candidate systems with experimentally determined protein-complex structures in the Protein Data Bank (PDB), and then traced the corresponding primary literature to collect experimentally measured mutation effects for the relevant partners. Exact full-length sequence identity between the biochemical assay construct and the crystallographic construct was not required automatically. When the assay and structural constructs differed, we manually checked whether the mutation could be mapped unambiguously, whether the mutation-centered binding interface was preserved, and whether the construct difference was likely to affect interpretation of the measured interaction. Entries were retained when the available structure remained representative of the interface used for the mutation-partner comparison.

Residue and chain mapping. For retained entries, mutation numbering was cross-mapped among the PDB coordinates, UniProt sequence, and the numbering used in the experimental report. The focal and partner chains used for the structural comparison were recorded explicitly.

Dataset 2 composition. Dataset 2 contains 11 mutated-WT protein systems, 18 mutation cases, 26 WT-partner relationships, and 42 mutation-partner observations. Its family-level composition is summarized in Supplementary Table S9, and the observation-level structural and assay annotations are listed in Supplementary Table S10.

Table S9: Composition of Dataset 2 by mutated-WT protein system.

Mutated WT	UniProt	Mutation cases	Unique partners	Partner observations
BIM / BCL2L11	O43521	4	2	8
EVA-P974	A0A023FDY8	2	2	4
EVA-A	P1243	2	4	8
BbKI	P83052	2	3	5
BMP2	P12643	2	2	4
TP53	P04637	1	2	2
Human IgG1 Fc	P01857	1	3	3
ActRIIB / Acvr2b	P27040	1	2	2
BLIP	P35804	1	2	2
ShPI-1	P31713	1	2	2
E. coli PotF	P31133	1	2	2
Total	—	18	26 WT–partner relationships	42

Table note. “Mutation cases” denotes unique amino-acid substitutions within a focal-protein system; “Unique partners” denotes distinct WT-partner relationships represented for that system; and “Partner observations” denotes individual mutation-partner measurements included in Dataset 2.

Table S10: Curated mutation-partner complexes and structural mappings for Dataset 2. (a) Structural mapping.

Group	Focal protein	UniProt	PDB ID	Mutation (PDB)	Mutation (UniProt)	Mutation (literature)	Focal chain	Partner chain
1	EVA-P974	A0A023FDY8	7S58	A:F23A	F31A	F31A	A	H
1	EVA-P974	A0A023FDY8	7S4N	A:F23A	F31A	F31A	A	B
2	EVA-P974	A0A023FDY8	7S58	A:Y4W	Y12W	Y12W	A	H
2	EVA-P974	A0A023FDY8	7S4N	A:Y4W	Y12W	Y12W	A	B
4	EVA-A	P1243	7SCU	A:Y44A	Y44A	Y44A	A	B
4	EVA-A	P1243	7SCS	A:Y44A	Y44A	Y44A	A	B
4	EVA-A	P1243	7SCT	A:Y44A	Y44A	Y44A	A	B
4	EVA-A	P1243	7SCV	A:Y44A	Y44A	Y44A	A	B
5	EVA-A	P1243	7SCU	A:L39P	L39P	L39P	A	B
5	EVA-A	P1243	7SCS	A:L39P	L39P	L39P	A	B
5	EVA-A	P1243	7SCT	A:L39P	L39P	L39P	A	B
5	EVA-A	P1243	7SCV	A:L39P	L39P	L39P	A	B
6	TP53	P04637	1YCR	B:P27A	P27A	P27A	B	A
6	TP53	P04637	3DAB	B:P27A	P27A	P27A	B	A
7	BbKI-R64F	P83052	7JQV	I:F64R	R64F	R64F	I	E
7	BbKI-R64F	P83052	7JR1	G:F64R	R64F	R64F	G	A
7	BbKI-R64F	P83052	7JRX	I:F64R	R64F	R64F	I	A/B/C
8	BbKI-R64M	P83052	7JQN	I:M64R	R64M	R64M	I	E
8	BbKI-R64M	P83052	7JR2	G:M64R	R64M	R64M	G	A
9	BLIP	P35804	7S5S	B:E73W	E73W	E73W	B	A
9	BLIP	P35804	1JTG	B:E73W	E73W	E73W	B	A
10	Human IgG1 Fc	P01857	4ZNE	E/J:G236A	G236A	EU G236A	E/J	A
10	Human IgG1 Fc	P01857	9MCX	A/B:G236A	G236A	EU G236A	A/B	C
10	Human IgG1 Fc	P01857	9MCY	A/B:G236A	G236A	EU G236A	A/B	E
11	BIM / BCL2L11	O43521	4QVF	B:I65A	I155A	I95A	B	A
11	BIM / BCL2L11	O43521	2PQK	B:I13A	I155A	I95A	B	A
12	BIM / BCL2L11	O43521	4QVF	B:F69A	F159A	F99A	B	A
12	BIM / BCL2L11	O43521	2PQK	B:F17A	F159A	F99A	B	A
13	BIM / BCL2L11	O43521	4QVF	B:Y73K	Y163K	Y103K	B	A
13	BIM / BCL2L11	O43521	2PQK	B:Y21K	Y163K	Y103K	B	A
14	BIM / BCL2L11	O43521	4QVF	B:I65F	I155F	I3dF (Bim I95F)	B	A
14	BIM / BCL2L11	O43521	2PQK	B:I13F	I155F	I3dF (Bim I95F)	B	A

Table S10 (continued)

Group	Focal protein	UniProt	PDB ID	Mutation (PDB)	Mutation (UniProt)	Mutation (literature)	Focal chain	Partner chain
15	ActRIIB / Acvr2b (mouse extracellular domain)	P27040	2H62	D:K37A	K55A	K37A	D	A/B
15	ActRIIB / Acvr2b (mouse extracellular domain)	P27040	1S4Y	A:K33A;C:K33A	K55A	K37A	A/C	B/D
18	BMP2	P12643	2H62	A:S85R;B:S85R	S367R	S85R	A/B	D
18	BMP2	P12643	2GOO	A:S85R;D:S85R	S367R	S85R	A/D	C/F
19	BMP2	P12643	2H62	A:L100K; B:L100K	L382K	L100K	A/B	D
19	BMP2	P12643	2GOO	A:L100K; D:L100K	L382K	L100K	A/D	C/F
20	ShPI-1	P31713	3M7Q	B:K13L	K13L	K13L	B	A
20	ShPI-1	P31713	3T62	D:K15L	K13L	K13L	D	A
21	E. coli PotF	P31133	6YE8	B:F276W	F276W	F276W	B	L
21	E. coli PotF	P31133	6YE0	B:F276W	F276W	F276W	B	L

Table S10: Curated mutation-partner complexes and structural mappings for Dataset 2. (b) Binding partners and experimental annotations.

Group	PDB ID	Binding partner	Endpoint	Measurement method	Change vs WT	Reference
1	7S58	CCL7	Binding affinity (KD)	Surface plasmon resonance (SPR); KD obtained by fitting the binding response with a 1:1 kinetic model.	Stronger	doi:10.1073/pnas.2122105119
1	7S4N	CCL17	Binding affinity (KD)	Surface plasmon resonance (SPR); KD obtained by fitting the binding response with a 1:1 kinetic model.	Weaker	doi:10.1073/pnas.2122105119
2	7S58	CCL7	Binding affinity (KD)	Surface plasmon resonance (SPR); KD obtained by fitting the binding response with a 1:1 kinetic model.	Weaker	doi:10.1073/pnas.2122105119
2	7S4N	CCL17	Binding affinity (KD)	Surface plasmon resonance (SPR); KD obtained by fitting the binding response with a 1:1 kinetic model.	Stronger	doi:10.1073/pnas.2122105119
4	7SCU	CCL7	Binding affinity (KD)	Surface plasmon resonance (SPR); KD obtained by fitting the binding response with a 1:1 kinetic model.	Stronger	doi:10.1038/s41467-023-39879-3
4	7SCS	CCL11	Binding affinity (KD)	Surface plasmon resonance (SPR); KD obtained by fitting the binding response with a 1:1 kinetic model.	Stronger	doi:10.1038/s41467-023-39879-3
4	7SCT	CCL16	Binding affinity (KD)	Surface plasmon resonance (SPR); KD obtained by fitting the binding response with a 1:1 kinetic model.	Weaker	doi:10.1038/s41467-023-39879-3
4	7SCV	CCL17	Binding affinity (KD)	Surface plasmon resonance (SPR); KD obtained by fitting the binding response with a 1:1 kinetic model.	Weaker	doi:10.1038/s41467-023-39879-3
5	7SCU	CCL7	Binding affinity (KD)	Surface plasmon resonance (SPR); KD obtained by fitting the binding response with a 1:1 kinetic model.	Stronger	doi:10.1038/s41467-023-39879-3
5	7SCS	CCL11	Binding affinity (KD)	Surface plasmon resonance (SPR); KD obtained by fitting the binding response with a 1:1 kinetic model.	Stronger	doi:10.1038/s41467-023-39879-3
5	7SCT	CCL16	Binding affinity (KD)	Surface plasmon resonance (SPR); KD obtained by fitting the binding response with a 1:1 kinetic model.	Weaker	doi:10.1038/s41467-023-39879-3

Table S10 (continued)

Group	PDB ID	Binding partner	Endpoint	Measurement method	Change vs WT	Reference
5	7SCV	CCL17	Binding affinity (KD)	Surface plasmon resonance (SPR); KD obtained by fitting the binding response with a 1:1 kinetic model.	Weaker	doi:10.1038/s41467-023-39879-3
6	1YCR	MDM2 / Q00987	Binding affinity (KD)	SPR-based competition binding assay; equilibrium KD obtained by nonlinear fitting of peptide competition data.	Stronger	doi:10.1016/j.jmb.2010.03.005
6	3DAB	MDMX / O15151	Binding affinity (KD)	SPR-based competition binding assay; equilibrium KD obtained by nonlinear fitting of peptide competition data.	Stronger	doi:10.1016/j.jmb.2010.03.005
7	7JQV	KLK4 / Q9Y5K2	Inhibitory potency (Ki)	Protease inhibition assay; Ki determined from concentration-dependent inhibition of enzyme activity.	Stronger	doi:10.1107/S2059798321006483
7	7JR1	Bovine trypsin / P00760	Inhibitory potency (Ki)	Protease inhibition assay; Ki determined from concentration-dependent inhibition of enzyme activity.	Stronger	doi:10.1107/S2059798321006483
7	7JRX	Alpha-chymotrypsin / P00766	Inhibitory potency (Ki)	Protease inhibition assay; Ki determined from concentration-dependent inhibition of enzyme activity.	Weaker	doi:10.1107/S2059798321006483
8	7JQN	KLK4 / Q9Y5K2	Inhibitory potency (Ki)	Protease inhibition assay; Ki determined from concentration-dependent inhibition of enzyme activity.	Stronger	doi:10.1107/S2059798321006483
8	7JR2	Bovine trypsin / P00760	Inhibitory potency (Ki)	Protease inhibition assay; Ki determined from concentration-dependent inhibition of enzyme activity.	Stronger	doi:10.1107/S2059798321006483
9	7S5S	CTX-M-15 / A0A5R8T042	Inhibitory potency (Ki)	Nitrocefin hydrolysis inhibition assay; Ki obtained by fitting residual activity data to the Morrison equation.	Stronger	doi:10.1016/j.jbc.2025.110850
9	1JTG	TEM-1 / P62593	Inhibitory potency (Ki)	Nitrocefin hydrolysis inhibition assay; Ki obtained by fitting residual activity data to the Morrison equation.	Stronger	doi:10.1016/j.jbc.2025.110850
10	4ZNE	Human Fc gamma RI / FCGR1A / P12314	Binding affinity (KD)	Biacore SPR; KD obtained by global fitting to a Langmuir binding model.	Weaker	doi:10.1158/1535-7163.MCT-08-0201
10	9MCX	Human Fc gamma RIIa-H131 / FCGR2A / P12318	Binding affinity (KD)	Biacore SPR; KD obtained by global fitting to a Langmuir binding model.	Stronger	doi:10.1158/1535-7163.MCT-08-0201

Table S10 (continued)

Group	PDB ID	Binding partner	Endpoint	Measurement method	Change vs WT	Reference
10	9MCY	Human Fc gamma RIIa-R131 / FCGR2A / P12318	Binding affinity (KD)	Biacore SPR; KD obtained by global fitting to a Langmuir binding model.	Stronger	doi:10.1158/1535-7163.MCT-08-0201
11	4QVF	BCL-xL / Q07817	Competitive binding affinity (Ki)	Competitive fluorescence polarization assay using the Bim(86-103) 18-mer; Ki derived from competition curves.	Stronger	doi:10.1110/ps.032896.107
11	2PQK	MCL-1 / Q07820	Competitive binding affinity (Ki)	Competitive fluorescence polarization assay using the Bim(86-103) 18-mer; Ki derived from competition curves.	Weaker	doi:10.1110/ps.032896.107
12	4QVF	BCL-xL / Q07817	Competitive binding affinity (Ki)	Competitive fluorescence polarization assay using the Bim(86-103) 18-mer; Ki derived from competition curves.	Weaker	doi:10.1110/ps.032896.107
12	2PQK	MCL-1 / Q07820	Competitive binding affinity (Ki)	Competitive fluorescence polarization assay using the Bim(86-103) 18-mer; Ki derived from competition curves.	Stronger	doi:10.1110/ps.032896.107
13	4QVF	BCL-xL / Q07817	Competitive binding affinity (Ki)	Competitive fluorescence polarization assay using the Bim(86-103) 18-mer; Ki derived from competition curves.	Stronger	doi:10.1110/ps.032896.107
13	2PQK	MCL-1 / Q07820	Competitive binding affinity (Ki)	Competitive fluorescence polarization assay using the Bim(86-103) 18-mer; Ki derived from competition curves.	Weaker	doi:10.1110/ps.032896.107
14	4QVF	BCL-xL / Q07817	Competitive binding potency (IC50)	Competitive fluorescence polarization assay using a recombinant Bim BH3 peptide; IC50 obtained from dose-response fitting.	Stronger	doi:10.1002/pro.329
14	2PQK	MCL-1 / Q07820	Competitive binding potency (IC50)	Competitive fluorescence polarization assay using a recombinant Bim BH3 peptide; IC50 obtained from dose-response fitting.	Weaker	doi:10.1002/pro.329
15	2H62	BMP2 / P12643	Binding affinity (apparent KD)	SPR biosensor assay with immobilized ligand; apparent KD obtained from equilibrium binding analysis.	Weaker	doi:10.1186/1472-6807-7-6
15	1S4Y	Activin A / INHBA / P08476	Binding affinity (apparent KD)	SPR biosensor assay with immobilized ligand; apparent KD obtained from equilibrium binding analysis.	Stronger	doi:10.1186/1472-6807-7-6

Table S10 (continued)

Group	PDB ID	Binding partner	Endpoint	Measurement method	Change vs WT	Reference
18	2H62	ActRIIB extracellular domain / mouse P27040 (human-equivalent domain)	Binding affinity (apparent KD)	SPR biosensor assay with immobilized E. coli-derived BMP2; apparent KD obtained from equilibrium binding analysis.	Weaker	doi:10.1186/1472-6807-7-6
18	2GOO	ActRIIA / mouse P27038	Binding affinity (apparent KD)	SPR biosensor assay with immobilized E. coli-derived BMP2; apparent KD obtained from equilibrium binding analysis.	Stronger	doi:10.1186/1472-6807-7-6
19	2H62	ActRIIB extracellular domain / mouse P27040 (human-equivalent domain)	Binding affinity (apparent KD)	SPR biosensor assay with immobilized E. coli-derived BMP2; apparent KD obtained from equilibrium binding analysis.	Stronger	doi:10.1186/1472-6807-7-6
19	2GOO	ActRIIA / mouse P27038	Binding affinity (apparent KD)	SPR biosensor assay with immobilized E. coli-derived BMP2; apparent KD obtained from equilibrium binding analysis.	Stronger	doi:10.1186/1472-6807-7-6
20	3M7Q	Bovine cationic trypsin / P00760	Inhibitory potency (Ki)	Competitive reversible tight-binding enzyme inhibition assay; Ki determined from residual substrate hydrolysis.	Weaker	doi:10.1074/jbc.M115.647586
20	3T62	Bovine alpha-chymotrypsin / P00766	Inhibitory potency (Ki)	Competitive reversible tight-binding enzyme inhibition assay; Ki determined from residual substrate hydrolysis.	Stronger	doi:10.1074/jbc.M115.647586
21	6YE8	Spermidine / PDB CCD SPD (non-polymer)	Binding affinity (KD)	Isothermal titration calorimetry (ITC); KD obtained by fitting the binding isotherm.	Stronger	doi:10.1016/j.jbc.2021.101419
21	6YE0	Putrescine / PDB CCD PUT (non-polymer)	Binding affinity (KD)	Isothermal titration calorimetry (ITC); KD obtained by fitting the binding isotherm.	Weaker	doi:10.1016/j.jbc.2021.101419

Table note for Supplementary Table S10. Task IDs and internal mutant-PDB filenames used by the inference pipeline are omitted. PDB, UniProt, and literature mutation columns provide the residue-numbering crosswalk used during curation. “Change vs WT” denotes the experimentally reported direction of binding or inhibitory potency relative to the corresponding wild type.

G.4 PARTNER-CONTEXT ABLATION

To evaluate the dependence of model predictions on local binding-partner context, we progressively cropped the partner protein around the mutated residue while keeping the focal protein unchanged. Starting from the 20 Å-cropped complex used as the structural input, partner-chain residues were retained only if they contained atoms within a predefined distance of the mutated residue. Distance cutoffs of 3, 5, 10, 15, and 20 Å were evaluated. The 20 Å condition therefore corresponds to the complete partner context available in the original pre-cropped structural input, whereas progressively smaller cutoffs remove increasingly distal partner information while preserving the focal protein.

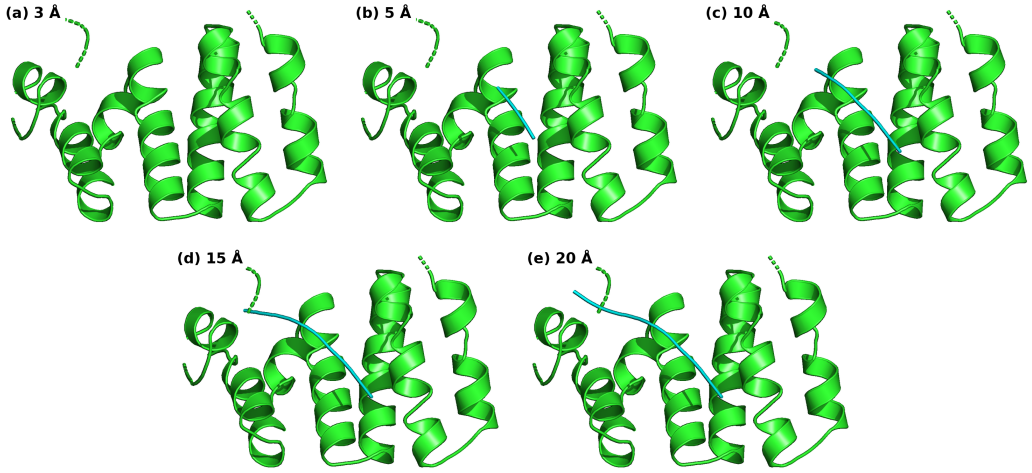


Figure S6: Progressive partner-context cropping around the mutated residue. Within the 20 Å-cropped complex used as the starting structural input, the focal protein is kept unchanged (green), whereas residues from the binding partner (cyan) are retained only when they fall within (a) 3 Å, (b) 5 Å, (c) 10 Å, (d) 15 Å, or (e) 20 Å of the mutated residue. Increasing the cutoff progressively restores the local structural context of the partner, from a minimal contact neighborhood to the complete partner context available in the 20 Å input. All panels are rendered from the same viewpoint and at the same scale to directly illustrate the effect of partner-context ablation.

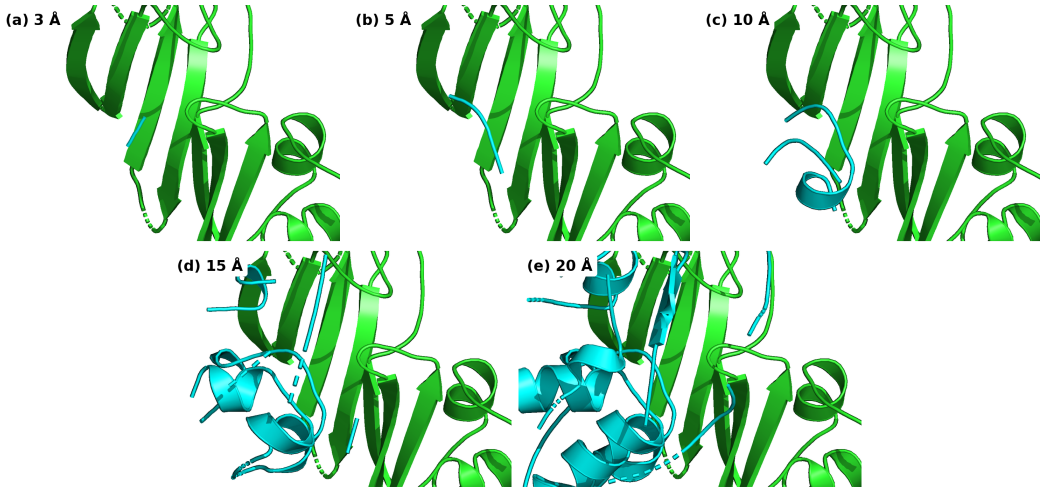


Figure S7: Second representative example of progressive partner-context cropping around the mutated residue. The focal protein is shown in green and is retained unchanged across all conditions, whereas the binding-partner context is shown in cyan and is progressively restricted according to its distance from the mutated residue: (a) 3 Å, (b) 5 Å, (c) 10 Å, (d) 15 Å, and (e) 20 Å. The series illustrates how increasing the distance cutoff progressively restores local partner structural context, from a minimal interface fragment to the full partner context available in the 20 Å pre-cropped input. All panels are rendered from the same viewpoint and at the same scale.