Accepted for publication in the proceedings of the 2026 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), CSBW 2026 workshop. © 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Plausible, Confident, and Wrong: A Preregistered Test of Mutation Response in AI-Generated Protein-Ligand Complexes

Jason L. Volk
Invariant Research, Garland, Texas, USA
jason@invariant.pro

***Abstract*—A generated protein-ligand complex can satisfy every check a modern pipeline applies to it and still get the mutation wrong. We report a preregistered, computationally blinded evaluation that keeps three questions apart: per-structure plausibility, response to a declared perturbation across conditions, and agreement with an experimental reference. It uses SATYA, a deterministic verification framework organized around that separation. For SARS-CoV-2 main protease (Mpro) with ensitrelvir, Boltz-2 generated 25 wild-type and 25 M49I complexes. PoseBusters, MolProbity and SATYA's own reference-free check advanced all 50; generator confidence advanced 25, of which 24 were wild type. After unblinding, no pocket-aligned ligand RMSD exceeded 1.01 Å, yet the cohort failed the preregistered mutation-site test: at the site where the minimum residue-49-to-ligand heavy-atom distance is 1.29 Å shorter in the mutant reference, the generated condition medians differed by 0.03 Å, and every wild-type candidate placed residue 49 closer to the ligand than the wild-type crystal structure does. Across the agreement descriptors recorded at freeze time, agreement was granularity dependent, and the preregistered residue-level endpoint was the most favorable of them. The frozen reference-free contrast rule declares a sampling floor below which it returns no classification; at five seeds per condition it was not applicable to Mpro. For KRAS G12D versus G12D/Y96S with MRTX1133 it returned RESOLVED at thirty seeds under its declared descriptor, and a post hoc analysis reversed the sign and found the response broad across the pocket. No construct-matched reference exists for the G12D arm, so this study has one retrospective target with a reference-backed failure and one prospective target with no reference-backed result. Passing per-structure plausibility checks does not, by itself, establish mutation response.**

***Index Terms*—cofolding, protein-ligand complexes, structure prediction evaluation, mutation response, computational drug discovery, deterministic verification, preregistration**

I. Introduction

A cofolding model returns a structure for every input it is given, and nothing in that output reports whether the sequence change encoded in the input made any difference. Models such as Boltz-2 generate protein-ligand complexes from sequence and ligand chemistry alone [1], [2], and a generated complex is a compound hypothesis: the protein adopts a conformation, the ligand occupies a pose, and the two form a particular set of relationships. Before such a complex informs design it is commonly screened for physical plausibility. We use *per-structure plausibility check* throughout for an automated screen that receives one structure and emits a pass-or-hold decision from it, with no experimental reference and no information about any other sequence condition; the frozen rules here emit ADVANCE or HOLD, and those names are used only for the rules themselves. PoseBusters tests chemical validity, geometry and steric overlap of a pose [3]; MolProbity validates all-atom protein geometry [4]; generator confidence estimates the model's own reliability; and where an experimental structure exists, ligand RMSD and contact recovery measure agreement with it. None was designed to establish the structural consequence of a sequence change, and the argument here is not that any fails at its own task, but that a workflow consuming their outputs must track which question each answered. Throughout, *mutation response* means a change in a declared structural descriptor between two sequence conditions, specified numerically in advance.

Recent benchmarks show these forms of evidence answer different questions. On 2225 human kinase chains, KinConf-Bench found favorable ligand-pose geometry did not reliably indicate the correct kinase conformational state, and reported mode collapse and drift toward apo-like states for holo complexes [5]. In a prospective study of several hundred Mac1-ligand complexes determined after the models' training cutoffs, AlphaFold 3, Boltz-2 and Chai-1 [6] each reproduced more than half of the ligand poses within 2 Å RMSD [7], while common ligand-associated rearrangements, including a peptide flip and a large loop opening, were not recapitulated. Perturbing cofolding inputs in chemically motivated ways produces predictions that diverge from the expected physical response, which those authors read as reliance on the training corpus rather than on the interactions themselves [8]. In each case the shortfall was invisible to the checks applied, because the structures were acceptable on their own terms.

A related question arises whenever a claim concerns a change. Resistance mutations and variant-specific binding ask whether a structural difference caused by a sequence difference has been reproduced. The question predates cofolding: single-structure prediction was reported largely insensitive to the consequences of missense substitutions [9], while other work found localized deformation between paired predictions does track experiment when the comparison is made between conditions rather than within one structure [10]. Those studies predict protein structure with no bound ligand; here the descriptor is a protein-ligand distance in a cofolded complex.

TABLE I
RELATED EVALUATION EFFORTS ON TWO AXES: WHETHER AN EXPERIMENTAL REFERENCE IS REQUIRED, AND WHETHER TWO SEQUENCE CONDITIONS ARE COMPARED WITH EACH OTHER.

Effort	What it establishes	Ref.	A vs. B
Per-structure checks [3], [4], confidence	For one structure: chemistry, geometry, clashes, and the model's own reliability estimate	no	no
State benchmarks [5], [7]	Whether the predicted protein state matches experiment	yes	no
Perturbation [8]	Whether predictions respond to perturbed inputs as physics implies	no	inputs
FoldX [11], ddG [12]	Substitution effect from a solved structure of one condition	yes	yes
This study	Plausibility, an A-versus-B contrast and reference agreement, reported separately	per layer	yes

What is asked of the model is a contrast, and a complex can be plausible, confident and globally accurate while failing that test, because the relevant information lies between the conditions rather than in any single structure. Table I places these efforts on the two axes that matter here. The consequence of a substitution is also the object of a mature class of methods, including FoldX [11] and Rosetta ddG [12], which begin from an experimental structure of one condition and estimate the effect of changing it. The question here is different: given two generated ensembles and no experimental structure for either, can the declared perturbation be shown to have produced a structural response at all. A reference-free contrast is not a substitute for those methods and is not evaluated against them.

We examine this with a preregistered, computationally blinded protocol and SATYA, a deterministic verification framework that treats three questions separately: whether a structure is admissible on its own, whether a declared perturbation produces a resolvable response, and whether a structure agrees with a matched reference. Each is answered by a different layer with its own evidence requirements, and the answers are not combined into a single score; each decision is recorded under an explicit policy together with whether the required evidence was available, which is what makes the basis and limits of each conclusion inspectable.

Our contributions are as follows. First, on Mpro with ensitrelvir, three per-structure plausibility checks, including SATYA's own, advanced all 50 generated complexes, while under the preregistered generation protocol the cohort did not reproduce the crystallographically observed residue-49 response, a test stated numerically before any candidate was scored. Second, we report every agreement descriptor recorded at freeze time rather than the preregistered endpoint alone, and show that agreement is granularity dependent. Third, we characterize why these per-structure evaluations could not establish the mutation response, from each tool's native local output. Fourth, we report the frozen reference-free contrast rule on two targets: it declined to classify Mpro because the five-seed design fell below its sampling floor, and on KRAS G12D with MRTX1133 at thirty seeds per condition it resolved a condition-associated difference under its declared descriptor, where reference agreement was not evaluable, and where the sign of the response depends on whether the descriptor includes atoms of the substituted side chain. The targets are asymmetric by construction, and the consequence is worth stating plainly: this study has one retrospective target with a reference-backed failure and one prospective target with no reference-backed result. The records behind every number reported here are deposited (Data and Code Availability).

II. VERIFICATION FRAMEWORK

A. Three layers, three questions

Layer A asks whether a single complex satisfies reference-free structural constraints. Layer B asks whether two ensembles generated under a declared sequence perturbation differ at a declared descriptor by more than a declared margin. Layer C asks whether a complex agrees with a matched experimental reference. Layers A and B consume no experimental structure; Layer C requires one. They are parallel questions with different inputs, not stages of a cascade. Each layer keeps measurements (distances, scores), decisions (thresholded outcomes under a declared, hashed policy) and availability (whether a measurement could be computed) distinct, and a measurement that cannot be computed is recorded as not computed rather than passed or failed. Layer B's status is orthogonal to the Layer A verdict and never converts an absent response into a structural failure, because some real mutations produce little structural change.

B. Determinism

We use the term operationally: for fixed candidate coordinates, reference, correspondence, policy, implementation version and execution environment, the verifier produces identical decisions and identical evidence. Policies are serialized and hashed, and the hash is recorded with every result. Where a computation uses Monte Carlo resampling, as in the Layer B null calibration, the random generator, seed and resample count are part of the declared specification. The claim concerns the verifier, not the generator, and computational repeatability is distinct from scientific validity.

C. Layer A: reference-free plausibility

Layer A (SATYA Co-Fold) evaluates a single complex against reference-free constraint channels and reports each violated constraint as a localized obstruction carrying a structural region and a kind. The channels are of two families: protein covalent geometry, comprising backbone bond lengths, backbone bond angles and peptide-bond planarity against standard stereochemical tolerances; and protein-ligand interface channels, comprising interface quality, which tests for intermolecular steric overlap, and interface sufficiency, which tests for a minimum contact density. Aggregation follows a cellular-sheaf consistency formulation over the molecular graph [14], in which local constraints on overlapping neighborhoods must agree; no claim is made here about its incremental sensitivity over classical geometric checks. The native PASS, REVIEW or FAIL verdict is computed under a hashed policy from the

aggregate channel scores and the count of critical violations, and the obstruction list is a separate localization output, not a verdict input. For cross-check comparison the preregistered mapping is PASS to ADVANCE and REVIEW or FAIL to HOLD.

D. Layer C: reference agreement

Layer C measures agreement with a matched experimental structure; the frozen specification names it "reference correctness." The comparator (version 0.3.0; policy SHA-256 prefix `e03ca4e2`) selects homodimer chain correspondence by a rule frozen before unblinding, the mapping with the lower whole-complex Cα RMSD with a declared tie-break, and records that mapping and its margin per candidate. Ligand RMSD is name-matched and computed after pocket alignment. Contacts are residue-level at a 4.0 Å heavy-atom cutoff summarized by Jaccard similarity; the comparator also records atom-pair contact Jaccard, interaction-fingerprint recall, a hydrogen-bond proxy recall, key-contact recall, the contact residues lost and gained per site, and pocket backbone and side-chain RMSD. Per-residue minimum heavy-atom distances are reported for pocket residues within 5.0 Å of the ligand in the reference; because a binary contact changes whenever a residue crosses the cutoff, continuous distances are the more stable observable. The layer establishes agreement with the selected structure, not biological correctness.

E. Layer B: perturbation consistency

Layer B follows a specification frozen before the KRAS candidates were generated. Condition A is the baseline and condition B the perturbed sequence. The primary descriptor is the minimum heavy-atom distance from the mutated residue to the ligand, per candidate and per site. Diffusion samples are summarized by their median within each seed, so the generator seed is the unit of stochastic replication, and homodimer sites are reported separately and never averaged. Samples from a seeded run are nested observations, so the reported n is the number of seeds rather than of generated models. Across seed summaries the effect is Hedges' g [13]:

$$g = J\,\frac{\bar{x}_A - \bar{x}_B}{s_p}, \quad J = 1 - \frac{3}{4(n_A + n_B) - 9}, \qquad (1)$$

with s_p the pooled within-condition standard deviation and standard error $\mathrm{SE}(g) = [(n_A + n_B)/(n_A n_B) + g^2/(2(n_A + n_B - 2))]^{1/2}$, giving a 95% interval $g \pm 1.96\,\mathrm{SE}(g)$. The status is COLLAPSED if the interval lies entirely within $(-\delta, +\delta)$, equivalence within the declared margin; RESOLVED if entirely outside it; INDETERMINATE if it spans both regimes, so that neither conclusion is supported; and NOT_APPLICABLE if the descriptor is undefined or fewer than 25 valid seeds are available in either condition. The margin $\delta = 1.0$ pooled standard deviation is a declared convention, not a derived constant. A null distribution is computed per condition by resampling seed summaries at the full ensemble size (PCG64 generator, fixed seed, 100,000 resamples); noise separation is adequate when the larger of the two per-condition 95th percentiles is below δ.

The frozen specification, which is deposited, sets the floor at 25 valid seeds per condition and fixed the KRAS design of 30 before generation. The floor's motivation is visible in $\mathrm{SE}(g)$: at $n_A = n_B = 5$ and $g = 0$ it equals $\sqrt{10/25} = 0.632$, so the interval half-width is 1.24 pooled standard deviations, already beyond $\delta = 1.0$, and COLLAPSED is unreachable at that design whatever the data. The specification's planning analysis reports definite-call probabilities of approximately 0.94 and 0.91 at 30 seeds for standardized effects of zero and twice the margin. Those probabilities are stated for declared standardized effects and depend on no measurement from either benchmark, so deriving the floor after the Mpro design was fixed does not make it circular; they are operating characteristics under the stated simulation assumptions, not independently reproduced here, and not a general sample-size guarantee. Below the floor the rule returns NOT_APPLICABLE rather than a classification. The rule as applied is stated in full above; the specification supplies the provenance of the floor and the planning rates, not the content of the rule.

III. Preregistered Protocol

A. Targets and references

The primary target is the SARS-CoV-2 main protease homodimer bound to ensitrelvir [15], wild type (PDB 8DZ0) and the M49I polymorphism (PDB 8DZ1), for which reduced ensitrelvir potency and the corresponding crystal structures have been reported [16]. In these structures the minimum residue-49-to-ligand heavy-atom distance is shorter in the mutant at both protomers: at binding site B it is 4.985 Å against 3.693 Å (−1.291 Å), creating a contact at the 4.0 Å definition, and at site A it is 5.447 Å against 4.248 Å (−1.199 Å), without crossing it. The second target is KRAS G12D with MRTX1133 [17], comparing G12D with G12D/Y96S, a substitution at a residue of the MRTX1133 binding pocket, with GDP and Mg^{2+} in the generation inputs of both arms. An experimental structure of the G12D/Y96S complex has been deposited [18].

B. Blind generation

Candidates were generated with Boltz-2 version 2.2.1 [1] without templates, pocket or contact constraints, or inference-time potentials, using hash-pinned multiple sequence alignments so that no network request occurred during production. Boltz-2 was chosen for an operational reason rather than a claim of representativeness: among current cofolding models it provides open weights and local inference code that can be run air-gapped with hash-pinned alignments, which is what this protocol's freeze discipline requires, and whether other generators behave the same way on these targets is not tested here. The blinding was computational. The references informed the preregistered hypothesis and the experimental residue-49 distances appear in the preregistration, but for both targets their coordinates were excluded from candidate generation and from every reference-free computation until the blind outputs were frozen. Blinding here concerns the evaluation, not the generator's prior exposure: 8DZ0 and 8DZ1 were released on 2022-10-05, before the 2023-06-01 structural

TABLE II

PREREGISTERED BLIND DECISION RULES. THE MODEL-CONFIDENCE THRESHOLD IS THE MEDIAN OF THE POOLED 50-CANDIDATE MPRO POPULATION, SO IT IS A RELATIVE RANKING FILTER OVER THE COHORT BEING SCORED, NOT AN ABSOLUTE CRITERION.

Check	ADVANCE if
Model confidence	score $\geq$ median of the 50 candidates
MolProbity	MolProbity score ≤ 2.0 and clashscore ≤ 10
PoseBusters	all reference-free checks pass
SATYA Layer A	native verdict PASS
Combined workflow	every check ADVANCE

TABLE III

BLIND CHECK OUTCOMES (CANDIDATES ADVANCED / SCORED). CONFIDENCE WAS RECORDED FOR ALL 300 KRAS CANDIDATES BUT CARRIED NO THRESHOLD THERE, SO NO COMBINED DECISION IS DEFINED FOR THAT BENCHMARK.

Check	Mpro WT	Mpro M49I	KRAS, both arms
Model confidence	24/25	1/25	not used
MolProbity	25/25	25/25	300/300
PoseBusters	25/25	25/25	300/300
SATYA Layer A	25/25	25/25	300/300
Combined workflow	24/25	1/25	not used

training cutoff stated in the Boltz-2 report [1], so both lie inside the training window, and what the benchmark tests is whether the generated ensembles differentiate the two sequences, not whether the generator could have seen either structure. The targets differ here: the KRAS reference was released on 2026-04-01 [18], after the same cutoff, so the KRAS arm is prospective with respect to the generator while the Mpro arm is not. For Mpro, five seeds with five diffusion samples each per condition produced 50 candidates; for KRAS, thirty seeds with five samples each per arm produced 300. Candidate populations were frozen by hash before any scoring.

C. Blind checks and decision rules

Four per-structure checks were run with references sealed, each emitting ADVANCE or HOLD under a rule fixed before scoring (Table II). PoseBusters ran in its reference-free configuration, and MolProbity through CCTBX [19] on receptor-only structures with hydrogens placed by Reduce [20]. The combined workflow decision is the conjunction of all four. Blind results were frozen and hash-verified before any reference-backed metric was computed.

D. Unblinding and endpoints

After the blind freeze, wild-type candidates were compared with 8DZ0 and M49I candidates with 8DZ1 through the frozen comparator. Residue-level contact Jaccard per site remained the preregistered primary endpoint and pocket-aligned ligand RMSD the secondary endpoint. Per-residue minimum distances were contextual measurements, used to evaluate the separately prespecified mutation-site predictions, stated numerically in advance. P1: wild-type candidates should place residue 49 at approximately 5.0 to 5.4 Å from the ligand, and M49I candidates at 3.7 to 4.2 Å. P2: the site A versus site B asymmetry should be reproduced, with the mutant gaining a residue-49 contact at site B and not at site A. P3, a prespecified anticipated failure mode: if wild-type and M49I candidates are statistically indistinguishable at residue 49, the model has not captured the mutation's structural consequence; P3 is stated here as the preregistered prediction, not as a demonstrated result. The plan prohibited new primary metrics, threshold changes, candidate exclusion and averaging of sites A and B after unblinding, so residue 49 was a prespecified test site, not a discovery. Section IV-C reports additional descriptors recorded at freeze time but not preregistered as endpoints; they are reported as recorded observables and the endpoint hierarchy above is unchanged.

IV. RESULTS

A. Every Mpro candidate passed every per-structure check

The two targets exercise different routes: Mpro was preregistered for the reference-backed route at five seeds per condition, KRAS for the reference-free contrast at thirty. PoseBusters advanced 50 of 50 candidates (100 of 100 ligand-site evaluations), MolProbity advanced 50 of 50, and SATYA Layer A returned PASS for 50 of 50 (Table III). Model confidence advanced 25 of 50, comprising 24 wild-type and one M49I candidate; because its threshold is the median of the pooled population, this split reflects a condition-level difference in confidence and provides almost no contrast within either condition. The combined workflow advanced the same 25. The preregistered binary decisions from the three structural checks contributed no discrimination.

B. Global agreement with the references was high

Whole-complex Cα RMSD did not exceed 0.78 Å for any candidate. Pocket-aligned ligand RMSD did not exceed 1.01 Å in any of the 100 site evaluations, well under the 2 Å criterion conventionally used for pose success [3]; medians were 0.30 Å (site B) and 0.34 Å (site A) for wild type and 0.94 Å and 0.63 Å for M49I, so the M49I medians exceed the wild-type medians at both sites even though every value is well inside the conventional criterion. Residue-level contact Jaccard, the preregistered primary endpoint, had medians between 0.85 and 0.95 across conditions and sites. These are continuous summaries with no preregistered pass threshold; on each of them, agreement was high.

For all 50 candidates the two homodimer chain mappings differed in Cα RMSD by less than 0.062 Å and are therefore marked ambiguous: the assignments are close to equally good, so deposited chain labels do not identify sites and every site statement has to be permutation-aware. The references are not symmetric at the test residue, which sits at 4.985 Å and 5.447 Å at the two sites of 8DZ0; this study does not investigate the origin of that asymmetry. Site-level statements below refer to the reference site selected by the frozen rule.

C. Agreement was granularity dependent

The comparator recorded several agreement descriptors beyond the preregistered endpoint, and reporting only the endpoint would present the most favorable of them. The values that follow are medians over the 25 candidates in each condition and reference site. Residue-level contact Jaccard runs 0.85 to 0.95; the same contact sets compared at atom-pair rather than residue granularity run 0.39 to 0.57, the lowest being the M49I arm at site B. Interaction-fingerprint recall runs 0.66 to 0.83 and the hydrogen-bond proxy recall 0.50 to 0.83, again lowest for M49I at site B, while key-contact recall is 1.00 in all four cells. Pocket backbone RMSD stays between 0.38 and 0.54 Å, but pocket side-chain RMSD exceeds 1.10 Å in every cell and reaches 1.37 Å for M49I at site B. These descriptors were measured at freeze time and are reported as recorded observables: none is a preregistered endpoint, the preregistration prohibits introducing a new primary metric after unblinding, and no threshold is applied to any of them. Read together they qualify the preceding subsection rather than contradict it. Agreement is high at the granularity the preregistered endpoint used, and the receptor side chains in the pocket, which is where a substitution acts, are the part that agrees least. Solvent-accessible surface area was not measured in the frozen record and is not computed or reported here.

D. The generated cohort failed the mutation-site test

At reference site B, where the minimum residue-49-to-ligand distance is 1.291 Å shorter in 8DZ1 than in 8DZ0, wild-type candidates placed residue 49 at a median of 3.833 Å and M49I candidates at 3.801 Å, a difference of 0.032 Å (Table IV, Fig. 1). At the declared replication unit the wild-type seed-level mean is 3.921 Å $[3.695, 4.148]$ and the M49I mean 3.805 Å $[3.657, 3.952]$, a difference of $+0.117$ Å $[-0.108, +0.341]$ against a pooled standard deviation of 0.154 Å. These seed-level intervals are descriptive and sit outside the preregistered primary analysis, which prohibits new inferential statistics after unblinding; they are reported because the reference contrast they are set against, 1.291 Å, lies far outside all of them. Seed-level medians ranged from 3.71 to 4.12 Å for wild type and 3.71 to 3.99 Å for M49I, and all five wild-type seed medians lie below the 4.985 Å wild-type reference distance. All 25 wild-type candidates placed residue 49 closer to the ligand than 8DZ0 does (range 3.27 to 4.43 Å), and 16 of 25 formed a residue-49 contact at the 4.0 Å definition that is absent from the wild-type crystal structure, while the M49I candidates lay near their own reference with a median 0.11 Å above 3.693 Å. Under the preregistered generation protocol the cohort placed residue 49 at a mutant-like minimum distance for both sequences and did not reproduce the crystallographically observed change. Prediction P1 failed for wild type. The small descriptive contrast is consistent with the failure mode anticipated by P3, but the five-seed analysis does not establish statistical equivalence between conditions. Failure of the mutation-site test here denotes failure of the prespecified expectation P1 for wild type together with failure to reproduce the 1.291 Å reference contrast, both at reference site B, the only site at which a wild-type residue-49 distance is recorded. The confidence check had advanced 24 of these 25 wild-type candidates, so the cohort it preferentially advanced is the one in which every candidate places residue 49 closer to the ligand than its own reference does.

TABLE IV

RESIDUE 49 TO ENSITRELVIR MINIMUM HEAVY-ATOM DISTANCE AT REFERENCE SITE B (Å). CANDIDATES: MEDIAN [RANGE], 25 PER CONDITION NESTED IN 5 SEEDS. SEEDS: MEAN [95% INTERVAL] OVER THE 5 SEED MEDIANS, DESCRIPTIVE AND OUTSIDE THE PRIMARY ANALYSIS. THE CRYSTAL DIFFERENCE IS COMPUTED UNROUNDED, SO IT DIFFERS BY 0.001 Å FROM SUBTRACTING THE ROUNDED ENTRIES.

	Crystal	Candidates	Seeds	<4.0 Å
Wild type	4.985	3.833 [3.271, 4.428]	3.921 [3.695, 4.148]	16/25
M49I	3.693	3.801 [3.103, 5.001]	3.805 [3.657, 3.952]	20/25
WT minus M49I	1.291	0.032	0.117 [−0.108, +0.341]	

The descriptor spans a substitution, so the comparisons it supports are worth separating. Each candidate-to-reference comparison is within a condition and therefore between the same residue type, so it carries no cross-residue chemistry term, and the wild-type one is where the prediction failed. The 0.032 Å candidate difference does span methionine against isoleucine, but so does the 1.291 Å crystal difference it is set against, so the chemistry enters both and the finding is the comparison between them. The comparator records the minimum distance and not the atom achieving it, so which atom drives each value is not established here.

Two coverage limits remain, one partly offset by another recorded column. Every recorded wild-type value lies below both wild-type reference distances (4.985 Å and 5.447 Å), so the discrepancy is not an artifact of comparing against the smaller; alternate-mapping sensitivity was not recomputed. In 8DZ0 residue 49 lies outside the 5.0 Å reference pocket at site A, so no wild-type site-A distance is recorded, which is why the alternate mapping could not be evaluated and why P2 is not fully evaluable as a distance comparison. The contact-change records do cover that site: residue 49 appears in the gained-contact set in 15 of 25 wild-type rows at site A, where the reference has it at 5.447 Å and so outside the contact definition, and in 16 of 25 at site B; in the M49I arm it is gained in 15 of 25 rows at site A, where the reference sits at 4.248 Å just outside the cutoff, and lost in 5 of 25 at site B. The direction is the same at both protomers in both conditions: the cohort pulls residue 49 toward the ligand relative to the matched reference. This is contact-set membership at a single cutoff, so it carries direction and not magnitude, and it does not convert P2 into an evaluable distance prediction. Finally, the 25 candidates per condition are nested within five seeds, so candidate-level counts and ranges are descriptive rather than 25 independent replicates.

E. Local information near residue 49 was present and unused

Layer A passed every candidate while recording 888 localized obstructions across the 50 (795 backbone angle, 77 peptide-bond planarity, 16 backbone bond), all in the protein

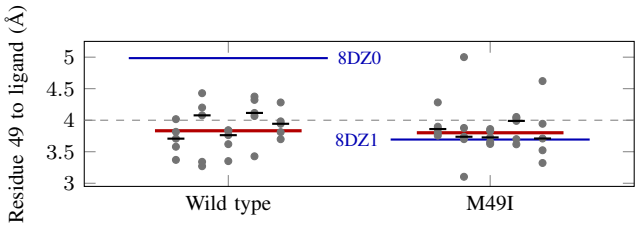


Fig. 1. Every wild-type candidate places residue 49 closer to the ligand than the wild-type crystal structure does. Points are grouped by seed; short black bars mark seed medians, red bars condition medians, blue lines the 8DZ0 and 8DZ1 crystal values, and the dashed line the 4.0 Å cutoff. The 25 points per condition are nested within five seeds.

region and none concerning a protein-ligand relationship; its interface-quality and interface-sufficiency channels were 1.0 throughout. Because the verdict is a function of the aggregate channel scores and the critical-violation count, while the obstruction list is a separate localization output, every candidate's aggregates fell inside the policy's PASS region. A PASS beside a nonzero obstruction count is the declared policy behaving as constructed rather than a failure suppressed at reporting time, and retaining localization for passing structures is what lets an advanced candidate be examined afterwards at a named residue. MolProbity, on receptor-only structures, retained clash records involving residue 49 in 49 of 50 candidates, 25 of 25 M49I and 24 of 25 wild type; at most 40 were retained per receptor, so the count is descriptive, the records are protein-internal rather than protein-ligand, and the counts are similar across conditions, establishing no condition-specific difference or cause. The preregistered MolProbity rule operates on global scores and advanced every candidate. Local information near residue 49 was therefore present in at least one tool's native output, but neither a clash nor a satisfied interface constraint establishes the observed position of residue 49 for a given sequence. Layer A's interface channels test whether a sufficient and physically reasonable interface exists; without a reference they cannot test whether a residue lies at the experimentally observed distance for the sequence modeled. In this benchmark the failure is the reference-backed one defined in Section IV-D, not a demonstrated equivalence between the two generated ensembles.

F. The reference-free rule declined on Mpro and resolved on KRAS

Layer B addresses the contrast directly, without a reference (Table V). For Mpro, with five seeds per condition, the frozen rule returns NOT_APPLICABLE: the design falls below the floor and no production classification is made. The five-seed design was preregistered for Mpro's reference-backed endpoints, and Mpro is the development case from which Layer B was designed, the floor being derived afterwards from that rule's operating characteristics, so the Mpro sampling was not a choice made against a floor that existed at the time. A development-only diagnostic that bypasses the floor gives $g = +0.52$ $[-0.75, +1.78]$, spanning both regimes; it is reported for transparency and is not load-bearing, since the production status is NOT_APPLICABLE either way and the Mpro failure is established by the reference-backed comparison above. Its M49I summaries are medians over both sites rather than per site; site-B recomputation of both arms gives $g = +0.68$ $[-0.60, +1.97]$, the same classification.

KRAS demonstrates prospective application of the frozen rule; its two-arm reference-backed endpoint was not fully evaluable. With thirty seeds per arm the frozen descriptor, residue 96 to MRTX1133 minimum heavy-atom distance, resolved: the baseline-minus-perturbed shift was -0.448 Å, $g = -10.52$ $[-12.50, -8.54]$, and noise separation was adequate. The pooled standard deviation of the seed medians is 0.042 Å, so a standardized effect beyond 10 reflects very small between-seed dispersion rather than a large physical displacement; the raw shift is accordingly the biophysical observable and g the decision statistic the rule thresholds, and both are in Table V so neither is read as the other. The seed is the replication unit because diffusion samples from one seed are not independent draws of the generator's response, and that choice was fixed before generation rather than selected once the dispersion was visible. PoseBusters, MolProbity and Layer A again advanced every candidate. A post hoc development analysis showed the sign of this response depends on the descriptor: in the G12D arm the nearest residue-96 atom is predominantly the tyrosine hydroxyl, which has no counterpart in serine, and restricting the descriptor to atoms present in both residues (N, Cα, C, O, Cβ) gives $+0.288$ Å and $g = +7.11$ $[+5.72, +8.50]$, the opposite sign. Residue 96 ranked first of 24 pocket residues under this descriptor, but 12 of the other 23 also exceeded the margin, so the response is broad rather than confined to the mutation site. That breadth is a statement about the margin rather than about displacement: the 12 control residues clear it on raw shifts of at most 0.095 Å, against per-residue pooled standard deviations between 0.016 and 0.083 Å, while residue 96 shifts 0.288 Å, about three times the largest control shift. Whether those small collective shifts reflect a common motion of the ligand is not determined here; that would require comparing ligand placement between the arms, which the blind design did not include. Layer B therefore establishes a condition-associated structural response under its declared descriptor; it does not by itself establish reference agreement or biological correctness.

The reason the two-arm endpoint is not evaluable is recorded in the preregistration. An audit of all 19 deposited structures containing MRTX1133 found no construct-matched reference for the G12D arm: every deposit carrying G12D also carries either an engineered substitution absent from the perturbed reference or a second mutation, and the wild-type entries carry no G12D at all. That arm is marked reference-incompatible upstream of the evaluator and is never scored, while the perturbed arm's reference matches its generated core exactly. The loss is permanent rather than a temporary gap, and any reference-backed result from this benchmark is a single-arm result. This manuscript reports the Layer B result for KRAS and does not claim completion of the broader preregis-

TABLE V
LAYER B CONTRAST AT THE MUTATED RESIDUE (MPRO RESIDUE 49, KRAS RESIDUE 96; SEED LEVEL, NO REFERENCE). SHIFT IS THE BASELINE-MINUS-PERTURBED DIFFERENCE OF SEED SUMMARIES; g IS THE SAME CONTRAST IN POOLED STANDARD DEVIATIONS. THE MPRO DEVELOPMENT-ONLY ROW IS NOT LOAD-BEARING.

Analysis	Shift (Å)	g [95% CI]	Status
Mpro, 5 seeds, frozen rule	not computed		NOT_APPLICABLE
Mpro, 5 seeds, development only	$+0.092$	$+0.52$ [-0.75, $+1.78$]	diagnostic
KRAS, 30 seeds, frozen descriptor	-0.448	-10.52 [-12.50, -8.54]	RESOLVED
KRAS, 30 seeds, common atoms	$+0.288$	$+7.11$ [$+5.72$, $+8.50$]	post hoc

tered endpoint set; the absence of a compatible reference is an availability limitation rather than a failed endpoint result, and a result in one evaluable arm would not establish two-arm confirmation. A separate evidence extractor, written for this study and not an external replication, recomputed every Mpro blind decision and its hash from each check's native output, verified ten freeze-linked source digests, and reproduced the KRAS seed-level arithmetic from the 300 recorded values; it did not rerun the structural engines or the null calibration. The seed-level statistics and recorded descriptors in Sections IV-C and IV-D were measured from the frozen per-site record by a standard-library producer that asserts the record's shape and fails closed otherwise, and that producer, its input digest and its output record are in the deposit.

V. DISCUSSION AND LIMITATIONS

The Mpro case separates three properties that are often treated as one. Every candidate passed every per-structure plausibility check, every global agreement summary was favorable including the preregistered primary endpoint, and yet the candidates were discordant with the references at the one residue whose behavior defines the mutation, in the same direction at both protomers. The discordance was silent in a specific sense: every check returned its most favorable outcome, so nothing in the workflow's own output signalled that the mutation response was absent, and a pipeline consuming only that output would have advanced the cohort. The confidence check, the only one that discriminated at all, preferentially advanced the wild-type arm, which is the arm in which every candidate misplaces the test residue. Plausibility, global agreement and mutation response are different properties, and evidence for the first two is not sufficient evidence for the third, consistent with independent reports that ligand-pose accuracy and the observed protein state can come apart [5], [7]. Reporting the full recorded descriptor set sharpens the point from another direction: the descriptor the preregistration named as primary was the most favorable one available, and the pocket side chains, where a substitution acts, were the least.

For practice, conclusions that depend on a sequence change should be supported by evidence about the change itself. Where references exist for both conditions that means site-resolved comparison at residues the change is expected to affect, reported as continuous distances, and reporting agreement at more than one granularity so the choice of endpoint is visible. Where no reference exists it means generating both conditions with enough independent seeds for a contrast rule to classify; five per condition fell below the frozen floor here, while 30 satisfied it and returned RESOLVED under the declared descriptor on KRAS. A descriptor restricted to atoms present in both residues isolates displacement; one spanning the whole side chain measures a different quantity that includes the substitution's chemistry. On the evidence here the common-atom descriptor is the better default whenever the substitution changes atom composition. Recording both is what made the reversal visible at all, and a pipeline reporting a single contrast with the descriptor left implicit gives a reader no way to recover that dependence.

The claim is one of insufficiency, and it rests on what a per-structure check receives rather than on which generator produced the cohort. A check handed one structure, no reference, no condition-specific expected state and no information about the other condition has no representation of a cross-condition expectation; the limitation is a property of the evidence available to that class of check and applies to any workflow that screens on plausibility alone. What this study adds empirically is one cohort that passes all three evaluated checks while failing the reference-backed mutation-site test.

The limits are clear, and the targets are asymmetric by construction: one retrospective target with a reference-backed failure, and one prospective target with no reference-backed result. The study covers two targets, one generator and one protocol, so it does not characterize the generator in general, and whether AlphaFold 3 or Chai-1 behave the same way on Mpro is untested. The blinding was computational, and both Mpro references predate the generator's stated training cutoff, so the residue-49 result is not evidence that this complex is beyond the generator's reach. The benchmark shows that the ensembles did not reproduce the observed wild-type-to-M49I separation at residue 49; it does not determine why. An attracting-state account of the kind reported for kinase conformational states [5] is one candidate, but one distance at one site cannot distinguish it from other explanations, and separating them would need the conformational distribution of both ensembles, which the blind design did not record. Residue 49 was prespecified, so this evaluates a declared site rather than demonstrating unprompted localization. The mapping, coverage and nesting limits stated in Section IV-D apply here too, and a 30-seed Mpro extension would have to be reported as a post-unblinding extension with its own status rather than folded into the preregistered result. The descriptors in Section IV-C were recorded but not preregistered, and solvent-accessible surface area was never measured; recording which atom achieves each minimum distance, and adding a solvent-accessible surface descriptor, belong to the next frozen specification. The crystal structures represent single states with their own uncertainty, the KRAS sign reversal and breadth

come from post hoc development analysis, and KRAS reference agreement was not evaluable. SATYA's own reference-free check did not discriminate, and these experiments do not test whether any component of its constraint machinery adds information beyond simpler measurements.

VI. Conclusion

In a preregistered, computationally blinded benchmark, three per-structure plausibility checks advanced every generated Mpro-ensitrelvir complex and global agreement with the references was high, while under the preregistered generation protocol the cohort did not reproduce the crystallographically observed effect of the M49I substitution at residue 49. A per-structure check, without condition-specific expectations or cross-condition context, does not carry the information such a response would require. The two routes demonstrated here are not interchangeable and were not both available on either target: site-resolved reference comparison exposed the prespecified discrepancy on Mpro, where the reference-free rule returned NOT_APPLICABLE at five seeds, while reference-free ensemble comparison resolved a condition-associated difference on KRAS at thirty seeds under its declared descriptor, where a post hoc analysis reversed the sign and found the response broad across the pocket and where agreement with the experimental response was not evaluated. The discrepancy was not detected by any check in the workflow; it was detected only by a comparison the workflow lacked. Passing per-structure plausibility checks does not, by itself, establish mutation response.

Data and Code Availability

The analysis is bound to frozen records identified by SHA-256 digests, including the Mpro candidate population, the preregistration and its amendments, the blind results, the reference comparator policy, the frozen verification specification, the unblinded comparator outputs, the per-candidate and per-seed measurement tables for both targets, the KRAS Layer B record, and the evidence extraction and measurement scripts with their source snapshots. These records are deposited as a public archive at doi:10.5281/zenodo.24168325, https://doi.org/10.5281/zenodo.24168325. The archive carries a manifest listing every file with its digest and origin path, and a verifier that recomputes every digest using only the Python standard library. Components that are not deposited are named in the manifest with the reason. Information about SATYA is available at https://invariant.pro.

Acknowledgment

Anthropic Claude and OpenAI ChatGPT were used to assist with analysis scripting, figure-generation code, manuscript drafting and revision, and independent consistency checking. The author reviewed and verified generated text, code, citations, and reported values against the frozen source records and primary literature and assumes responsibility for the final content.

Declarations

Competing interests: The author is the founder of Invariant Research and the developer of SATYA, the system evaluated in this study. This is an author-developed method evaluated by its developer, and no claim of comparative advantage is made anywhere in this work.

Funding: No external funding was received for this work.

References

[1] S. Passaro *et al.*, "Boltz-2: Towards accurate and efficient binding affinity prediction," bioRxiv, 2025, doi:10.1101/2025.06.14.659707.

[2] J. Abramson *et al.*, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, vol. 630, pp. 493–500, 2024.

[3] M. Buttenschoen, G. M. Morris, and C. M. Deane, "PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences," *Chem. Sci.*, vol. 15, pp. 3130–3139, 2024.

[4] C. J. Williams *et al.*, "MolProbity: More and better reference data for improved all-atom structure validation," *Protein Sci.*, vol. 27, pp. 293–315, 2018.

[5] K. Sun and T. Head-Gordon, "A curated benchmark for cofolding models on kinase conformational states," *npj Drug Discov.*, vol. 3, art. 33, 2026.

[6] Chai Discovery Team, "Chai-1: Decoding the molecular interactions of life," bioRxiv, 2024, doi:10.1101/2024.10.10.615955.

[7] J. Kim *et al.*, "Large scale prospective evaluation of co-folding across 557 Mac1-ligand complexes and three virtual screens," *eLife*, vol. 15, reviewed preprint RP110475, 2026, doi:10.7554/eLife.110475.1.

[8] M. R. Masters, A. H. Mahmoud, and M. A. Lill, "Investigating whether deep learning models for co-folding learn the physics of protein-ligand interactions," *Nat. Commun.*, vol. 16, art. 8854, 2025.

[9] M. A. Pak *et al.*, "Using AlphaFold to predict the impact of single mutations on protein stability and function," *PLOS ONE*, vol. 18, no. 3, art. e0282689, 2023.

[10] J. M. McBride, K. Polev, A. Abdirasulov, V. Reinharz, B. A. Grzybowski, and T. Tlusty, "AlphaFold2 can predict single-mutation effects," *Phys. Rev. Lett.*, vol. 131, no. 21, art. 218401, 2023.

[11] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, "The FoldX web server: an online force field," *Nucleic Acids Res.*, vol. 33, no. Web Server issue, pp. W382–W388, 2005.

[12] E. H. Kellogg, A. Leaver-Fay, and D. Baker, "Role of conformational sampling in computing mutation-induced changes in protein structure and stability," *Proteins*, vol. 79, no. 3, pp. 830–838, 2011.

[13] L. V. Hedges, "Distribution theory for Glass's estimator of effect size and related estimators," *J. Educ. Stat.*, vol. 6, pp. 107–128, 1981.

[14] J. Hansen and R. Ghrist, "Toward a spectral theory of cellular sheaves," *J. Appl. Comput. Topol.*, vol. 3, pp. 315–358, 2019.

[15] Y. Unoh *et al.*, "Discovery of S-217622, a noncovalent oral SARS-CoV-2 3CL protease inhibitor clinical candidate for treating COVID-19," *J. Med. Chem.*, vol. 65, no. 9, pp. 6499–6512, 2022.

[16] G. D. Noske *et al.*, "Structural basis of nirmatrelvir and ensitrelvir activity against naturally occurring polymorphisms of the SARS-CoV-2 main protease," *J. Biol. Chem.*, vol. 299, no. 3, art. 103004, 2023.

[17] X. Wang *et al.*, "Identification of MRTX1133, a noncovalent, potent, and selective $KRAS^{G12D}$ inhibitor," *J. Med. Chem.*, vol. 65, no. 4, pp. 3123–3133, 2022.

[18] Protein Data Bank entry 9U8L, KRAS G12D/Y96S in complex with MRTX1133, GDP and Mg^{2+}, doi:10.2210/pdb9U8L/pdb.

[19] R. W. Grosse-Kunstleve, N. K. Sauter, N. W. Moriarty, and P. D. Adams, "The Computational Crystallography Toolbox," *J. Appl. Cryst.*, vol. 35, pp. 126–136, 2002.

[20] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson, "Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation," *J. Mol. Biol.*, vol. 285, pp. 1735–1747, 1999.