CATGEOBENCH: MECHANISM-GUIDED CATALYTIC GEOMETRY FOR ENZYME–LIGAND CO-FOLDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing benchmarks for protein–ligand structure prediction and docking rely on crystal structures and evaluate binding geometry and physical plausibility, but do not assess whether predicted enzyme complexes satisfy the geometric prerequisites of catalysis. We present CatGeoBench, a benchmark for catalytic geometric compatibility based on mechanisms from M-CSA. It includes a core set built directly from M-CSA and an extended set that covers related enzymes with alternative substrates. For each mechanism we build electron-flow chains and extract reacting atom pairs as the intermolecular bond-forming atom pairs collected up to the first intermolecular structure-changing electron-flow event. CatGeoBench scores predictions with mechanism-derived distance criteria based on scaled van der Waals radii of these pairs, rather than by comparison to crystal ligand poses. The core set comprises 239 entries with 320 reacting atom pairs, and the extended set comprises 123 entries with 161 pairs. Each entry provides a curated enzyme functional unit, validated substrate and cofactor inputs, explicit atom mappings for downstream structure evaluation, and the associated distance thresholds. We evaluate representative co-folding models under a unified protocol. Models achieve high success under the distance-only criterion on the core set but degrade on the extended set, where core performance does not reliably indicate success on related enzymes with alternative substrates.

1 INTRODUCTION

Existing protein–ligand structure-prediction and docking benchmarks primarily evaluate geometric accuracy and physicochemical plausibility, while standardized evaluation of mechanism-specific catalytic geometry in enzyme–substrate–cofactor complexes remains limited. Benchmarks such as PoseBusters (Buttenschoen et al., 2024), PoseBench (Morehead et al., 2026), and PoseX (Jiang et al., 2026) primarily assess the geometric accuracy and physical plausibility of binding poses, but do not specifically evaluate whether predicted enzyme complexes with substrates and cofactors satisfy atomistic geometric requirements of catalysis. Building a benchmark for this purpose raises two central challenges.

First, experimental structures that simultaneously contain substrates and cofactors consistent with the catalytic mechanism remain scarce. Although RCSB PDB (Berman et al., 2002) contains a large number of enzyme-related entries, many are apo structures, catalytic-residue mutants, or complexes containing substrate analogues, products, inhibitors, or cofactors alone. Recent resources such as SKiD (Krishnan et al., 2025) and EnzyBind (Song et al., 2026) provide large enzyme–ligand collections for kinetic modeling and enzyme design. However, many entries are computationally posed complexes or experimental structures that lack cognate substrates or cofactors, and therefore are not directly suitable for evaluating mechanism-specific catalytic geometry.

Second, catalysis involves multistep chemical transformations and conformational changes that go beyond generic ligand binding. A catalytic cycle typically comprises substrate binding, chemical transformation through intermediates, and dynamical rearrangement of the protein conformational ensemble. A predicted enzyme complex therefore cannot be evaluated using conventional pose-reproduction metrics alone. In particular, a unique experimentally resolved reactive ligand pose is often unavailable, making ligand RMSD ill-suited to evaluating mechanism-specific catalytic geometry. Classical enzymology instead examines whether reacting atoms adopt geometries associated

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.

with the near-attack conformation (NAC) (Bruice & Lightstone, 1999; Bruice, 2002; Zhang et al., 2024) and relates the frequency of such geometries in molecular dynamics to catalytic activity. Recent AI-driven enzyme design pipelines (Lauko et al., 2025; Butcher et al., 2025) likewise impose mechanism-based geometric filters, including active-site preorganization and distances relevant to hydrogen bonding and nucleophilic attack. Tools such as PLACER (Anishchenko et al., 2025) further support rapid ensemble assessment of intermediate-state catalytic geometry. These approaches rely on system-specific mechanistic knowledge and have largely remained case-specific, without a reusable and systematic evaluation resource.

We address this gap with CatGeoBench through the following design choices:

- **Mechanism-derived geometric criteria.** Guided by electron-flow annotations in M-CSA (Ribeiro et al., 2018), we construct electron-flow chains and extract reacting atom pairs from the initial state up to the first intermolecular structure-changing event. We then define distance thresholds from chemically motivated priors to evaluate catalytic geometric compatibility.
- **Curated catalytic initial states.** Catalytic residues are validated; substrate and cofactor identities and counts are cross-checked against the overall reaction and stepwise mechanism, ensuring that each benchmark input represents a structurally and mechanistically consistent catalytic initial complex.
- **Threshold validation and model evaluation.** We assess the resulting geometric criterion on crystal complexes with reliably mapped reacting atom pairs and apply it to representative co-folding models under a unified evaluation protocol.
- **Extensibility.** The benchmark is extended to related enzymes with conserved catalytic chemistry and matched reaction-center architectures but alternative substrates, enabling evaluation of generalization beyond the core set.

2 Related Work

Protein–ligand docking benchmarks. Recent benchmarks for protein–ligand structure prediction and docking emphasize geometric accuracy and physical plausibility. PoseBusters added chemical and steric checks and showed that many deep learning docking methods yield poses that fail basic physical criteria (Buttenschoen et al., 2024). PLINDER provided a large-scale, leakage-aware resource for protein–ligand structure prediction with standardized datasets and evaluation protocols, enabling systematic assessment of modern structure-prediction and docking methods (Durairaj et al., 2024). PoseBench extended evaluation to apo-to-holo and multi-ligand settings and compared co-folding models (AlphaFold 3, Boltz) with classical and learning-based docking (Morehead et al., 2026). PoseX studied self- versus cross-docking and the role of energy minimization (Jiang et al., 2026). These benchmarks primarily target binding geometry and physical realism and are not specifically designed to evaluate mechanism-derived catalytic geometry in enzyme–substrate–cofactor systems.

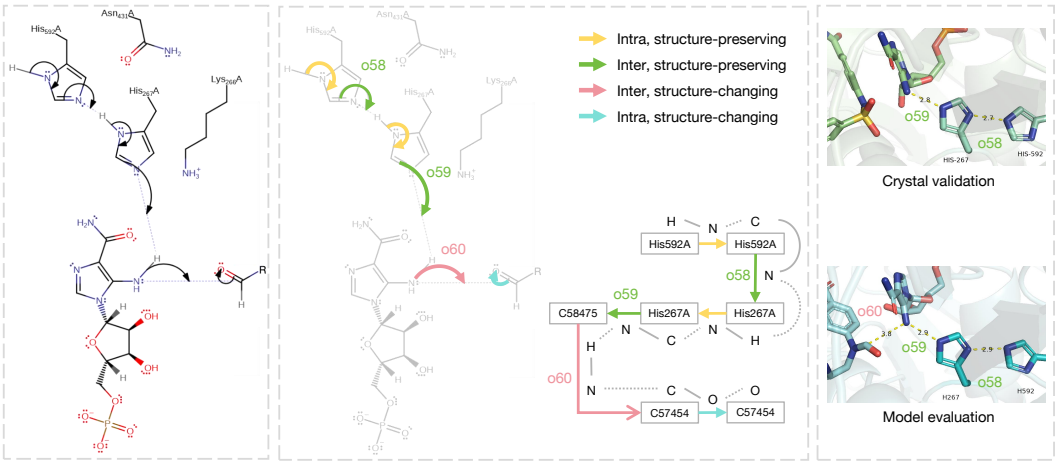
Enzyme–ligand resources and benchmarks. SKiD links kinetic parameters from BRENDA to three-dimensional structures for structure–kinetics analysis, while EnzyControl introduces EnzyBind, an enzyme–ligand collection for substrate-specific enzyme design (Krishnan et al., 2025; Song et al., 2026). These collections are valuable for training and downstream applications, yet many complexes are docked or modeled, ligands need not be the cognate substrates implied by a catalytic mechanism, and cofactors are often incomplete. They are assembled for association, design, or learning rather than as standardized tests of catalytic geometric compatibility. The Atomic Motif Enzyme (AME) benchmark (Ahern et al., 2026) is built from M-CSA active sites for which the experimental structure contains the relevant reactants, yielding 41 atomic catalytic motif cases. It evaluates whether generative methods can scaffold these motifs without prespecified residue indices. The task is active-site scaffolding on the design side; success is measured mainly by how accurately the motif atoms are recapitulated in the designed structure (with additional filters such as ligand clashes), not by scoring co-folded enzyme–substrate–cofactor complexes against mechanism-derived reacting-atom distances. Taken together, neither these enzyme–ligand resources nor AME provide a standardized test of catalytic geometric compatibility for predicted enzyme complexes.

Mechanism annotation and NAC geometry. M-CSA is a curated database of enzyme mechanisms with catalytic residues, cofactors, and stepwise electron flow (Ribeiro et al., 2018). Near-attack conformations (NACs) describe ground-state geometries in which reacting atoms approach van der Waals contact in a transition-state-like orientation (Bruice & Lightstone, 1999; Bruice, 2002).

Existing structure-prediction benchmarks do not evaluate mechanism-derived catalytic geometry. CatGeoBench bridges this gap by extracting reacting atom pairs from M-CSA mechanisms and applying chemically informed distance criteria based on NAC geometry and scaled van der Waals radii to assess catalytic geometric compatibility.

3 Dataset

CatGeoBench is derived from M-CSA and comprises a core set and an extended set. For each core entry, the structure-prediction input consists of the sequence of a catalytic functional unit together with SMILES or CCD representations of substrates and cofactors. Using the curated electron-flow description of the mechanism, we identify reacting atom pairs and assign chemically informed distance cutoffs to test whether a predicted complex is geometrically compatible with that step. Dataset assembly follows several design principles; after filtering, the **core set** contains 239 entries and 320 reacting atom pairs, and the **extended set** contains 123 entries and 161 reacting atom pairs. Sections 3.1–3.5 describe the main stages of core construction; Section 3.6 describes the extended set, which applies the same scoring protocol to related enzymes with alternative substrates. Detailed principles and processing statistics are provided in Appendix A.



(a) Mechanism from M-CSA (b) Building Electron-flow Chains and Reacting Atom Pairs (c) Validation and Evaluation

Figure 1: Overview of CatGeoBench. (a) M-CSA mechanism with curly-arrow electron flow. (b) Electron-flow chain with step classification. Solid and dashed segments mark bond cleavage and bond formation (or electron-pair localization, if a dashed loop lies on one atom). Geometric checks use bond-forming atoms of electron flows o58–o60, where o60 is the stopping point. Yellow arrows denote aromatic resonance and are treated as structure-preserving. (c) Geometric checks on crystal structures and co-folding model predictions.

3.1 Building Electron-flow Chains and Reacting Atom Pairs

We retrieve entry records and stepwise MRV files from the M-CSA API. MRV files encode 2D molecular structures together with curly-arrow electron flow for each mechanistic step. For the first step of each mechanism, we construct electron-flow chains as ordered sequences of flow events, linking consecutive arrows when the destination (to) atom set of one arrow intersects the source (from) atom set of the next. Most first steps yield linear chains, while a small number form cycles or chain-cycle combinations (see Appendix A.3).

Each flow event is classified by its intermolecular/intramolecular nature and whether it changes structure (heavy-atom connectivity or bond order), resulting in four classes: (i) intermolecular and

structure-changing; (ii) intermolecular and structure-preserving; (iii) intramolecular and structure-changing; and (iv) intramolecular and structure-preserving. For linear chains, we traverse events from the start and stop at the first intermolecular structure-changing event (class (i)). All intermolecular bond-forming atom pairs between the start and this event (inclusive) are retained as **reacting atom pairs** for geometric evaluation. Figure 1 illustrates the procedure.

Cycles are traversed completely, and chain-cycle cases are processed by considering both components. If the first mechanism step contains no structure-changing event, we construct chains for subsequent steps and repeat the search.

3.2 Mapping of Reacting Atoms

Molecular graphs are reconstructed from atom and bond records in the MRV files. Reacting atoms are annotated such that each electron-flow step is associated with reacting atom indices and their parent residue or ligand. For residue atoms, the local chemical graph is matched to standard residue templates to assign PDB atom names. For substrate and cofactor atoms, the mechanism graph is matched to the corresponding ligands in co-folding model predictions to transfer reacting atom indices. These mappings locate reacting atom pairs in predicted structures for downstream geometric evaluation.

3.3 Distance Thresholds

For heavy-atom bond formation, we use the sum of van der Waals radii as a physically motivated reference scale for NAC-like pre-reaction proximity, following the near-attack conformation (NAC) perspective (Zhang et al., 2024). Since our evaluation targets individual predicted structures rather than conformational ensembles sampled by molecular dynamics, we adopt a deliberately permissive tolerance to accommodate structural uncertainty while retaining mechanistic relevance. Specifically, we define the threshold as $d \leq 1.4(r_i + r_j)$, where r_i and r_j are the van der Waals radii of the reacting atoms. This criterion provides a catalytic geometric compatibility assessment rather than a sufficient condition for catalysis: structures exceeding this upper bound are less likely to represent physically plausible pre-reaction arrangements. The criterion is further assessed on experimentally determined crystal complexes with reliably mapped reacting atoms (Section 4.3).

3.4 Validation of Residues, Substrates, and Cofactors

Catalytic residue indices and chain assignments in M-CSA may be inconsistent across numbering schemes. We cross-check these annotations and correct them to ensure that each catalytic residue maps to the correct chain and sequence position. Protein sequences are obtained from the M-CSA reference structures, with engineered mutations reverted to wild type. Modified residues are retained only when they are specified as catalytic in the mechanism; other modifications are reverted.

Substrate and cofactor identities and counts from the M-CSA API are reconciled with the overall reaction and stepwise mechanism to define the catalytic initial complex. Species present in the first mechanistic step are required to match the reaction annotation, while molecules appearing only in later steps are excluded. Remaining discrepancies are manually resolved. When the mechanism-derived molecular graph differs from the associated ChEBI SMILES, the structure is manually verified and harmonized with the mechanism representation.

3.5 Active-site multiplicity

Many reference assemblies contain multiple equivalent catalytic sites. Supplying only a single copy of each substrate or cofactor may cause a co-folding model to distribute ligands across different sites, leading to failed reacting-atom distance checks unrelated to the quality of individual catalytic sites and thereby underestimating performance. We therefore identify the number of equivalent catalytic sites in the reference structure by matching the mechanism-defined catalytic residue set based on residue identity and local geometry. Substrate and cofactor copy numbers are scaled accordingly so that each catalytic site is paired with a complete ligand set.

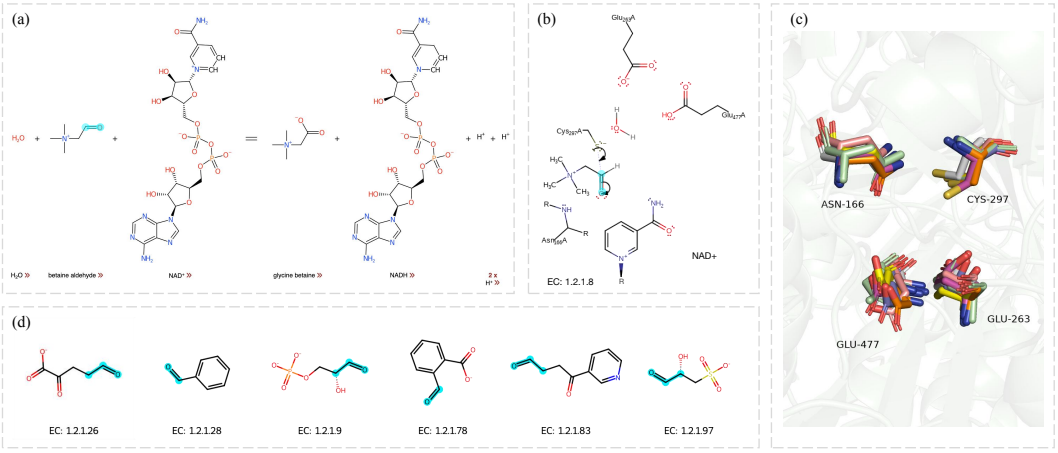


Figure 2: Illustration of core-to-extended transfer for M-CSA entry 100 (M100). (a) Overall reaction of the core entry M100. The reaction center (cyan) is defined as the radius-1 local chemical environment around the reacting atoms. (b) First mechanistic step of M100, showing catalytic residues (ASN-166, GLU-263, CYS-297, GLU-477), the substrate betaine aldehyde, and the cofactor NAD^+. (c) Superposition of catalytic residues from representative enzymes retrieved by a Folddisco (Kim et al., 2026) search using a catalytic-residue query derived from the M100 active site. (d) Representative substrates from extended entries linked to M100. Cyan highlights the matched reaction center together with the substructure obtained after element-compatible growth from the seed match.

3.6 Extended set

The core set links M-CSA mechanisms to reacting atom pairs, enabling direct evaluation of co-folding predictions against catalytic distance criteria. However, each core entry is associated with an experimental protein structure, and most include a substrate, cofactor, or close analogue complex. To capture related enzymes acting on alternative substrates, including cases without experimental holo structures, we construct an extended set that applies the same reacting-atom distance protocol to enzymes whose catalytic chemistry and active-site geometry can be aligned to a core entry. Each extended entry follows the core input format, with matched ligand counts and explicit reacting-atom mappings.

The extended set is constructed to preserve the applicability of distance criteria defined for core entries. Candidate proteins are required to share the first three EC digits with the core entry while differing in the full EC number, and their catalytic residues must match the core entry in residue identity and spatial arrangement. This ensures that residue-side reacting atoms remain interpretable under the same geometric evaluation. Swiss-Prot enzymes are linked to Rhea reaction SMILES and EC annotations and filtered by the EC criterion; catalytic-residue queries from each core entry are then used to identify compatible active-site matches by Folddisco (Kim et al., 2026). Hits are retained when residue identities agree and local active-site geometry is sufficiently similar.

Ligand correspondence is established at reacting centers rather than whole-molecule similarity. For each reacting atom in the core entry, a local chemical environment is used as a substructure query against candidate reactants. A candidate center is accepted only when it matches the reference atom in element, degree, hybridization, hydrogen count, aromaticity, and ring membership. For multiple reacting atoms, jointly consistent assignments are ranked by the extension of the seed match into neighboring graph regions under element-compatible growth. Only assignments recovering all required reacting atoms are retained; reverse reactions are considered when forward reactants do not yield a valid map. Non-reacting substrates and cofactors are assigned by fingerprint similarity to candidate reaction reactants, preserving core ligand counts.

This procedure yields 123 extended entries with 161 reacting atom pairs and expands evaluation beyond the original enzyme–substrate pairs while maintaining the same mechanism-derived geometric scoring protocol. The distribution of associated extended enzymes across core entries is shown in Appendix Fig. 7.

4 EXPERIMENTS

We conduct three evaluations: (i) consistency of the distance thresholds with experimental geometry; (ii) co-folding success on the core set; and (iii) matched core–extended comparison under the same geometric protocol.

4.1 EVALUATION METRICS

We evaluate predicted structures using mechanism-derived geometric criteria and ligand physical plausibility checks.

Distance. For each reacting atom pair in the benchmark, we measure the distance between the mapped atoms in the prediction and compare it with the cutoff $1.4(r_i + r_j)$, where r_i and r_j are the van der Waals radii of the reacting atoms. A reacting pair is considered satisfied when the predicted distance does not exceed this threshold, indicating compatibility with the mechanism-derived geometric criterion for the annotated reaction step.

PB-Valid. The physicochemical validity and structural plausibility of predicted ligand poses are assessed using the PoseBusters test suite (PB-Valid), which evaluates standard chemical and steric consistency criteria.

Success rate. We report success at both the pair and complex levels. **Pair success rate** is the fraction of reacting atom pairs satisfying the distance criterion, with an optional stricter variant requiring the associated ligand pose to also pass PB-Valid. **Complex success rate** is the fraction of entries for which all reacting atom pairs satisfy the distance criterion; a stricter variant additionally requires PB-Valid for complexes containing ligands.

4.2 EXPERIMENTAL SETUP

We evaluate Chai-1 (team et al., 2024), Boltz-1/1x/2/2x (Wohlwend et al., 2025; Passaro et al., 2025), AlphaFold 3 (Abramson et al., 2024), and Protenix-v1/v2 (Team et al., 2026; Zhang et al., 2026) using their default inference settings. All methods receive the same inputs: curated functional-unit sequences and substrate/cofactor SMILES or CCD codes, without templates or pocket constraints.

For each entry and each random seed, we generate five independent samples. Top-1 evaluates the highest-ranked sample, whereas Top-5 considers a pair or complex successful if at least one of the five samples satisfies the evaluation criterion. We report the average success rates across three random seeds.

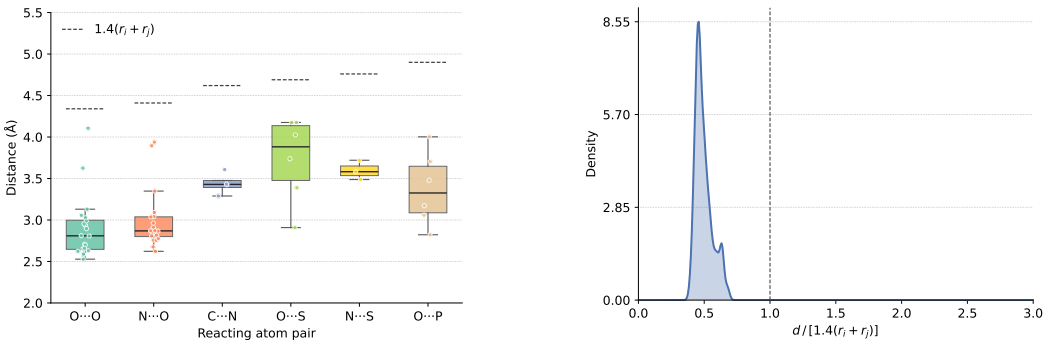
4.3 CRYSTAL STRUCTURE VALIDATION

To assess whether the proposed distance thresholds are consistent with experimentally observed catalytic geometries, we collect, for each entry, the M-CSA reference PDB together with additional crystal structures sharing the same UniProt accession, identified through SIFTS (Dana et al., 2019). For each reacting atom pair, distances are evaluated across all structures in which the pair can be reliably assigned, and the median observed distance is retained to reduce sensitivity to individual crystal conformations.

A pair is included only when both sides of the reacting atoms can be reliably mapped. On the residue side, catalytic positions must remain unmutated in the target structure and map to the correct chain and sequence indices through UniProt residue correspondence. Hydrogen atoms specified in the mechanism are represented by their parent heavy atoms, and equivalent atoms (e.g., carboxylate oxygens) are resolved by selecting the nearest contact. On the ligand side, the crystal structure must contain the cognate substrate, cofactor, or a close analogue. Candidate ligands are required to have Morgan-fingerprint Tanimoto similarity above 0.7 and a maximum common substructure that establishes a valid correspondence for the reacting atoms. Distances are computed only when all ligand components involved in the reacting pair are successfully matched.

Under these criteria, 69 reacting atom pairs can be validated. As shown in Figure 3, all validated pairs satisfy the proposed upper-bound criterion $1.4(r_i + r_j)$, and their element-specific distance distributions are consistent with near-contact geometries. These observations support the use of this

threshold as a catalytic geometric compatibility criterion. Coverage is limited by ligand completeness, mutations, and unsuccessful mappings; therefore, this analysis evaluates threshold behavior on the experimentally observable subset rather than validating every pair in the benchmark.



(a) (b)

Figure 3: Assessment of the proposed distance criterion on crystal structures. (a) Observed distances of validated reacting atom pairs grouped by element type; dashed lines indicate the thresholds $1.4(r_i + r_j)$. (b) Distribution of normalized distances $d/[1.4(r_i + r_j)]$.

4.4 Co-folding evaluation on the core set

Models are evaluated under a unified protocol: curated functional-unit sequences and substrate/cofactor SMILES as inputs, with reacting pairs extracted from predicted coordinates using precomputed atom mappings. Success rate is reported as the fraction of reacting atom pairs satisfying the distance thresholds. Table 1 reports co-folding success rates on the core set.

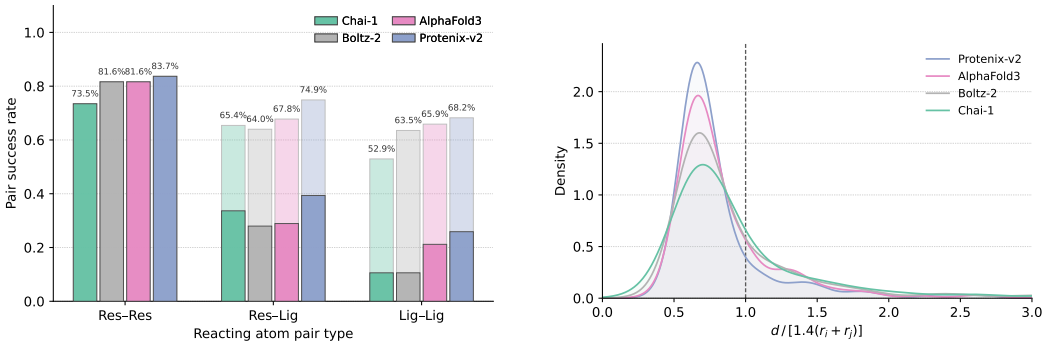
Under the distance-only criterion, Protenix-v2 achieves the highest pair- and complex-level success rates at Top-1 (79.57% / 73.64%). At Top-5, it achieves the highest pair success rate (83.28%), while its complex-level success rate (78.66%) is tied with Boltz-1. Allowing five samples improves success rates across all models, indicating that additional sampling can recover a fraction of failures observed in individual predictions.

Table 1: Co-folding success rates (%) on the core set (N=239 complexes, 320 reacting pairs). Pair@K and Cplx@K denote pair- and complex-level success rates at Top-K. w/o PB-Valid and w/ PB-Valid denote results without and with the PB-Valid constraint, respectively. Bold and underline indicate the best and second-best results; ties receive the same marking.

	w/o PB-Valid				w/ PB-Valid			
Model	Pair@1	Pair@5	Cplx@1	Cplx@5	Pair@1	Pair@5	Cplx@1	Cplx@5
Chai-1	67.80	78.02	60.67	72.80	35.91	51.08	23.01	39.75
Boltz-2	70.90	80.19	64.44	76.57	33.44	44.58	22.18	35.98
Boltz-2x	71.21	79.26	65.69	74.90	**49.54**	**69.97**	**39.33**	**63.60**
Boltz-1	73.07	<u>81.73</u>	68.62	**78.66**	29.41	38.70	17.57	27.62
AlphaFold 3	73.99	80.80	68.20	<u>76.99</u>	36.84	51.39	24.27	42.68
Protenix-v1	73.99	81.11	67.78	<u>76.99</u>	39.32	47.99	26.78	38.49
Boltz-1x	<u>74.61</u>	79.26	<u>69.46</u>	75.73	<u>45.20</u>	<u>65.02</u>	<u>34.31</u>	<u>59.00</u>
Protenix-v2	**79.57**	**83.28**	**73.64**	**78.66**	<u>45.20</u>	54.49	32.64	43.93

Imposing PB-Valid changes the ranking of models. Boltz-2x achieves the highest performance across all columns with PB-Valid, while Boltz-1x performs best among the remaining models. This pattern is consistent with the design goals of the physics-inspired inference-time steering mechanisms used by the Boltz-x models. In contrast, several models that are competitive under distance-only evaluation, including Protenix-v2 and Boltz-1, show substantial decreases when physical plausibility is additionally required. Distance-based evaluation and PB-Valid capture complementary aspects of prediction quality.

On standard single-ligand docking benchmarks, Boltz-1x retains relatively high performance when PB-Valid is imposed (Jiang et al., 2026). In our enzyme–substrate/cofactor setting, however, requiring PB-Valid together with catalytic geometry substantially reduces success rates (Pair@1: 74.6% $\rightarrow$ 45.2% for Boltz-1x; 71.2% $\rightarrow$ 49.5% for Boltz-2x). This indicates that simultaneously satisfying catalytic geometry and physical ligand validity remains challenging in enzyme–substrate/cofactor systems, where predictions often place multiple molecular components simultaneously and accommodate catalytic sites that span subunit interfaces while preserving chemically plausible ligand conformations. These additional structural and geometric constraints leave substantial room for model improvement.



(a) (b)

Figure 4: Core-set analysis under the distance criterion (Top-1) for representative models (Chai-1, Boltz-2, AlphaFold 3, Protenix-v2). (a) Pair success rate by reacting-pair type (unfilled bar: without PB-Valid; filled bar: with PB-Valid). (b) Distribution of normalized distances $d/[1.4(r_i + r_j)]$.

Pair success is strongly stratified by reacting-pair type (Figure 4a). Residue–residue pairs are easiest (about 73–84% across the representative models), followed by residue–ligand pairs (about 65–75%), while ligand–ligand pairs are the most challenging (about 53–68%). Difficulty increases with the involvement of ligand atoms in the reacting pair. Protenix-v2 achieves the highest success rates across all three strata, with the largest margin observed for residue–ligand pairs. Normalized distances $d/[1.4(r_i + r_j)]$ exhibit similar distributions across representative models (Figure 4b): the density concentrates below the threshold, with a long tail extending beyond it. This suggests that many geometric failures correspond to moderate deviations from the threshold rather than extreme separation from the catalytic site. Differences between models are mainly reflected in the concentration near the threshold and the tail distribution, rather than qualitatively different distance patterns.

4.5 Co-folding evaluation on the extended set

Extended-set evaluation. The extended set evaluates catalytic geometry on enzymes that share the same EC sub-subclass as a core entry but differ in full EC number and act on alternative substrates (Section 3.6). Let $\mathcal{S}$ denote the core entries with at least one extended enzyme, and let $\mathcal{U}(e)$ be the corresponding set of extended enzymes. Evaluation is restricted to $\mathcal{S}$.

For a predicted complex x, let $r_K(x)$ denote the fraction of reacting pairs that satisfy the distance threshold under Top-K, where a pair succeeds if any of the K samples passes. We use $K = 5$. For each $e \in \mathcal{S}$, we define

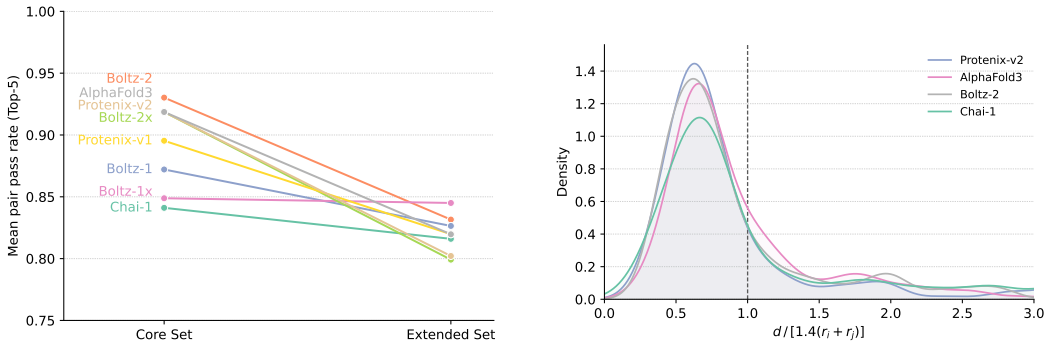
$$R_K^{\text{core}}(e) = r_K(x_e), \qquad R_K^{\text{ext}}(e) = \frac{1}{|\mathcal{U}(e)|} \sum_{u \in \mathcal{U}(e)} r_K(x_u),$$

and report their averages over $\mathcal{S}$ as the **mean pair pass rate**. This weighting gives each core entry equal weight, preventing large extension families from dominating the score.

All models achieve lower pass rates on the extended set than on the matched core subset (Figure 5a), although the magnitude of the decrease varies across models. Protenix-v2 shows a marked decrease, whereas Boltz-1x shows the smallest decrease and the highest extended-set rate among the evaluated

models. Thus, model performance changes between the matched core and extended sets, indicating that core-set performance alone does not fully capture generalization to related enzymes with alternative substrates.

Normalized distances $d/[1.4(r_i + r_j)]$ on the extended set show a similar overall distribution to the core set (Figure 5b), with density concentrated below the threshold and a long tail beyond it. A larger fraction of pairs lies above the threshold than in the matched core subset, indicating that the extended set is more challenging under the same distance criterion.



(a) (b)

Figure 5: Generalization from core to extended enzymes. (a) Per-entry mean pair pass rate (Top-5) on $\mathcal{S}$ for all evaluated models, comparing the matched core subset and associated extended enzymes. (b) Distribution of normalized distances $d/[1.4(r_i + r_j)]$ (Top-1) for representative models.

5 Conclusion

We present CatGeoBench, a reusable benchmark for assessing *catalytic geometric compatibility* under mechanistic constraints. Starting from electron-flow annotations in M-CSA, we extract intermolecular bond-forming reacting atom pairs from the catalytic initial state through the first intermolecular structure-changing event, and define distance thresholds from scaled van der Waals radii. Predicted co-folded complexes are scored against these thresholds directly. The benchmark comprises a core set aligned with curated mechanisms and an extended set of related enzymes that act on alternative substrates while preserving transferable distance criteria. A unified evaluation of co-folding models yields two main findings. First, models often achieve high success under the distance criterion on the core set, yet performance drops sharply when ligand poses are additionally required to pass PoseBusters physicochemical checks; even Boltz-1x and Boltz-2x, which apply physics-inspired inference-time steering, remain far from saturated under this joint standard, indicating that satisfying catalytic geometry and ligand physical plausibility together is still difficult for enzyme–substrate–cofactor systems. Second, the extended set is consistently harder, and strong core performance does not reliably predict success on related enzymes with alternative substrates, so the extended protocol is a necessary complement for assessing generalization. CatGeoBench thus provides a mechanism-grounded protocol for structure prediction in catalytic settings.

Limitations and Future Work. Several limitations remain for CatGeoBench: (i) its scale is constrained by the number of curated enzyme reactions in M-CSA that contain complete electron-flow mechanisms, and the coverage of reaction types remains incomplete; (ii) the extended set is built under strict matching criteria: EC sub-subclass, catalytic residue types and spatial arrangements, and reaction-center correspondence. These criteria are intended to preserve the applicability of the distance criteria defined on a core entry after expansion, but at the cost of excluding some mechanistically similar enzyme pairs whose reaction-center local chemical environments differ (e.g., amide versus carboxylic ester hydrolysis), or whose conjugate acid-base residues vary among multiple candidates within a sub-subclass; and (iii) the current criteria consider only distances between reacting atom pairs, and future work will introduce mechanism-derived angular constraints to more fully characterize catalytic geometric compatibility.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript writing assistance, including language polishing and drafting. We have not used generative AI tools for research ideation or experimental design, and other generative-AI-assisted research tasks are not applicable to this work. We have reviewed and verified all AI-assisted text for accuracy, consistency with the underlying work, and appropriate attribution. We take responsibility for the final content of this work, including all text, claims, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The dataset and code are available at `https://anonymous.4open.science/r/catgeobench-42E0`.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Woody Ahern, Jason Yim, Doug Tischer, Saman Salike, Seth M Woodbury, Donghyo Kim, Indrek Kalvet, Yakov Kipnis, Brian Coventry, Han Raut Altae-Tran, et al. Atom-level enzyme active site scaffolding using rfdiffusion2. *Nature Methods*, 23(1):96–105, 2026.

Ivan Anishchenko, Yakov Kipnis, Indrek Kalvet, Guangfeng Zhou, Rohith Krishna, Samuel J Pellock, Anna Lauko, Gyu Rie Lee, Linna An, Justas Dauparas, et al. Modeling protein–small molecule conformational ensembles with placer. *Proceedings of the National Academy of Sciences*, 122(45):e2427161122, 2025.

Helen M Berman, Tammy Battistuz, Talapady N Bhat, Wolfgang F Bluhm, Philip E Bourne, Kyle Burkhardt, Zukang Feng, Gary L Gilliland, Lisa Iype, Shri Jain, et al. The protein data bank. *Biological Crystallography*, 58(6):899–907, 2002.

Thomas C Bruice. A view at the millennium: the efficiency of enzymatic catalysis. *Accounts of Chemical Research*, 35(3):139–148, 2002.

Thomas C Bruice and Felice C Lightstone. Ground state and transition state contributions to the rates of intramolecular and enzymatic reactions. *Accounts of chemical research*, 32(2):127–136, 1999.

Jasper Butcher, Rohith Krishna, Raktim Mitra, Rafael I Brent, Yanjing Li, Nathaniel Corley, Paul T Kim, Jonathan Funk, Simon Mathis, Saman Salike, et al. De novo design of all-atom biomolecular interactions with rfdiffusion3. *bioRxiv*, 2025.

Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.

Jose M Dana, Aleksandras Gutmanas, Nidhi Tyagi, Guoying Qi, Claire O'Donovan, Maria Martin, and Sameer Velankar. Sifts: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic acids research*, 47(D1):D482–D489, 2019.

Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Gabriel Studer, Daniel Kovtun, et al. Plinder: The protein-ligand interactions dataset and evaluation resource. *BioRxiv*, pp. 2024–07, 2024.

Yize Jiang, Xinze Li, Yuanyuan Zhang, Jin Han, Youjun Xu, Ayush Pandit, Zaixi Zhang, Mengdi Wang, Mengyang Wang, Chong Liu, et al. Posex: Ai defeats physics-based methods on protein ligand cross-docking. In *International Conference on Learning Representations*, volume 2026, pp. 88211–88245, 2026.

Hyunbin Kim, Rachel Seongeun Kim, Milot Mirdita, Jaewon Yoon, and Martin Steinegger. Structural motif search across the protein universe with folddisco. *Nature Biotechnology*, pp. 1–5, 2026.

Sowmya Ramaswamy Krishnan, Nishtha Pandey, Rajgopal Srinivasan, and Arijit Roy. A structure-oriented kinetics dataset of enzyme-substrate interactions. *Scientific Data*, 12(1):1489, 2025.

Anna Lauko, Samuel J Pellock, Kiera H Sumida, Ivan Anishchenko, David Juergens, Woody Ahern, Jihun Jeung, Alexander F Shida, Andrew Hunt, Indrek Kalvet, et al. Computational design of serine hydrolases. *Science*, 388(6744):eadu2454, 2025.

Alex Morehead, Nabin Giri, Jian Liu, Pawan Neupane, and Jianlin Cheng. Assessing the potential of deep learning for protein–ligand docking. *Nature Machine Intelligence*, 8(1):32–41, 2026.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

António J M Ribeiro, Gemma L Holliday, Nicholas Furnham, Jonathan D Tyzack, Katherine Ferris, and Janet M Thornton. Mechanism and catalytic site atlas (m-csa): a database of enzyme reaction mechanisms and active sites. *Nucleic acids research*, 46(D1):D618–D623, 2018.

Chao Song, Zhiyuan Liu, Han Huang, Liang Wang, Qiong Wang, Jian-Yu Shi, Hui Yu, Yihang Zhou, and Yang Zhang. Enzycontrol: adding functional and substrate-specific control for enzyme backbone generation. *Advances in Neural Information Processing Systems*, 38:153176–153206, 2026.

Chai Discovery team, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhonikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *BioRxiv*, pp. 2024–10, 2024.

Protenix Team, Yuxuan Zhang, Chengyue Gong, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, Jiaqi Guan, Lan Wang, Yanping Yang, et al. Protenix-v1: Toward high-accuracy open-source biomolecular structure prediction. *bioRxiv*, pp. 2026–02, 2026.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, pp. 2024–11, 2025.

Chuanxi Zhang, Yinghui Feng, Yiting Zhu, Lei Gong, Hao Wei, and Lujia Zhang. Nac4ed: A high-throughput computational platform for the rational design of enzyme activity and substrate selectivity. *Mlife*, 3(4):505–514, 2024.

Yuxuan Zhang, Chengyue Gong, Jinyuan Sun, Jiaqi Guan, Milong Ren, Song Xue, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, et al. Protenix-v2: Broadening the reach of structure prediction and biomolecular design. *bioRxiv*, pp. 2026–04, 2026.

A Dataset Construction and Statistical Analysis

- **Mechanistic checkpoint.** A full catalytic cycle may proceed through multiple intermediates and structural rearrangements, whereas the benchmark evaluates a single static predicted complex. We therefore focus on the earliest intermolecular electron-flow event that changes molecular structure, defined as an event that alters heavy-atom connectivity or bond order. Reacting atom pairs involved in this event, together with intermolecular bond-forming pairs from preceding electron-flow events, constitute the geometric evaluation targets. Electron-flow steps that only redistribute electrons without changing molecular structure, such as aromatic resonance, are treated as structure-preserving.
- **Local reaction geometry.** We retain mechanistic steps only when their relevant transformations can be represented by local atom-pair geometry. Processes that cannot be expressed in this form, such as long-range electron tunneling, are excluded because they do not admit a direct static geometric criterion.
- **Catalytic initial-state composition.** Substrate and cofactor identities and stoichiometries are reconciled with both the overall balanced reaction and the stepwise mechanism. The input contains only molecular species present in the reactant-side initial state of the evaluated step; species introduced only in later mechanistic steps are excluded, including in ping-pong and related mechanisms.
- **Input and structural compatibility.** Protein sequences, catalytic residues, engineered mutations, and modified residues are curated to ensure consistency between the reference structure and the annotated mechanism. Ligand inputs are restricted to discrete non-polymeric components that can be represented by the evaluated co-folding pipelines, including metal ions and organic cofactors. Entries requiring nucleic acids, peptides, or gaseous reactants such as O_2 and CO_2 are excluded.
- **Equivalent-site multiplicity.** When a reference assembly contains multiple equivalent catalytic sites, substrate and cofactor copy numbers are scaled to the number of sites identified by matching the mechanism-defined catalytic residues in identity and local geometry. This ensures that each catalytic site is supplied with a complete ligand set and avoids penalizing predictions because equivalent ligands are distributed across different sites.

A.1 Dataset Construction Process

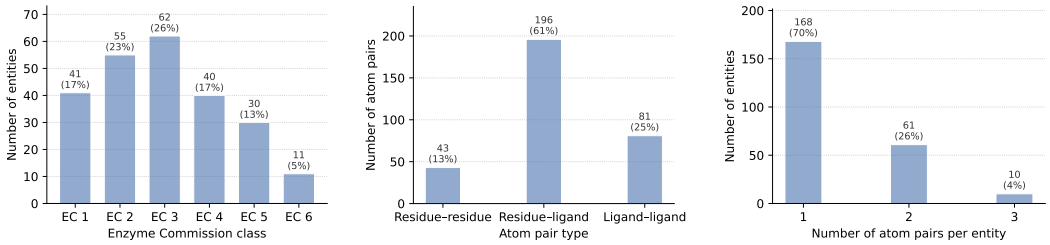
Table 2: Construction of the CatGeoBench.

Selection Step	Remaining entities
Entity record available via M-CSA API	1003
Exactly one curated reaction mechanism per entry	880
Mechanism length fewer than 10 steps	841
Exclude EC class 7 (translocases)	831
First mechanistic step provides a downloadable MRV file	572
Reactants and cofactors outside the excluded ChEBI list (Table 3)	475
No reactant SMILES containing wildcard atoms (*)	325
All reactants with molecular weight $\leq 900\,\mathrm{Da}$	315
No radical character in any step description	309
At least one valid electron-flow chain yielding reacting atom pairs	273
Manual M-CSA curation	270
Exclude heteromeric assemblies	256
Exclude assemblies with more than 8 protein chain copies	252
Exclude entries exceeding the input size limit	239

Table 3: Excluded reactant ChEBI identifiers

Name	ChEBI ID
dioxygen	15379
carbon dioxide	16526
superoxide	18421
double-stranded DNA	4705
single-stranded DNA	9160
RNA (poly(A))	8756
messenger RNA	33699
transfer RNA	17843
DNA 8-oxoguanine	137052
3-methyladenine DNA	137218
poly[(GA)]	137464
2'-deoxyadenosine	17256
dipeptide	46761
acyl-carrier protein	13534
pectin	17309
di-μ-sulfido-diiron(2+)	33737
di-μ-sulfido-diiron(1+)	33738
Fe_8S_7 iron–sulfur cluster	21143
tetra-μ_3-sulfido-tetrairon	49883
iron–sulfur–molybdenum cluster	48796
Cu(I)–S–Mo(IV)(=O)O–NBIC cluster	134062
Mo(=O)-bis(molybdopterin guanine dinucleotide)(4−)	60539

A.2 Statistical Characteristics



(a) EC class distribution (b) Reacting atom-pair types (c) Number of reacting atom pairs

Figure 6: Statistical overview of the CatGeoBench. (a) Distribution of entries by Enzyme Commission (EC) class. (b) Distribution of reacting atom pairs by type (residue–residue, residue–ligand, ligand–ligand). (c) Distribution of the number of reacting atom pairs per entry.

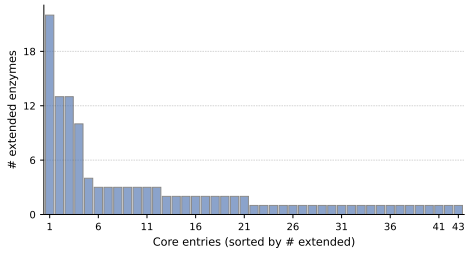
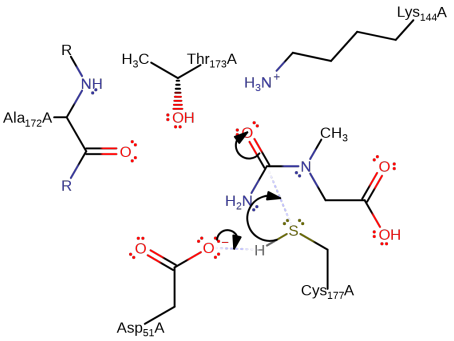
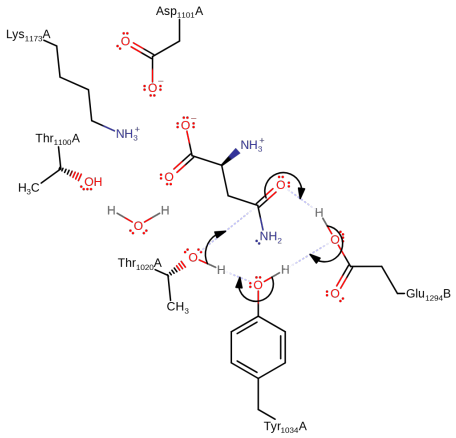


Figure 7: Distribution of associated extended enzymes across core entries. Each bar represents a core entry, with the bar height indicating the number of associated extended enzymes. Core entries are sorted by the number of associated extended enzymes.

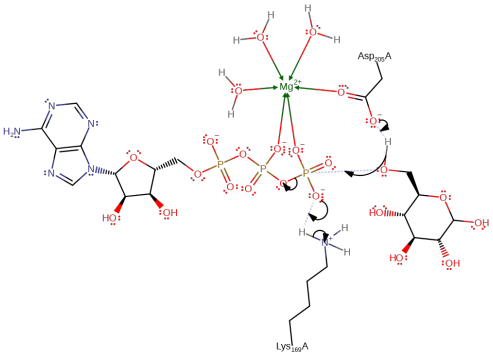
A.3 Electron-flow Chain Topologies



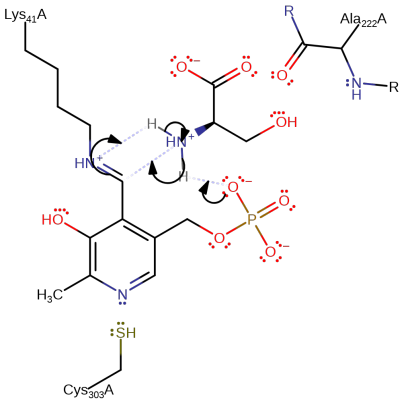
(a) Single linear chain (94%)



(b) Single cyclic chain (3%)



(c) Multiple linear chains (2%)



(d) Mixed linear and cyclic chains (1%)

Figure 8: Electron-flow chain topologies from the first mechanistic step of each entry. Subpanels show representative single linear chains, single cycles, multiple linear chains, and chain–cycle combinations; percentages are given in the subpanel captions.