STRUPAIR: A STRUCTURE-ONLY CLASSIFIER AND BENCHMARK FOR MULTI-STATE PROTEIN DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Designing proteins with multiple functional states, including allosterically regulated switches, molecular sensors, and enzymes, requires generating physically plausible conformations that represent the distinct states needed to carry out the intended function. However, evaluating whether computationally generated structural states correspond to realistic and functionally meaningful conformational changes remains challenging, as commonly used structural similarity metrics do not fully capture the diverse modes and scales of conformational change observed in natural proteins. Here, we introduce StruPair, a structure-only framework for classifying and benchmarking protein conformational changes using structure pairs collected from experiments, molecular dynamics (MD) simulations, and computational design. We train a hybrid classifier that combines structural comparison features with graph neural network representations. The trained StruPair classifier assigns each structure pair to one of five ordered classes based on the extent of conformational change and the preservation or reorganization of secondary and tertiary structure. Using only paired backbone structures, without amino acid identities or sequence embeddings, allows StruPair to evaluate candidate pairs before sequence design or optimization. We provide the evaluation set and the trained classifier through a web interface and API.

1 INTRODUCTION

1.1 MOTIVATION AND PROBLEM SETTING

Protein function often depends on multiple conformational states. Ligand binding, conformational change near the active sites, and relative domain movements can reorganize a protein into distinct states associated with different functional outcomes. Such conformational changes are central to natural and designed systems including protein switches, sensors, and allosterically regulated assemblies (Langan et al., 2019; Praetorius et al., 2023; Guo et al., 2025; Quijano-Rubio et al., 2021; Pillai et al., 2024; Lee et al., 2026; Guo et al., 2026). Designing proteins with multiple functional states requires assessing both the quality of each generated structure and the relationship between the intended states to be appropriate for the intended function (Leaver-Fay et al., 2011; Abrudan et al., 2025). If two target structures are nearly identical, they may not express sufficiently distinct functional states. Conversely, a large structural difference does not necessarily imply desirable change. Large-scale motions can arise from coherent movement of preserved domains, and structurally similar global changes can also involve substantial local deformation. We therefore need to assess state pairs considering not only how much the structures differ, but also which structural elements remain preserved, which regions reorganize, and whether the observed differences can be described as coordinated movements of relatively intact structural units.

This assessment is particularly important in protein-design workflows in which conformation pairs can be generated before their amino acid sequence has been determined or optimized. For example, generative structure-design methods can first produce candidate backbone conformations (Watson et al., 2023; Krishna et al., 2024; Butcher et al., 2025; Bauer et al., 2025; Jing et al., 2026), after which a shared sequence is designed to be compatible (Dauparas et al., 2022; 2025; Shuai et al., 2025). In this scenario, many candidate backbone pairs may be geometrically valid as individual structures but differ in whether their relationship represents a plausible multi-state design target.

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.

StruPair is designed for this setting to operate on paired backbone structures alone, allowing conformational relationships of the pairs to be evaluated either before or after sequence design.

1.2 Limitations of Existing Structure-Pair Evaluation

RMSD and TM-score represent global structural similarity (Zhang & Skolnick, 2004), but they cannot fully describe the modes of conformational change. For example, a large relative rotation of domains can preserve internal structure (Gerstein et al., 1994; Hayward & Berendsen, 1998), and a small global difference may involve locally reorganized contacts across several regions. Changes confined to a small region may also be weakly reflected in global-structure-based similarity scores (Mariani et al., 2013). To evaluate conformational changes not only based on the magnitude of structural difference, we define five classes that categorize structure pairs according to different patterns and extents of conformational change guided by human judgments. Existing structural resources provide important examples of conformational diversity, but their conditions and motion categories do not directly correspond to this classification task; we discuss them in Section 2. Building a machine-learning model for this task also requires clear class definitions, annotation procedures, and appropriate data splits. Finally, two observed endpoint structures alone do not establish a transition pathway or kinetics. In this study, we limit StruPair to classifying the structural relationship between the observed endpoint conformations for multi-state protein design applications.

1.3 Goals and Contributions

StruPair classifies structural relationships from small structural variation to extensive reorganization into Classes 4–0. Basic features used for classifying include the extent of change, preservation of shared structure, relative movement of structural units, and internal reorganization.

This study makes three contributions. First, we assemble a structure-pair dataset from multiple sources, including publicly available structural resources and evaluation sets, and annotate residue correspondences and conformational-change class labels for each pair. Second, we establish a benchmark for predicting the conformational-change classes from backbone structure pairs without sequence information. Third, we develop and provide a hybrid classifier that combines structural comparison features with graph neural network representations to classify the relationship between candidate structural states.

The preferred class of conformational change can depend on the multi-state protein design objective (Tuinstra et al., 2008; Porter & Looger, 2018). StruPair provides a framework for evaluating and prioritizing candidate state pairs according to the type and extent of conformational change appropriate for a particular design problem (Figure 1).

2 Related Work

2.1 Related Datasets and Benchmarks

Databases of macromolecular motion and conformational diversity provide a starting point for studying structure pairs (Gerstein & Krebs, 1998). CoDNaS 2.0 links multiple experimentally observed conformations to conditions and protein properties (Monzon et al., 2016). PSCDB organizes ligand-free and ligand-bound structure pairs into motion types and specifies how to judge cases with more than one type of motion (Amemiya et al., 2011; 2012). These resources describe states and motions, while StruPair asks how well methods reproduce human judgments of preservation and reorganization across five ordered classes.

MD resources provide a different source of structural variation. mdCATH contains trajectories and derived structural quantities for diverse protein domains (Mirarchi et al., 2024). Simulation-derived structures should remain distinguishable from experimentally determined structures because their sampling conditions differ (Piana et al., 2011). StruPair records whether pairs come from experiments, MD simulations, or computational design, and reports the source composition of the training corpus and evaluation set separately.

Annotation resources also show the value of recording how labels were produced. DisProt links curated judgments to publications, residue regions, and evidence (Quaglia et al., 2024), while MobiDB

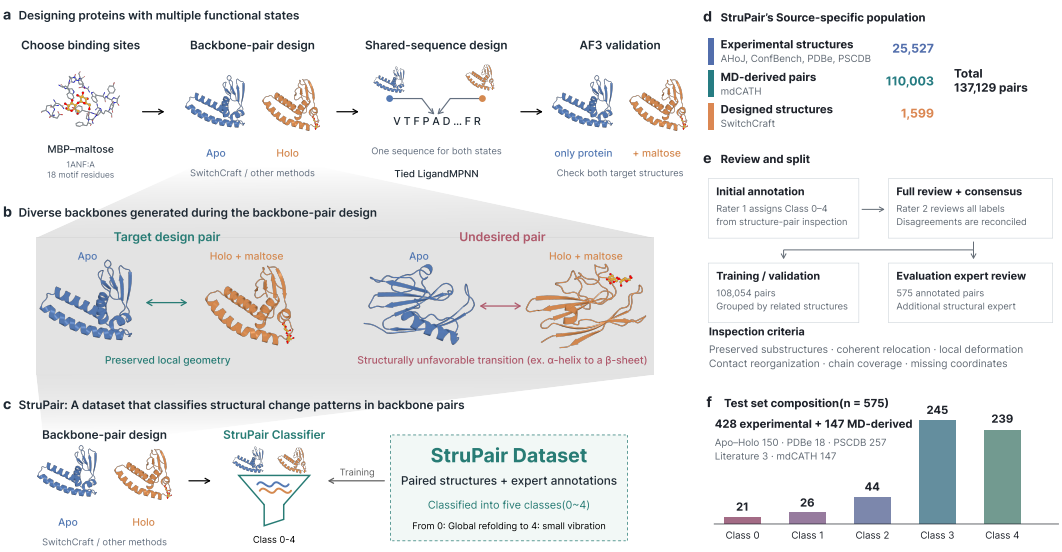

Figure 1: StruPair data and its use in multi-state design. (a) Backbone-pair generation, shared-sequence design (Dauparas et al., 2025), and validation of both target states (Abramson et al., 2024). (b) Candidate pairs with different patterns of structural preservation. (c) Classification of annotated backbone pairs into Classes 0–4. (d) Sources of the internal corpus. (e) Annotation review and data splitting. (f) Source and class composition of the 575-pair evaluation set.

distinguishes manual annotations, derived information, and predictions (Piovesan et al., 2021). StruPair likewise separates source observations, human labels, and model predictions. For benchmark design, ATOM3D, FLIP, and ProteinGym illustrate the need to state tasks, splitting units, measurements, and comparison procedures (Townshend et al., 2021; Dallago et al., 2021; Notin et al., 2023). We report these elements so that external methods can be assessed on the same structure-pair task.

2.2 Protein Structure Representations and Structure-Pair Comparison

A structure pair can be represented by explicit comparison features or by learned representations of its two structures. RMSD and TM-score summarize global relationships, whereas contact preservation and per-residue displacement describe more local differences. We train a classifier on conventional structural features as a baseline. Thus, our conventional-feature baseline is a trained structure-pair classifier, not an individual metric threshold.

GeoBPE represents backbone geometry with structural tokens at several scales (Sun et al., 2026). We combine representations of the two endpoints and train a classifier for StruPair Classes 0–4. ENSEMBITS learns from structural ensembles and can provide a representation for a single structure (Shi & Oliver, 2026). We pass each endpoint as one frame through the frozen pretrained model and use its representations for the same pair-classification task. Neither original method was designed to directly predict StruPair classes; Appendix C specifies our pair-combination and classifier settings.

2.3 Differences from Existing Work

StruPair turns human judgments of structural change into a common prediction and evaluation task. It tests whether existing features and learned representations explain these judgments and identifies class boundaries that remain difficult. Because the input contains backbone structures but no amino acid identities or sequence embeddings, the task also applies to design candidates before sequence selection.

3 DATASET CONSTRUCTION

3.1 DESIGN PRINCIPLES AND SCOPE

Each sample consists of two backbone structures, A and B, the correspondence between the residues being compared, and the sources and annotations of the structures. The evaluation scope is fixed to a selected chain or a predefined region, and we record the residues observed in both structures and any missing regions. The model input consists of geometric features computed from coordinates and residue connectivity, without amino acid identities or sequence embeddings.

The structure-only condition applies to the model's prediction inputs. The source data may include sequence information, which can be used during dataset construction to verify residue correspondences, determine whether structures belong to the same protein or construct, and split data by homology. We thus distinguish metadata used for dataset construction from inputs used for prediction.

The five categories describe both the extent of change and how existing structures are preserved and reorganized. Higher category values generally indicate smaller changes that preserve the existing fold, while lower values indicate more extensive reorganization. However, the intervals between categories do not represent equal physical differences. Also, domains may move substantially while preserving their internal structures, so we do not define categories as RMSD ranges.

3.2 DATA SOURCES AND COLLECTION

The data resources are divided into an internal corpus for training and development and an evaluation set to be released publicly. The current fixed version of the internal corpus contains 25,527 experimental structure pairs, 110,003 MD-derived structure pairs, and 1,599 computationally designed structure pairs, for a total of 137,129 pairs. For each pair, we record the source data, collection route, and type of evidence, categorized as experiment, simulation, or design.

Experimental pairs came from AHoJ (Feidakis et al., 2022), ConfBench (Qiao et al., 2025), PDBe (Ellaway et al., 2024), PSCDB (Amemiya et al., 2012), and individual publications; mdCATH (Mirarchi et al., 2024) and SwitchCraft (Jing et al., 2026) provided simulated and designed pairs, respectively. Source provenance and collection details are given in Appendix A.

The public evaluation resource contains 575 structure pairs sampled from the full dataset to preserve a distribution similar to that of the overall dataset. The evaluation set consists of 150 Apo–Holo pairs, 147 mdCATH pairs, 18 PDBe pairs, 257 PSCDB pairs, and three cases collected from individual publications. Of these, 428 are experimental pairs and 147 are MD-derived pairs. The three individual literature cases are GltPh, Mhp1, and LeuT. For source reporting, the 275 pairs collected from PDBe and PSCDB are grouped as EPC-Curated, whereas the three individual literature cases are grouped as EPC-Direct.

3.3 PREPROCESSING, QUALITY FILTERING, AND DEDUPLICATION

We standardized each structure pair to a selected chain and corresponding backbone residues, retaining missing coordinates as missing. Source-specific checks covered residue correspondence, observed-residue coverage, and backbone completeness before metric calculation and graph construction. Duplicate endpoint pairs were merged, and related samples were grouped for splitting. Of the 137,129 collected pairs, 28,500 were removed by these checks and by deduplication, leaving 108,629 pairs that entered the training, validation, and evaluation splits. Appendix A gives the exact processing rules and thresholds.

3.4 LABELS, ANNOTATIONS, AND EVALUATION TASK CONSTRUCTION

The labeling unit was a pair of backbone structures with defined residue correspondences, and each sample was assigned one category from Class 0 to Class 4. Judgments considered the extent of structural change, the preservation of shared fold organization, the relative rearrangement of structural elements, and whether extensive structural reorganization had occurred. Human annotations consisted of a categorical label for each structure pair.

Annotators inspected superposed structures with structural comparison metrics as supporting information. Disagreements between the two annotators required reconciliation for 16% of the total dataset.

Individual annotator choices were not logged, so a chance-corrected coefficient cannot be computed directly. Assuming both annotators shared the marginal distribution of the consensus labels, the observed 84% agreement corresponds to Cohen's $\kappa \approx 0.75$. This value depends on an unverifiable assumption about the annotators' marginals and should be read as indicative only.

Appendix B clarifies the class interpretation, and Appendix C describes the release model.

We define five categories of structural relationships based on the preservation of internal geometry, contact organization, and the relative arrangement of shared structural regions. Class 4 describes closely preserved folds with minor local variations. Class 3 describes coherent repositioning of structural regions, such as loops or domains, whose internal structures remain largely preserved. Class 2 describes regional rearrangements accompanied by appreciable internal deformation or contact reorganization, while substantial structure remains recognizable. Class 1 describes extensive reorganization with identifiable shared substructures that retain coherent internal geometry and contact relationships. Class 0 describes global reorganization in which little shared fold organization remains beyond isolated structural elements.

Annotation considers both the extent of the differences and how they are distributed within and between structural regions. The Class 4–Class 3 boundary depends on whether the differences amount to minor variations or form a coherent rearrangement of identifiable regions. For the Class 3–Class 2 distinction, the key question is whether relative repositioning of largely preserved regions adequately explains the observed differences. Changes in contacts between such regions, together with localized deformation at hinges or connecting segments, are compatible with Class 3. Class 2 applies when explaining the differences also requires appreciable changes in geometry or contact relationships within the regions.

The Class 2–Class 1 and Class 1–Class 0 boundaries depend on how much coherent structural organization remains. Class 2 retains substantial recognizable structure, whereas in Class 1 preservation is limited to identifiable substructures within a more extensively reorganized fold. To distinguish Class 1 from Class 0, we considered whether the remaining shared parts preserve meaningful internal relationships, including the packing and contacts among multiple secondary-structure elements. Individual helices or strands that retain their local shape do not, by themselves, establish preservation of a shared substructure.

These categories describe relationships observed between two endpoint structures; they do not establish a transition pathway or measure transition rates or energy barriers. Representative structure pairs are shown in Figure 2.

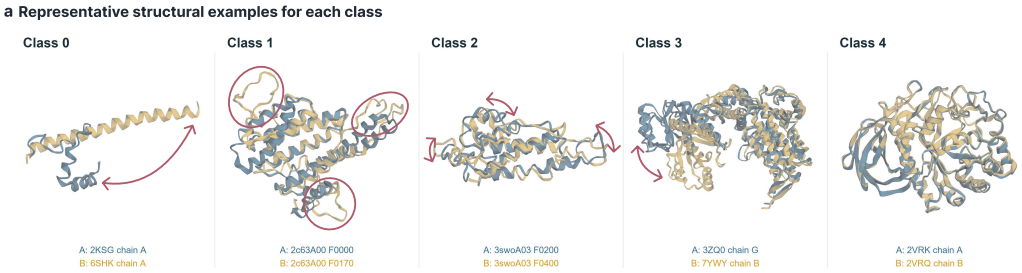

Figure 2: Representative structure pairs for the five structural relationship categories. From left to right, Class 0 through Class 4 illustrate increasing preservation of shared structural organization, from global reorganization to closely preserved folds with minor variations. Blue and yellow distinguish the two structures; arrows and circles highlight regions of structural difference.

3.5 Dataset Statistics and Characteristics

Figure 3 characterizes both the structural coverage of the evaluation set and the sequence-based family definition used for data splitting. As shown in Figure 3a, the 575 evaluation pairs span the major regions of the structural feature landscape represented by the full dataset. The first two principal components explain 69.2% and 14.4% of the variance, respectively, and the evaluation samples broadly overlap with the overall distribution while covering diverse regions of the feature space.

Figure 3b illustrates the sequence-based family criterion used to identify related structures. A sequence was considered to belong to the same family only when both sequence identity and alignment coverage satisfied the predefined thresholds of 30% and 80%, respectively. The highlighted examples demonstrate that satisfying either sequence identity or coverage alone was insufficient for family assignment. This joint criterion was used to reduce close sequence-level relationships between the evaluation set and the training and validation data.

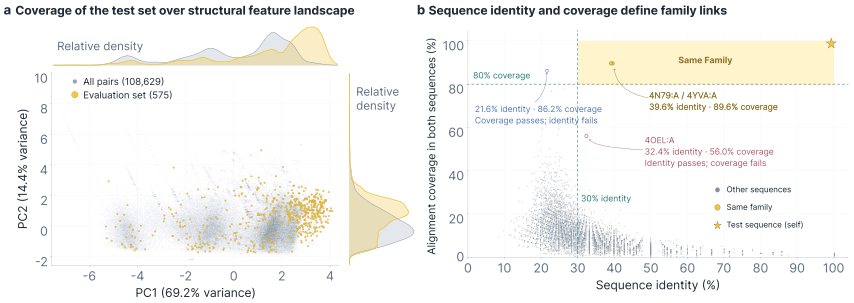

Figure 3: Structural feature distributions and sequence-based family relationships. (a) Projection of the 575-pair evaluation set and the 108,629 pairs labeled "All pairs" in the figure onto the first two principal components, with marginal relative densities. PC1 and PC2 explain 69.2% and 14.4% of the variance, respectively. (b) An example of family assignment using both sequence identity and alignment coverage in both sequences, with thresholds of 30% identity and 80% coverage. The highlighted examples show a pair that meets both thresholds and pairs that meet only one.

4 Dataset Evaluation and Analysis

4.1 Relation Between Structural Measures and Labels

StruPair labels describe both the extent of change between two structures and which structural units remain intact or reorganize. We first examine how these labels relate to familiar structure measures. Figure 4 shows the distributions of global displacement, structural similarity, contact preservation, and an approximate secondary-structure change measure across the five classes. The distributions describe class tendencies, but their overlap rules out treating any one measure as a fixed class boundary. Contact rewiring is defined as one minus contact Jaccard, so these two panels are not independent evidence.

The question for the predictive comparison is whether conventional features and learned backbone representations recover the same labels. A higher score for a combined representation may suggest useful information beyond a small set of features.

4.2 Comparison of Structural Representations

We compare six feature combinations on the same annotated 575-pair cohort: 54-conventional structural features, 148-expanded features, GeoBPE (Sun et al., 2026), ENSEMBITS (Shi & Oliver, 2026), and two combinations that add both learned representations to either the 54-feature A set or the 148-featured B set that expanded from A. The cohort contains

Table 1: Macro-F1 on 575 pairs.

Feature combination	Macro-F1
Conventional A (54)	0.8754
Expanded B (148)	0.9103
GeoBPE	0.6648
ENSEMBITS	0.7786
A (54) + G + E	0.9249
B (148) + G + E	0.9143

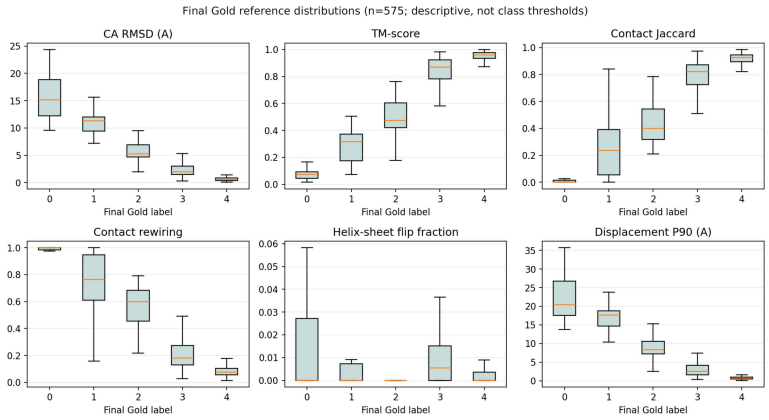

Figure 4: Distributions of structural comparison metrics across the five annotated classes in the 575-pair evaluation set. Box plots show Cα RMSD, a TM-based similarity score, contact Jaccard similarity, contact rewiring, a helix–sheet change proxy, and the 90th percentile (P90) of per-residue displacement for Class 0–4. These distributions provide descriptive references for the structural characteristics of each class and were not used as fixed thresholds for class assignment.

428 experimental pairs and 147 MD-derived pairs. Five-class macro-F1 gives each class equal weight and is the primary score in Table 1. For brevity, we denote GeoBPE as G and ENSEMBITS as E.

GeoBPE and ENSEMBITS alone score below the 54-feature A set. Adding both representations to the 54-feature A set gives 0.9249, while the combination with 148-feature B set gives 0.9143. These scores exceed the 54-feature A set by 0.0495 and 0.0389, respectively.

4.3 Hard-to-Distinguish Classes and Structural Examples

To understand which relationships become easier to classify as structural information is added, we examined pairs that a simpler readout missed but a richer readout classified correctly.

Class 4. A small region moves far, raising global displacement although most of the fold remains intact. **Class 3.** A large but coherent motion of preserved units can look severe in global measures even when their internal structures remain stable. **Class 2.** Modest overall displacement can hide several local deformations and contact changes, making the motion structurally complex. **Class 1.** Part of the structure is disrupted while other substructures remain, blurring the boundary with broader reorganization. **Class 0.** In a small protein, changes to only a few elements can reshape much of the fold, so the absolute size of the changed region can be misleading.

5 Classifier and Design Application

5.1 Hybrid Classifier and Label Prediction

The StruPair classifier combines a graph of the aligned backbone pair with seven global structural-comparison features (Figure 5a). Two graph layers encode residue-level changes, and pooled graph features are joined with the global features before an ordinal head predicts the five classes. We average class probabilities across three training seeds; Appendix C gives the implementation details.

Figure 5b reports macro-F1 of 0.875, 0.910, and 0.963 for the 54-feature, 148-feature, and hybrid classifiers, respectively. Figure 5c summarizes the structural-representation comparisons discussed in Section 4.2. GeoBPE and ENSEMBITS alone score below the conventional-feature readout, while the combined readouts score higher.

The hybrid classifier reaches a macro-F_1 of 0.9635 on the 575-pair evaluation set. All 22 errors fall between adjacent classes. No pair is misclassified by two or more classes (quadratic weighted κ = 0.9807). Table 4 in Appendix D gives the full confusion matrix with per-class confidence intervals.

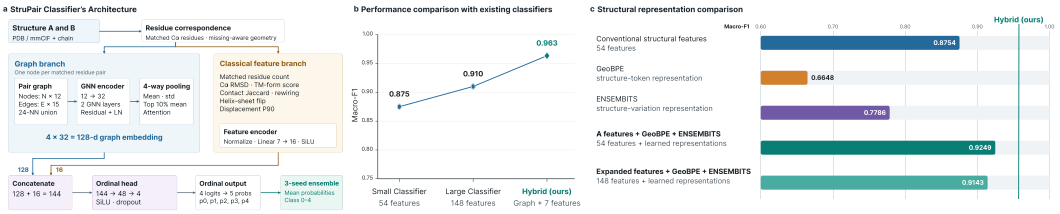

Figure 5: StruPair classifier and label-prediction results. (a) A pair graph and seven global structural features feed an ordinal classification head. (b) Macro-F1 for the 54-feature, 148-feature, and hybrid classifiers. (c) Macro-F1 for conventional features, GeoBPE, ENSEMBITS, and their combined readouts.

5.2 Selection of Multi-State Design Candidates

We tested whether StruPair can help select backbone pairs for a downstream design workflow. SwitchCraft (Jing et al., 2026) generated apo–holo candidate structures under ligand and motif constraints. We selected pairs assigned to Classes 2–3 and compared them with the unfiltered SwitchCraft candidates, assessing 100 pairs in each reported arm. LigandMPNN (Dauparas et al., 2025) designed one sequence for both states of each pair, and AlphaFold 3 (AF3) (Abramson et al., 2024) predicted the resulting apo and holo structures. We counted a pair as structurally recovered when both AF3 predictions had Cα RMSD at most 3 Å from their respective target backbones.

Under this criterion, 0 of 100 unfiltered candidates and 10 of 100 StruPair-selected Class 2–3 candidates recovered both target states (Figure 6a). Requiring pLDDT of at least 80 for both predictions left 2 StruPair-selected pairs. Figure 6b shows examples of the target and predicted structures. These counts support the use of StruPair for prioritizing candidates in this computational workflow; AF3 recovery alone does not establish an experimentally accessible transition or switching function.

6 Discussion

6.1 What StruPair Enables

StruPair turns expert judgments of structure pairs into a reusable five-class evaluation task. The classes describe both the extent of change and how shared structural units are preserved or rearranged. This common target supports future comparison of structural features and learned representations on the same pairs. The trained classifier provides a tool for screening candidate backbone pairs.

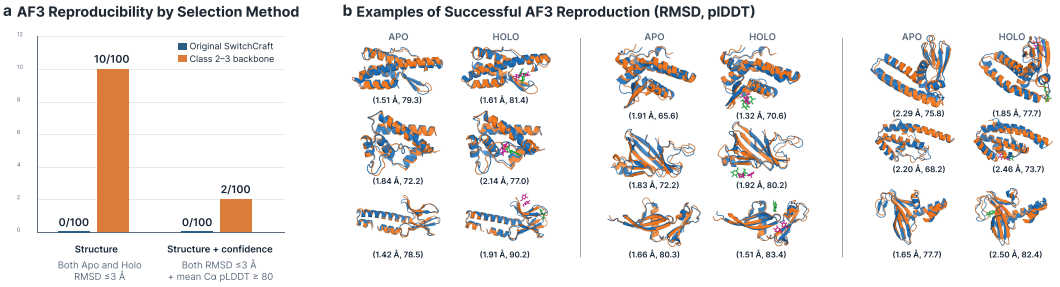

Figure 6: Computational assessment of apo–holo design candidates. (a) Counts for which both AF3 predictions recover their respective targets at Cα RMSD $\leq$ 3 Å, with and without a pLDDT $\geq$ 80 requirement for both states. (b) Example target backbones and AF3 predictions, with per-state RMSD and pLDDT.

In multi-state protein design, StruPair can help screen candidate backbone pairs before sequence design. It can distinguish candidates with too little structural difference from those whose shared structure is reorganized beyond the intended design goal. The preferred class depends on the target change, so a higher class is not always better. The public evaluation set and labeling guidelines also give other researchers a common resource for testing new methods, while the web interface and API make the trained classifier available for reviewing generated pairs.

6.2 Limits of Interpretation

The evaluation set is concentrated in Classes 3 and 4 and differs in composition by source. Overall accuracy can therefore hide errors in rare classes or near class boundaries; estimates for Classes 0 and 1 are especially uncertain. A single ordered class can also summarize distinct kinds of structural change.

Backbone-only input supports use before sequence design, but it omits sequence, side-chain interactions, ligands, environmental conditions, and some inter-chain effects. A class describes the relationship between endpoint structures, not whether or how a transition occurs. The private training corpus limits independent retraining and overlap audits, although the public evaluation set and API support comparisons on a shared task.

6.3 Future Directions

We plan to expand StruPair by incorporating additional rare and borderline cases from a broader range of protein families. The design-candidate experiment shown here assesses sequence design and AF3 backbone structure re-prediction, but these criteria may not fully establish whether the designed protein pairs can functionally interconvert. In future work, we will apply StruPair in our own design campaigns and evaluate through additional computational analyses and experimental characterization.

7 Conclusion

This study presents StruPair for describing and evaluating relationships between structure pairs generated in multi-state protein design. StruPair organizes structural rearrangement and the preservation or reorganization of secondary and tertiary structure into five classes. It also provides a benchmark and a hybrid classifier that predict the five structural categories from backbone structures alone.

The 575-pair public evaluation set and its annotation evidence can support future comparisons of structure-pair methods. In a computational design study, Class 2–3 selection recovered both AF3-predicted target states within 3 Å for 10 of 100 candidates, compared with 0 of 100 unfiltered SwitchCraft candidates. This result supports using the classifier to prioritize backbone pairs for further testing.

8 Data and Code Availability

An anonymous github repository is available at `https://anonymous.4open.science/r/StruPair-24DE`. It summarizes the task and figures; the 575 annotated evaluation pairs and benchmark results have already been released. The 108,054-pair training corpus, model weights, and complete training implementation are not public. The trained classifier API and its full documentation will be made publicly available upon acceptance.

Reproducibility Statement

We will document the API model version used in this paper, the required input coordinates and chain and residue mapping, preprocessing, output classes, and error handling. Access to the service does not reproduce the original training process.

Ethics Statement

This is a computational study based on publicly available protein structure data and on expert judgments of structural relationships. For the 575 pairs in the public evaluation set, we recorded the original source of each structure and verified the redistribution terms of each source.

Annotations were produced by 2 annotators with backgrounds in structural biology and computational protein design. They are authors of this paper. No personal data were collected from annotators, and the released annotations are not linked to individual annotator identities.

StruPair outputs are classifications of the relationship between two endpoint structures. They do not establish transition pathways, rates, or energy barriers, and we do not present them as validation of biological function. We note this because selecting design candidates on the basis of these classes could otherwise be read as evidence of switching behavior that has not been experimentally tested.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016): 493–500, 2024. doi: 10.1038/s41586-024-07487-w. URL https://doi.org/10.1038/s41586-024-07487-w.

Alex Abrudan, Sebastian Pujalte Ojeda, Chaitanya K. Joshi, Matthew Greenig, Felipe Engelberger, Alena Khmelinskaia, Jens Meiler, Michele Vendruscolo, and Tuomas P. J. Knowles. Multistate Protein Design with DynamicMPNN, 2025. URL https://arxiv.org/abs/2507.21938. Preprint version consulted.

T. Amemiya, R. Koike, A. Kidera, and M. Ota. PSCDB: a database for protein structural change upon ligand binding. *Nucleic Acids Research*, 40(D1):D554–D558, 2012. doi: 10.1093/nar/gkr966. URL https://doi.org/10.1093/nar/gkr966.

Takayuki Amemiya, Ryotaro Koike, Sotaro Fuchigami, Mitsunori Ikeguchi, and Akinori Kidera. Classification and Annotation of the Relationship between Protein Structural Change and Ligand Binding. *Journal of Molecular Biology*, 408(3):568–584, 2011. doi: 10.1016/j.jmb.2011.02.058. URL https://doi.org/10.1016/j.jmb.2011.02.058.

Magnus S. Bauer, Jason Z. Zhang, Kejia Wu, Gyu Rie Lee, Brian Coventry, Kody A. Klupt, Jiuhan Shi, Rafael I. Brent, Xinting Li, Carolina Moller, Nicole Roullier, Dionne K. Vafeados, Indrek Kalvet, Rebecca K. Skotheim, Siyu Zhu, Amir Motmaen, Luca C. Herrmann, Pascal Sturmfels, Doug Tischer, Han Raut Altae-Tran, David Juergens, Rohith Krishna, Woody Ahern, Jason Yim, Asim K. Bera, Alex Kang, Emily Joyce, Andrew Lu, Lance Stewart, Frank DiMaio, and David Baker. De novo design of phospho-tyrosine peptide binders. *bioRxiv*, pp. 2025.09.29.678898, 2025. doi: 10.1101/2025.09.29.678898. URL https://doi.org/10.1101/2025.09.29.678898.

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

Jasper Kenneth Veje Butcher, Rohith Krishna, Raktim Mitra, Rafael Isaac Brent, Yanjing Li, Nathaniel Corley, Paul Kim, Jonathan Funk, Simon Valentin Mathis, Saman Salike, Aiko Muraishi, Helen Eisenach, Tuscan Rock Thompson, Jie Chen, Yuliya Politanska, Enisha Sehgal, Brian Coventry, Odin Zhang, Bo Qiang, Kieran Didi, Maxwell Kazman, Frank DiMaio,

and David Baker. De novo Design of All-atom Biomolecular Interactions with RFdiffusion3. *bioRxiv*, pp. 2025.09.18.676967, 2025. doi: 10.1101/2025.09.18.676967. URL https://www.biorxiv.org/content/10.1101/2025.09.18.676967v1.

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

Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K. Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/2b44928ae11fb9384c4cf38708677c48-Abstract-round2.html.

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

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

Joseph I. J. Ellaway, Stephen Anyango, Sreenath Nair, Hossam A. Zaki, Nurul Nadzirin, Harold R. Powell, Aleksandras Gutmanas, Mihaly Varadi, and Sameer Velankar. Identifying protein conformational states in the Protein Data Bank: Toward unlocking the potential of integrative dynamics studies. *Structural Dynamics*, 11(3), 2024. doi: 10.1063/4.0000251. URL https://doi.org/10.1063/4.0000251.

Christos P Feidakis, Radoslav Krivak, David Hoksza, and Marian Novotny. AHoJ: rapid, tailored search and retrieval of apo and holo protein structures for user-defined ligands. *Bioinformatics*, 38 (24):5452–5453, 2022. doi: 10.1093/bioinformatics/btac701. URL https://doi.org/10.1093/bioinformatics/btac701.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

Mark Gerstein and Werner Krebs. A database of macromolecular motions. *Nucleic Acids Research*, 26(18):4280–4290, 1998. doi: 10.1093/nar/26.18.4280. URL https://doi.org/10.1093/nar/26.18.4280.

Mark Gerstein, Arthur M. Lesk, and Cyrus Chothia. Structural Mechanisms for Domain Movements in Proteins. *Biochemistry*, 33(22):6739–6749, 1994. doi: 10.1021/bi00188a001. URL https://doi.org/10.1021/bi00188a001.

Amy B. Guo, Deniz Akpinaroglu, Christina A. Stephens, Michael Grabe, Colin A. Smith, Mark J. S. Kelly, and Tanja Kortemme. Deep learning–guided design of dynamic proteins. *Science*, 388(6749), 2025. doi: 10.1126/science.adr7094. URL https://doi.org/10.1126/science.adr7094.

Zhong Guo, Oleh Smutok, Gyu Rie Lee, Zhenling Cui, Haocheng Qianzhu, Monika Kish, Cagla Ergun Ayva, Kejia Wu, Roxane Mutschler, Colin J. Jackson, Maria M. Fiorito, Andrew C. Warden, Oliver B. Smith, Alfredo Quijano-Rubio, Thomas Huber, Jonathan J. Phillips, Gottfried Otting, Evgeny Katz, David Baker, and Kirill Alexandrov. Artificial allosteric protein switches with machine-learning-designed receptors. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03081-9. URL https://doi.org/10.1038/s41587-026-03081-9.

Steven Hayward and Herman J.C. Berendsen. Systematic analysis of domain motions in proteins from conformational change: New results on citrate synthase and T4 lysozyme. *Proteins: Structure, Function, and Genetics*, 30(2):144–154, 1998. doi: 10.1002/(sici)1097-0134(19980201)30:2⟨144::aid-prot4⟩3.0.co;2-n. URL https://doi.org/10.1002/(sici)1097-0134(19980201)30:2%3C144::aid-prot4%3E3.0.co;2-n.

Bowen Jing, Mihir Bafna, Anisha Parsan, Heyuan Michael Ni, David Kwabi-Addo, Bryan Bryson, Adam Klivans, and Bonnie Berger. SwitchCraft: A Programmatic Framework for Designing State-Switching Proteins. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.31236. Author version available on arXiv.

W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. doi: 10.1107/s0567739476001873. URL https://doi.org/10.1107/s0567739476001873.

W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 34(5):827–828, 1978. doi: 10.1107/s0567739478001680. URL https://doi.org/10.1107/s0567739478001680.

Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. doi: 10.1002/bip.360221211. URL https://doi.org/10.1002/bip.360221211.

Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S. Morey-Burrows, Ivan Anishchenko, Ian R. Humphreys, Ryan McHugh, Dionne Vafeados, Xinting Li, George A. Sutherland, Andrew Hitchcock, C. Neil Hunter, Alex Kang, Evans Brackenbrough, Asim K. Bera, Minkyung Baek, Frank DiMaio, and David Baker. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, 384 (6693):eadl2528, 2024. doi: 10.1126/science.adl2528. URL https://doi.org/10.1126/science.adl2528.

Robert A. Langan, Scott E. Boyken, Andrew H. Ng, Jennifer A. Samson, Galen Dods, Alexandra M. Westbrook, Taylor H. Nguyen, Marc J. Lajoie, Zibo Chen, Stephanie Berger, Vikram Khipple Mulligan, John E. Dueber, Walter R. P. Novak, Hana El-Samad, and David Baker. De novo design of bioactive protein switches. *Nature*, 572(7768):205–210, 2019. doi: 10.1038/s41586-019-1432-8. URL https://doi.org/10.1038/s41586-019-1432-8.

Andrew Leaver-Fay, Ron Jacak, P. Benjamin Stranges, and Brian Kuhlman. A Generic Program for Multistate Protein Design. *PLoS ONE*, 6(7):e20937, 2011. doi: 10.1371/journal.pone.0020937. URL https://doi.org/10.1371/journal.pone.0020937.

Gyu Rie Lee, Samuel J. Pellock, Christoffer Norn, Doug Tischer, Justas Dauparas, Ivan Anishchenko, Jaron A. M. Mercer, Alex Kang, Asim K. Bera, Hannah Nguyen, Evans Brackenbrough, Banumathi Sankaran, Inna Goreshnik, Dionne Vafeados, Nicole Roullier, Hannah L. Han, Brian Coventry, Hugh K. Haddox, David R. Liu, Andy Hsien-Wei Yeh, and David Baker. Small-molecule binding and sensing with a designed protein family. *Nature Communications*, 17:4533, 2026. doi: 10.1038/s41467-026-70953-8. URL https://doi.org/10.1038/s41467-026-70953-8.

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013. doi: 10.1093/bioinformatics/btt473. URL https://doi.org/10.1093/bioinformatics/btt473.

Antonio Mirarchi, Toni Giorgino, and Gianni De Fabritiis. mdCATH: A Large-Scale MD Dataset for Data-Driven Computational Biophysics. *Scientific Data*, 11(1), 2024. doi: 10.1038/s41597-024-04140-z. URL https://doi.org/10.1038/s41597-024-04140-z.

Alexander Miguel Monzon, Cristian Oscar Rohr, María Silvina Fornasari, and Gustavo Parisi. CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database*, 2016:baw038, 2016. doi: 10.1093/database/baw038. URL https://doi.org/10.1093/database/baw038.

Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3): 443–453, 1970. doi: 10.1016/0022-2836(70)90057-4. URL https://doi.org/10.1016/0022-2836(70)90057-4.

Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Rose Orenbuch, Yarin Gal, and Debora S. Marks. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/cac723e5ff29f65e3fcbb0739ae91bee-Paper-Datasets_and_Benchmarks.pdf.

Stefano Piana, Kresten Lindorff-Larsen, and David E. Shaw. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophysical Journal*, 100(9):L47–L49, 2011. doi: 10.1016/j.bpj.2011.03.051. URL https://doi.org/10.1016/j.bpj.2011.03.051.

Arvind Pillai, Abbas Idris, Annika Philomin, Connor Weidle, Rebecca Skotheim, Philip J. Y. Leung, Adam Broerman, Cullen Demakis, Andrew J. Borst, Florian Praetorius, and David Baker. De novo design of allosterically switchable protein assemblies. *Nature*, 632(8026): 911–920, 2024. doi: 10.1038/s41586-024-07813-2. URL https://doi.org/10.1038/s41586-024-07813-2.

Damiano Piovesan, Marco Necci, Nahuel Escobedo, Alexander Miguel Monzon, András Hatos, Ivan Mičetić, Federica Quaglia, Lisanna Paladin, Pathmanaban Ramasamy, Zsuzsanna Dosztányi, Wim F Vranken, Norman E Davey, Gustavo Parisi, Monika Fuxreiter, and Silvio C E Tosatto. MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Research*, 49(D1):D361–D367, 2021. doi: 10.1093/nar/gkaa1058. URL https://doi.org/10.1093/nar/gkaa1058.

Lauren L. Porter and Loren L. Looger. Extant fold-switching proteins are widespread. *Proceedings of the National Academy of Sciences*, 115(23):5968–5973, 2018. doi: 10.1073/pnas.1800168115. URL https://doi.org/10.1073/pnas.1800168115.

Florian Praetorius, Philip J. Y. Leung, Maxx H. Tessmer, Adam Broerman, Cullen Demakis, Acacia F. Dishman, Arvind Pillai, Abbas Idris, David Juergens, Justas Dauparas, Xinting Li, Paul M. Levine, Mila Lamb, Ryanne K. Ballard, Stacey R. Gerben, Hannah Nguyen, Alex Kang, Banumathi Sankaran, Asim K. Bera, Brian F. Volkman, Jeff Nivala, Stefan Stoll, and David Baker. Design of stimulus-responsive two-state hinge proteins. *Science*, 381(6659):754–760, 2023. doi: 10.1126/science.adg7731. URL https://doi.org/10.1126/science.adg7731.

Zhuoran Qiao, Feizhi Ding, Thomas Dresselhaus, Mia A. Rosenfeld, Xiaotian Han, Owen Howell, Aniketh Iyengar, Stephen Opalenski, Anders S. Christensen, Sai Krishna Sirumalla, Frederick R. Manby, Thomas F. Miller, III, and Matthew Welborn. NeuralPLexer3: Accurate Biomolecular Complex Structure Prediction with Flow Models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0001. URL https://papers.neurips.cc/paper_files/paper/2025/hash/0010031a1b4910aa67edbda26a705518-Abstract-Conference.html.

Federica Quaglia, Anastasia Chasapi, Maria Victoria Nugnes, Maria Cristina Aspromonte, Emanuela Leonardi, Damiano Piovesan, and Silvio C E Tosatto. Best practices for the manual curation of intrinsically disordered proteins in DisProt. *Database*, 2024, 2024. doi: 10.1093/database/baae009. URL https://doi.org/10.1093/database/baae009.

Alfredo Quijano-Rubio, Hsien-Wei Yeh, Jooyoung Park, Hansol Lee, Robert A. Langan, Scott E. Boyken, Marc J. Lajoie, Longxing Cao, Cameron M. Chow, Marcos C. Miranda, Jimin Wi, Hyo Jeong Hong, Lance Stewart, Byung-Ha Oh, and David Baker. De novo design of modular and tunable protein biosensors. *Nature*, 591(7850):482–487, 2021. doi: 10.1038/s41586-021-03258-z. URL https://doi.org/10.1038/s41586-021-03258-z.

Kaiwen Shi and Carlos Oliver. ENSEMBITS: an alphabet of protein conformational ensembles, 2026. URL https://arxiv.org/abs/2605.13789. Preprint.

Richard W. Shuai, Tianyu Lu, Subhang Bhatti, Petr Kouba, and Po-Ssu Huang. Ensemble-conditioned protein sequence design with Caliby. *bioRxiv*, pp. 2025.09.30.679633, 2025. doi: 10.1101/2025.09.30.679633. URL https://www.biorxiv.org/content/10.1101/2025.09.30.679633v1.

Michael Sun, Weize Yuan, Gang Liu, Wojciech Matusik, and Marinka Zitnik. Protein Structure Tokenization via Geometric Byte Pair Encoding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/404df2480b6eef0486a1679e371894b0-Abstract-Conference.html.

Raphael J. L. Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander S. Powers, Yianni Laloudakis, Sidhika Balachandar, Bowen Jing, Brandon Anderson, Stephan Eismann, Risi Kondor, Russ B. Altman, and Ron O. Dror. ATOM3D: Tasks on Molecules in Three Dimensions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c45147dee729311ef5b5c3003946c48f-Abstract-round1.html.

Robbyn L. Tuinstra, Francis C. Peterson, Snjezana Kutlesa, E. Sonay Elgin, Michael A. Kron, and Brian F. Volkman. Interconversion between two unrelated protein folds in the lymphotactin native state. *Proceedings of the National Academy of Sciences*, 105(13):5057–5062, 2008. doi: 10.1073/pnas.0709518105. URL https://doi.org/10.1073/pnas.0709518105.

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2):243–246, 2024. doi: 10.1038/s41587-023-01773-0. URL https://doi.org/10.1038/s41587-023-01773-0.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023. doi: 10.1038/s41586-023-06415-8. URL https://doi.org/10.1038/s41586-023-06415-8.

wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, 47(D1):D520–D528, 2019. doi: 10.1093/nar/gky949. URL https://doi.org/10.1093/nar/gky949.

Chengxin Zhang, Morgan Shine, Anna Marie Pyle, and Yang Zhang. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature Methods*, 19(9):1109–1115, 2022. doi: 10.1038/s41592-022-01585-1. URL https://doi.org/10.1038/s41592-022-01585-1.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004. doi: 10.1002/prot.20264. URL https://doi.org/10.1002/prot.20264.

A Dataset Collection and Preprocessing

A.1 Source Provenance

Experimental structure pairs were collected from resources including AHoJ (Feidakis et al., 2022), ConfBench (Qiao et al., 2025), PDBe (Ellaway et al., 2024), and PSCDB (Amemiya et al., 2012), as well as from individual experimental publications. Among these data, internal experimental pairs integrating resources such as AHoJ and ConfBench are managed under the name Apo–Holo. Experimental coordinates were obtained from the PDB (Berman et al., 2000; wwPDB consortium, 2019) and linked to the structure IDs, chains, and state information provided by each source. When the same structure pair appeared in multiple sources, duplicate entries were removed while information from all associated sources was retained.

MD-derived structure pairs were constructed by pairing structures from different time points in mdCATH simulation trajectories (Mirarchi et al., 2024). Computationally designed structure pairs were obtained from SwitchCraft outputs (Jing et al., 2026). Although the original SwitchCraft data include designed sequences, our classifier does not use amino acid identities or sequence embeddings as input. We also distinguish structure and state information provided by the original sources from the training labels and human Class 0–4 annotations generated in this study.

A.2 Coordinate Extraction and Residue Correspondence

For experimental structures in mmCIF format, we selected the target chain using `label_asym_id` and retained residue positions in the full sequence using `label_seq_id`. We extracted the backbone atoms N, Cα, C, and O from the first structural model. In the existing coordinate cache and graph construction, we used the coordinates with the highest occupancy among atoms whose alternate-location identifier was blank or A. Unobserved coordinates were retained as missing. The selected experimental chains were converted into single-chain PDB files that preserved residue positions and were used for graph construction.

Residue correspondences were determined according to the data source and processing stage. Apo–Holo structure pairs with identical full sequences were matched using standardized sequence positions, and structural metrics for mdCATH frames were computed using the same trajectory residue indices. Quality checks and existing structural metric calculations for PDBe, PSCDB, and individual literature data used global alignments of the full sequences. These alignments were performed with Biopython PairwiseAligner (Cock et al., 2009), using a match score of $+2$, a mismatch score of -1, a gap-opening score of -10, and a gap-extension score of -0.5. The existing GNN inputs used sequence-position correspondence for Apo–Holo and Needleman–Wunsch alignment (Needleman & Wunsch, 1970) of the observed residue sequences for the other data sources. The latter used a match score of $+2$, a mismatch score of -1, and a gap penalty of -2 per residue. For each processing route, comparisons used matched residues for which Cα coordinates were observed in both structures.

A.3 Structural Superposition and Quality Filtering

RMSD and per-residue displacements were computed using Kabsch rigid-body superposition (Kabsch, 1976; 1978) over all corresponding Cα coordinates. Rotation and translation were optimized to minimize the sum of squared distances between corresponding coordinates, without allowing reflection. For these global measurements, we neither superposed domains separately nor iteratively removed residues with large displacements. The GNN distance and contact features were computed from coordinates within each structure and therefore did not require spatial superposition of the two structures.

Quality checks were based on file and chain validity, residue correspondence, the fraction of observed residues, and backbone completeness. For the Apo–Holo coordinate cache, the two structures were required to have identical full sequences, and the fraction of residues with Cα coordinates observed in both structures relative to the full sequence was required to be at least 0.95. In addition, the fraction of residues for which all four backbone atoms N, Cα, C, and O were observed in each structure was required to be at least 0.95. For data processed through direct PDB reading, including mdCATH and SwitchCraft, the sequence identity between the two structures was required to be at

least 0.95, the matched-residue fraction relative to the number of observed residues in both structures was required to be at least 0.95, and backbone atom completeness over observed residues was also required to be at least 0.95. For candidate structure pairs collected from PDBe, PSCDB, and individual publications, the identity of the full-sequence alignment was required to be at least 0.95, and the matched-residue fraction relative to the full sequence length of each structure was required to be at least 0.95. In addition, the observed-residue fraction in each chain was required to be at least 0.80, and each chain was required to contain at least 40 residues.

The existing cache-based structural metric calculations excluded pairs with fewer than three commonly observed Cα atoms or a common Cα observation fraction below 0.85 within the corresponding region. For graph construction, the minimum input requirements were a sequence identity of at least 0.80 and at least eight corresponding residues. These requirements were applied separately from the dataset-level quality-control criteria described above.

A.4 Deduplication and Related Samples

Duplicates among internal experimental structure pairs were identified using the PDB IDs and chain identifiers of the two endpoints. Pairs that differed only in A/B order and identical pairs present in multiple sources were merged, while information from all sources was retained. Related samples sharing the same protein, construct, endpoint structure, MD trajectory, or design scaffold were managed as connected groups and used during dataset splitting. Structurally similar endpoints were searched using Foldseek (van Kempen et al., 2024), and the nearest internal structures to external evaluation candidates were rechecked using US-align (Zhang et al., 2022). This structural search was used to examine similarity and split relationships between samples and is distinct from residue correspondence and Kabsch superposition within an individual structure pair.

A.5 Planned Evaluation Set Release

The released data will be accompanied by a datasheet following the structure proposed by Gebru et al. (2021), covering motivation, composition, collection process, preprocessing, uses, and distribution.

The 575 structure pairs used for evaluation will be released together with their annotations and source information. The public data will include the original structure IDs, chain information, residue correspondences, and annotations for each pair.

B Metrics and Judgment Criteria

The five classes summarize how much of the common structural organization remains and how preserved parts move. Class 4 describes small differences with the common fold and relative arrangement maintained. Class 3 describes a coherent movement of mostly preserved structural units. Class 2 describes a substantial common scaffold together with more complex rearrangement or local deformation. Class 1 describes broad reconstruction with some recognizable substructure retained. Class 0 describes a pair in which most common fold organization is lost or replaced. These are ordered descriptions, not equally spaced physical measurements or ranges of RMSD.

Order of judgment. First establish the chain, region, and residue correspondence and identify missing coordinates. Then inspect whether a shared fold, domain, or core remains; an isolated matching helix is not enough to establish a preserved scaffold. Next distinguish movement of intact units from deformation within those units. Finally decide whether the difference is a coherent change of state or only a small residual variation, and record the reason for choosing one of two neighboring classes.

Class 2 is a structural category, not an uncertainty bin. We apply no fixed 50% preservation rule. RMSD, TM-based scores, contact measures, displacement distributions, and approximate secondary-structure measures provide supporting evidence rather than automatic cutoffs. Global and local comparisons should be read together: a large global displacement can arise from movement of an internally preserved domain. Contact changes should distinguish contacts between domains from contacts within a domain. Contact rewiring defined as one minus contact Jaccard is a transformation

Table 2: Questions used to distinguish adjacent classes. No single numerical threshold defines a boundary.

Boundary	Main question
3 / 4	Is there a coherent rearrangement of a region or unit, beyond small residual variation, even if global RMSD is small?
2 / 3	After accounting for motion of preserved units, is there substantial internal deformation or more complex structural reorganization?
1 / 2	Is a substantial common scaffold rearranged, or is the shared architecture broadly rebuilt?
0 / 1	Does a meaningful multi-element substructure remain, rather than only isolated matching fragments?

of the same contact information, not independent evidence. A TM-like feature need not be an official TM-align score, and the backbone-angle helix–sheet proxy is not a DSSP assignment(Kabsch & Sander, 1983).

Missing coordinates are not zero displacement or evidence of unfolding. If correspondence, coverage, or structure quality is inadequate, the pair needs a data-quality decision before a class decision. Endpoint classes also do not specify a physical transition path; a linear interpolation between coordinates is only a visualization and does not demonstrate an accessible transition.

C Model Reproducibility Settings

Appendix C describes two separate models: the hybrid release classifier (Appendix C.1), which uses seven global features with a pair graph, and the feature-set readouts used for the Section 4.2 comparison (Appendix C.2), which use the 54- and 148-feature sets of Table 3.

C.1 Hybrid StruPair Classifier (Release Model)

Input features and graph. The seven global features are the number of matched residues, Cα RMSD, a TM-like score, contact Jaccard, contact rewiring, an angle-based helix–sheet change proxy, and the 90th percentile of matched-residue displacement. The TM-like score is computed from post-alignment displacements and is not an official TM-align score. Contacts use an 8-Å cutoff for matched Cα pairs separated by at least three sequence positions in both structures; contact rewiring is one minus Jaccard and therefore carries the same contact information. The helix–sheet proxy uses backbone angles, not DSSP assignments(Kabsch & Sander, 1983). Missing proxy values use the median saved with the release preprocessing configuration.

Each graph node represents a matched residue with 12 geometric and validity channels. Directed edges are formed from the union of each endpoint's 24 nearest neighbors and sequential neighbors. Their 15 channels summarize endpoint distances and differences, 8-Å contact state and change, sequence separation, and radial-basis distances. The node encoder maps 12 channels to 32, followed by two 32-channel message-passing blocks. Mean, standard-deviation, top-decile-mean, and attention pooling form a 128-dimensional graph summary. The seven global features pass through a 7-to-16 branch; the combined 144-dimensional vector passes through a 144-to-48-to-4 ordinal head. The model has 27,637 trainable parameters. From head outputs $r_0, \ldots, r_3$, cumulative logits satisfy $\ell_0 = r_0$ and $\ell_k = r_0 - \sum_{j=1}^{k} \mathrm{softplus}(r_j)$ for $k = 1, 2, 3$. With $q_k = \sigma(\ell_k)$, the five probabilities are $(1 - q_0, q_0 - q_1, q_1 - q_2, q_2 - q_3, q_3)$. Training minimizes the negative log probability of the target class.

Atomic input and preprocessing. The release input is a specified chain from a PDB or mmCIF structure. The first structural model is used; mmCIF chains are selected by `label_asym_id`. Deposited polymer sequence and `label_seq_id` define mmCIF residue positions, while observed residues define PDB correspondence. Global sequence alignment uses match $+2$, mismatch -1, gap-open -10, and gap-extension -0.5. Input pairs require 8–5,000 matched Cα residues and at least 0.8 sequence identity in the matched region. Missing coordinates are not generated. Backbone continuity checks residue positions, chain breaks, a 2.5–4.5-Å Cα step, and, when C and N are present, a 0.9–1.8-Å peptide step. Spatial nearest-neighbor edges do not imply peptide continuity.

Table 3: Raw geometric feature families for Classifier A (small) and Classifier B (large). A has 54 features; B adds 94, giving 148 in total.

Classifier	Family	Count	Examples
A	Basic size and global comparison	5	Matched-residue counts, RMSD, TM-like score, P90
A	Global displacement distribution	12	Mean, spread, percentiles, interquartile range
A	High-displacement tails	15	Top 5–50% residue summaries and squared-displacement share
A	Displacement threshold fractions	6	Fractions above 0.5, 1, 2, 4, 8, and 12 Å
A	Similarity summaries	5	GDT proxies, TM-like and official TM-score summaries
A	Global shape	11	Gyration eigenvalues, radius of gyration, anisotropy
B	Secondary-structure change proxy	1	Helix–sheet flip fraction
B	Preserved core	10	RMSD of the least-displaced 50–95% of residues
B	Changed-segment concentration	6	Longest run above displacement thresholds
B	Observed backbone continuity	1	Fraction of connected adjacent matched residues
B	Local windows	32	Independently aligned 5-, 9-, 17-, and 33-residue windows
B	Backbone angle changes	44	Circular differences in ϕ, ψ, ω, and Cα pseudo-dihedral

The final model's preprocessing statistics were re-estimated from the complete release list and saved with the model configuration.

Implementation checks and availability. The recorded environment was Python 3.12.14, PyTorch 2.6.0+cu124, CUDA 12.4, NumPy 2.3.3, SciPy 1.16.2, pandas 2.3.3, scikit-learn 1.7.2, Biopython 1.81, and Gemmi 0.7.5. Checks covered raw-structure-to-input agreement on 12 samples, A/B endpoint exchange, rigid transformations, invalid-chain rejection, missing connectivity, and CPU/GPU and batch inference. The training corpus, model weights, and complete training implementation are not part of the planned public release, so external retraining from the appendix alone is not possible.

C.2 Structural-Representation Comparison

The comparison in Section 4.2 used the frozen September 22, 2026 label version of 575 pairs (Class 0–4 counts: 21, 26, 44, 245, 239). Its development list contained 108,054 pairs. The same 91,384 training pairs and 16,670 validation pairs were used for each readout. Validation was split by leakage component into 7,904 pairs for early stopping and 8,766 for setting selection. Pair IDs, recorded leakage-component IDs, and exact endpoint-sequence hashes were checked across the splits. The earlier sequence exclusion was retained; no new exhaustive structural-similarity search was run for this comparison.

The selected 54-feature conventional classifier was LightGBM with 31 leaves; the selected combined classifier was LightGBM with 127 leaves. Their seed-17/29/43 stopping rounds were 113/140/90 and 118/145/193, respectively. GeoBPE used a 256–128 MLP (dropout 0.1, weight decay 0.001) with selected epochs 2/6/10; ENSEMBITS used a 512–256–128 MLP (dropout 0.2, weight decay 0.01) with epochs 4/6/4. The LightGBM learning rate was 0.04; the MLP optimizer was AdamW at learning rate 0.001. The selected setting for each readout was chosen by five-class macro-F1 on the setting-selection split, with log-loss to break ties. Three seed probabilities were averaged before the final class prediction. The public ENSEMBITS corpus shared PDB IDs with 113 evaluation pairs; this overlap alone does not establish that those pairs were in the encoder's pretraining split.

D DETAILED EVALUATION RESULTS

Table 4: Confusion matrix and per-class performance of the hybrid classifier on the 575-pair evaluation set. Rows are gold labels and the first five numeric columns are predicted classes; diagonal entries are in bold. Precision, recall, and F_1 are computed from the same matrix. Intervals are 95% bootstrap confidence intervals over 10,000 resamples of the 575 pairs.

	Predicted class					Per-class performance				
Gold	0	1	2	3	4	n	Prec.	Rec.	F_1	95% CI
Class 0	**20**	1	0	0	0	21	1.0000	0.9524	0.9756	0.9143–1.0000
Class 1	0	**26**	0	0	0	26	0.9286	1.0000	0.9630	0.9020–1.0000
Class 2	0	1	**43**	0	0	44	0.9348	0.9773	0.9556	0.9048–0.9905
Class 3	0	0	3	**227**	15	245	0.9913	0.9265	0.9578	0.9381–0.9751
Class 4	0	0	0	2	**237**	239	0.9405	0.9916	0.9654	0.9483–0.9805
Total	20	28	46	229	252	575	Macro-F_1 = 0.9635			