Appendix 1:

Prompt for agent to propose the PPI change criterion:

1. the prompt for agent to classify the PPI change to neo, hypo and native PPI with literature provided:

You are a bioinformatics assistant, based on the provided paper, classify single-point mutations in protein-protein complexes into one of three PPI perturbation categories. Classification categories: Neomorphic PPI: The mutation leads to a new binding mode and potentially increases protein–protein interactions. Hypomorphic PPI: The mutation disrupts the original interaction and may result in loss of function. Native PPI: The mutation causes no significant perturbation at the interface and has minimal impact on the interaction. Note: If the paper does not provide direct evidence about the mutation’s effect on the protein–protein interaction interface, explicitly state that. If the paper provides binding data or other experiment data, carefully check whether for the complex internal interaction or interaction of binding partners like substrate. Please answer in English, giving for each mutation:

1. the prompt for agent to define interaction change criterion:

You are a bioinformatics assistant. Explain how to determine whether a mutation changes a protein–protein interaction, considering interface residues, hydrogen bonds, salt bridges, and ΔpDockQ. Suggest complementary metrics if needed. Answer in English.

1. the prompt for agent to collect data for PPI analysis:

You are a bioinformatics assistant. Describe how to collect and curate wild-type and mutant protein–protein complexes for mutation impact analysis. Include structure retrieval, filtering, WT–MUT pairing, and optional predicted structures (e.g., AlphaFold2). Answer in English.

Appendix 2:

We have collect the output of PIPA and we show that there have some hallucination that highlight with human scientist correction.

1. Hallucination or Misclassification Cases

The comparison reveals two main types of discrepancies, totaling 27 cases:

(1)Fabricating or Lacking Supporting Evidence: PIPA inferred PPI effects without direct experimental support from the original papers.

(2)Misinterpreting Evidence and Functional Context: PIPA misclassified mutations by misinterpreting structural or functional data, often conflating catalytic or allosteric effects with true PPI changes.

Below we list the 27 mutation cases grouped into these two categories. Each entry includes the PDB pair, mutation, PIPA’s classification, the researcher’s correction, and a brief note on the discrepancy.

Table A2.1: Fabri

cating or Lacking Supporting Evidence

| **WT\_PDB** | **MUT\_PDB** | **Mutation** | **PIPA\_Class** | **Researcher\_Note** | **Discrepancy Type** |
| --- | --- | --- | --- | --- | --- |
| 1V1R | 1V1M | ASP295ALA | Hypomorphic | No direct PPI evidence in paper | Lack of support |
| 1WCI | 2BFB | GLN112ALA | – | No PPI impact discussed | Lack of support |
| 1WCI | 2BFD | TYR113ALA | – | No PPI impact discussed | Lack of support |
| 2GRN | 2GRO | ASN85GLN | Hypomorphic | No direct PPI evidence | Lack of support |
| 2GRN | 2GRP | TYR87ALA | Hypomorphic | No direct PPI evidence | Lack of support |
| 2GRN | 2GRQ | ASP127ALA | Hypomorphic | No direct PPI evidence | Lack of support |
| 2GRN | 2GRR | ASP127SER | Native | No direct PPI evidence | Lack of support |
| 2XBW | 2XBV | ARG150GLU | – | Mutation for crystallization only | Fabricated inference |
| 3DZ2 | 3DZ3 | PHE223ALA | Hypomorphic | Alters active site, not PPI | Misattributed effect |
| 3EP9 | 3EP3 | ASP174ASN | Hypomorphic | Alters active site, not PPI | Misattributed effect |
| 3EP9 | 3EP4 | GLU256GLN | Hypomorphic | Alters active site, not PPI | Misattributed effect |
| 3EP9 | 3EP5 | GLU178GLN | Hypomorphic | Alters active site, not PPI | Misattributed effect |
| 5IJ0 | 5IJ9 | ASP417HIS | Hypomorphic | No binding constant for tubulin-tubulin interface | Lack of support |
| 6X5J | 6X5P | ARG153ALA | – | No PPI data provided | Lack of support |
| 6XHB | 6XI7 | CYS118SER | Native | No direct PPI evidence | Lack of support |
| 8ROJ | 8ROI | GLU1144GLN | – | No PPI evidence provided | Lack of support |

Table A2.1 catalogs 16 instances where PIPA's classifications lacked direct experimental support. The primary error was a Lack of Supporting Evidence (12 cases), where the original publications did not discuss or provide data on PPI changes. In 4 cases, PIPA Misattributed effects on catalytic activity or ligand binding (active site alterations) to PPI changes. One case involved a Fabricated Inference, where a mutation introduced for crystallization purposes was incorrectly assigned a PPI effect. This highlights a tendency to generate unsupported conclusions.

Table A2.2: Misinterpreting Evidence and Functional Context

| **WT\_PDB** | **MUT\_PDB** | **Mutation** | **PIPA\_Class** | **Researcher\_Note** | **Discrepancy Type** |
| --- | --- | --- | --- | --- | --- |
| 1U5B | 1X7W | SER292GLN | Hypomorphic | Classified based on E1b-E2b interface, not internal E1b effect | Context error |
| 1U5B | 1X7X | SER292GLU | Hypomorphic | Same as above | Context error |
| 1U5B | 1X7Y | SER292ASN | Native | Same as above | Context error |
| 1U5B | 1X7Z | SER292ASP | Hypomorphic | Same as above | Context error |
| 1WCI | 2BFC | TYR113PHE | Neomorphic | No significant effect on substrate/LBD affinity; not neo | Overinterpretation |
| s5D1K | 5D1L | TYR165ALA | Hypomorphic | Kd tightens; new H-bonds; should be neo | Misclassification |
| 5D1K | 5D1M | PRO199ALA | Native | Minimal interface perturbation; should be native | Misclassification |
| 5GRE | 5GRF | LYS151ALA | Hypomorphic | Interface intact; no structural communication loss | Overinterpretation |
| 6LYY | 7DA5 | ASP309ASN | Native | Interface preserved; function lost due to conformational trap | Context error |
| 6XHB | 6XGU | GLN61ARG | Hypomorphic | Interface intact; affinity comparable to WT | Overinterpretation |
| 8TS7 | 8TS8 | HIS1047ARG | Hypomorphic | Weakened interface; correctly classified | Minor discrepancy |

Table A2.2 summarizes 11 cases where PIPA misinterpreted available evidence. A common error was Misinterpreted Context (5 cases), where PIPA applied PPI classification to mutations affecting internal protein dynamics or allostery rather than the intermolecular interface. Overinterpretation of Data occurred in 4 cases, where PIPA inferred PPI disruption from functional loss despite structural evidence showing an intact interface. There were 2 clear Misclassifications where PIPA's label contradicted biophysical evidence (e.g., tightened binding). One case was a Minor Discrepancy in wording. These errors underscore the challenge of distinguishing direct interface perturbations from indirect functional consequences.

1. Correct or Consistent Prediction Cases

In addition to the 27 hallucination cases, we identified 5 instances where the PIPA agent made correct or consistent predictions. These cases reflect the model's accuracy in certain situations.

Table A2.3: Correct or Consistent Predictions

| **WT\_PDB** | **MUT\_PDB** | **Mutation** | **PIPA\_Class** | **Researcher\_Note** | **Discrepancy Type** |
| --- | --- | --- | --- | --- | --- |
| 3KW5 | 3IFW | SER18TYR | Native | PIPA and researcher agree. The mutation does not significantly perturb the PPI interface. | Correct Classification |
| 3KW5 | 3KVF | ILE93MET | Native | PIPA and researcher agree. The PPI interface was not perturbed. | Correct Classification |
| 6EC0 | 6E2J | SER233LEU | Neomorphic | PIPA and researcher agree. The mutation mediates an aberrant interaction between tetramers by creating a new hydrophobic patch. | Correct Classification |
| 1OLS | 1OLU | HIS291ALA | Hypomorphic | Not discussed | Lack of Support |
| 1OLS | 1OLX | HIS146ALA | Hypomorphic | Not discussed | Lack of Support |

Table A2.3 contains 5 cases where PIPA's output was found to be either correct or consistent with the human scientist's findings. Three of these were Correct Classifications, where PIPA accurately identified the PPI type based on available evidence, reinforcing its capability to process and interpret structural and functional data. The final two cases, HIS291ALA and HIS146ALA, are categorized as having a "Lack of Support" in the original literature, a nuance PIPA's output did not capture. While PIPA's classification was technically an inference, the output was not a clear misinterpretation of existing data, and thus these cases are distinct from the hallucinations found in Tables A2.1 and A2.2.

Appendix 3:

Rationale and Definition of the Interface Conservation Index (ICI)}

A critical challenge in assessing protein–protein interaction (PPI) perturbations is that traditional metrics, such as ΔpDockQ, primarily quantify predicted changes in binding affinity but fail to fully capture structural rearrangements at the interface. Subtle perturbations, including loss or gain of interface residues, hydrogen bonds, or salt bridges, may not be reflected in pDockQ scores, potentially underestimating the functional impact of mutations.

To address this limitation, we introduced the Interface Conservation Index (ICI) as a complementary metric. The ICI quantifies the extent to which physical contacts at a PPI interface are preserved after a mutation, integrating multiple structural features into a single measure. The ICI was conceived based on the hypothesis that interface integrity, rather than binding affinity alone, is a key determinant of functional impact. By integrating structural and energetic information at the interface, ICI provides a more comprehensive assessment of mutation-induced PPI changes.

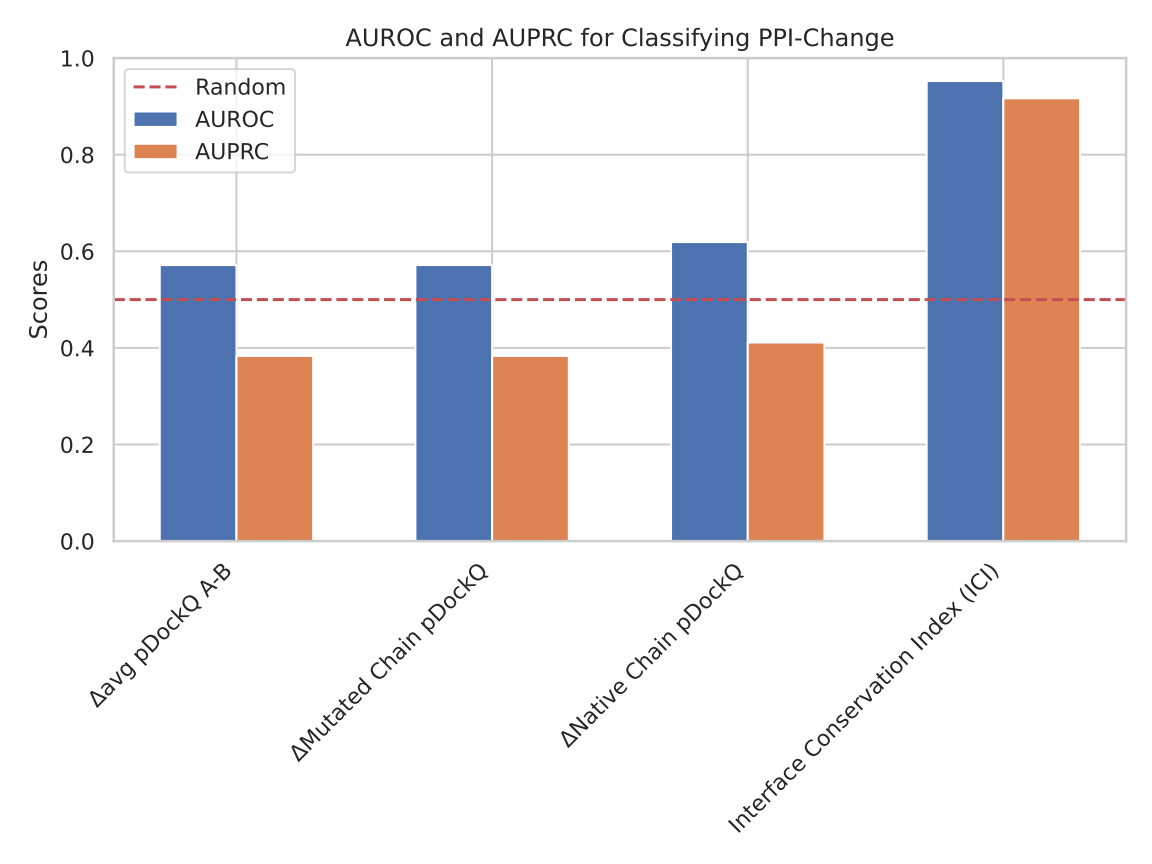


Figure1. AUROC and AUPRC for classifying PPI-change

Evaluation using the validated dataset: We computed ICI for all WT–MUT complex pairs and compared its performance with traditional ΔpDockQ scores. While ΔpDockQ alone showed limited discriminative power (AUROC = 0.52, AUPRC = 0.48), the ICI achieved significantly higher predictive performance (AUROC = 0.95, AUPRC = 0.91), demonstrating its effectiveness in capturing interface perturbations.

In practice, ICI values were combined with ΔpDockQ to classify mutations into categories of interaction perturbation, including:

hypo\_PPI – interaction weakening

non\_change – no significant change

This framework ensures that PIPA captures both binding affinity and structural consequences of mutations, thereby improving interpretability and biological relevance in downstream analyses.

Apppendix 4:

