**Supplementary Materials**

**AlloFusion: Allosteric Site Prediction Based on Language Models and Multi-Feature Fusion**

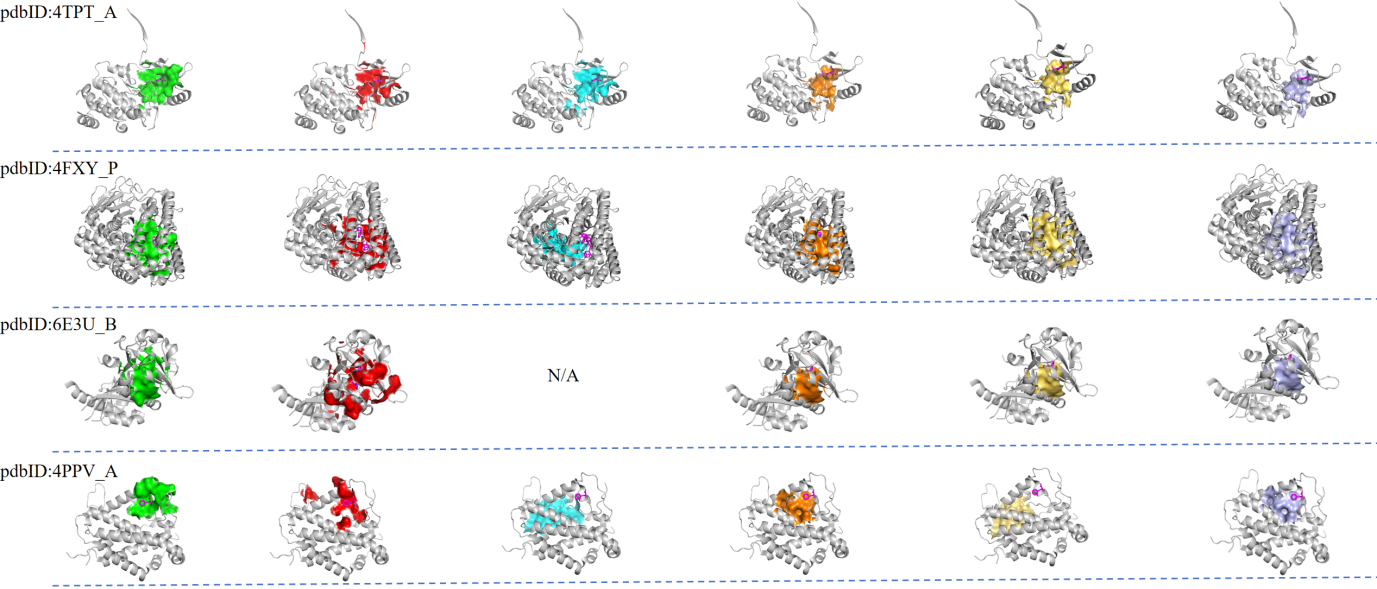
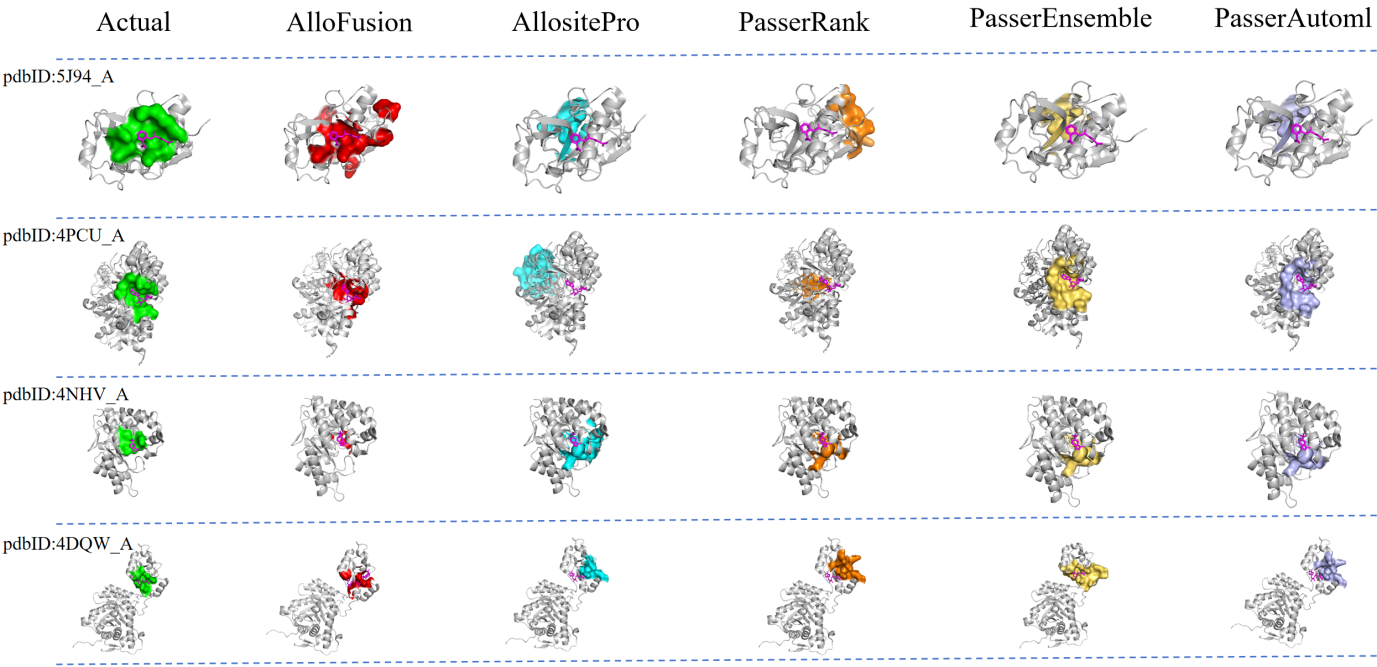
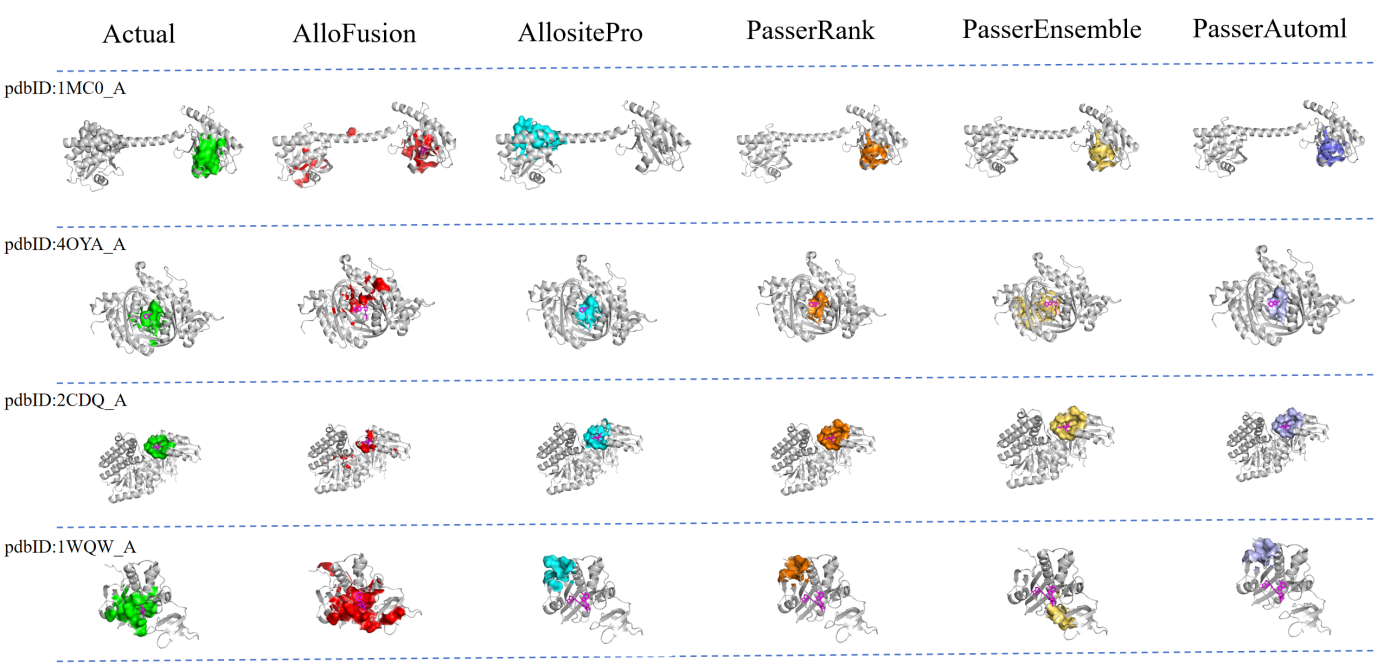
**Burak T. Kaynak†‡, Ivet Bahar†‡ and Pemra Doruker†**

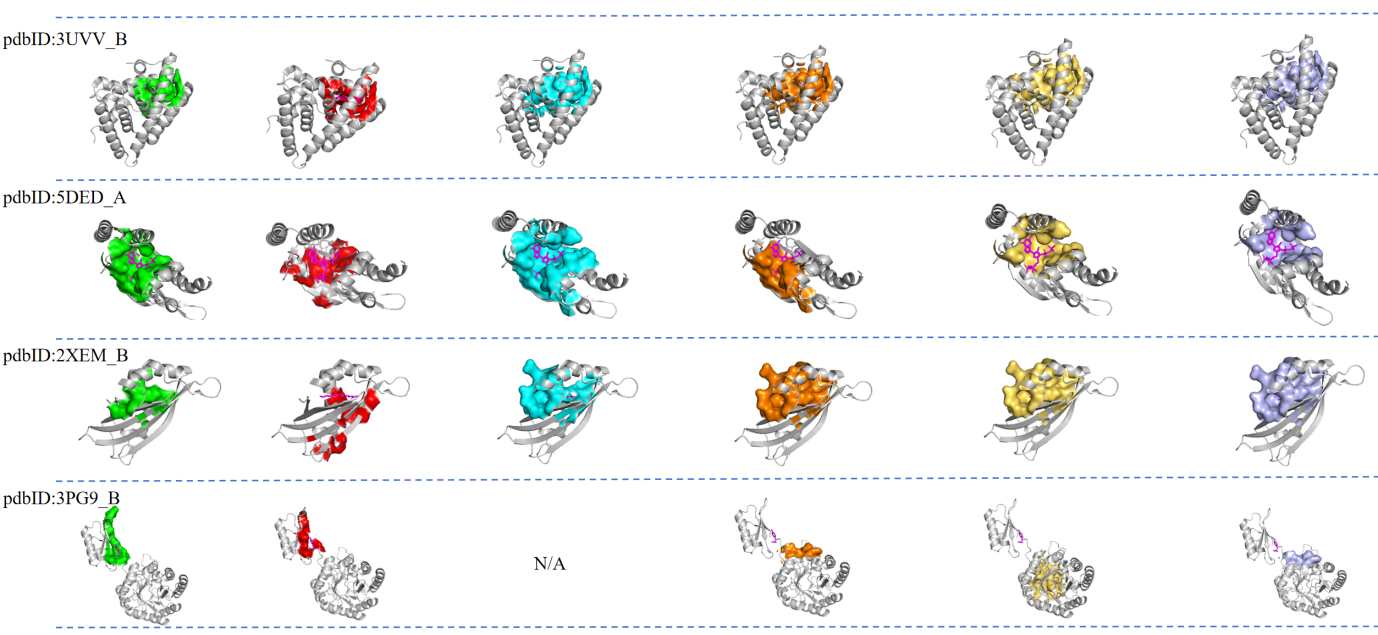
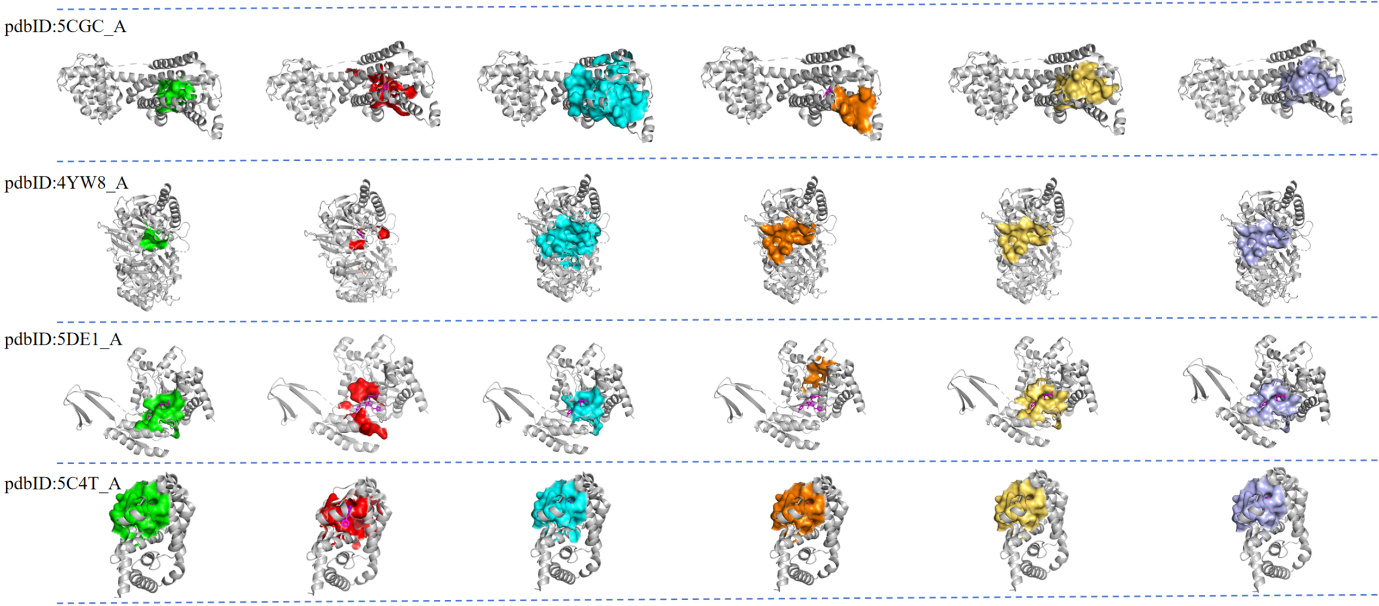
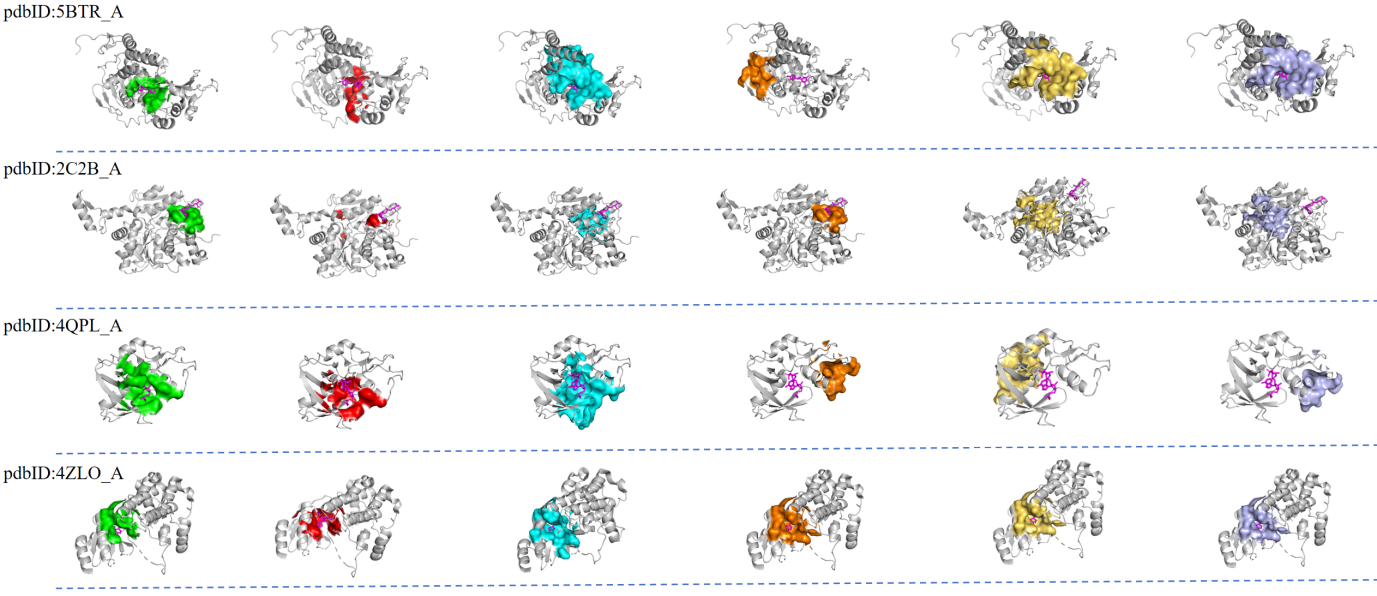
*†Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15232*

‡Corresponding authors: [bahar@pitt.edu](mailto:bahar@pitt.edu) and [burak.kaynak@pitt.edu](mailto:burak.kaynak@pitt.edu)

**Table S1**. The independent test set D24 of 24 allosteric proteins

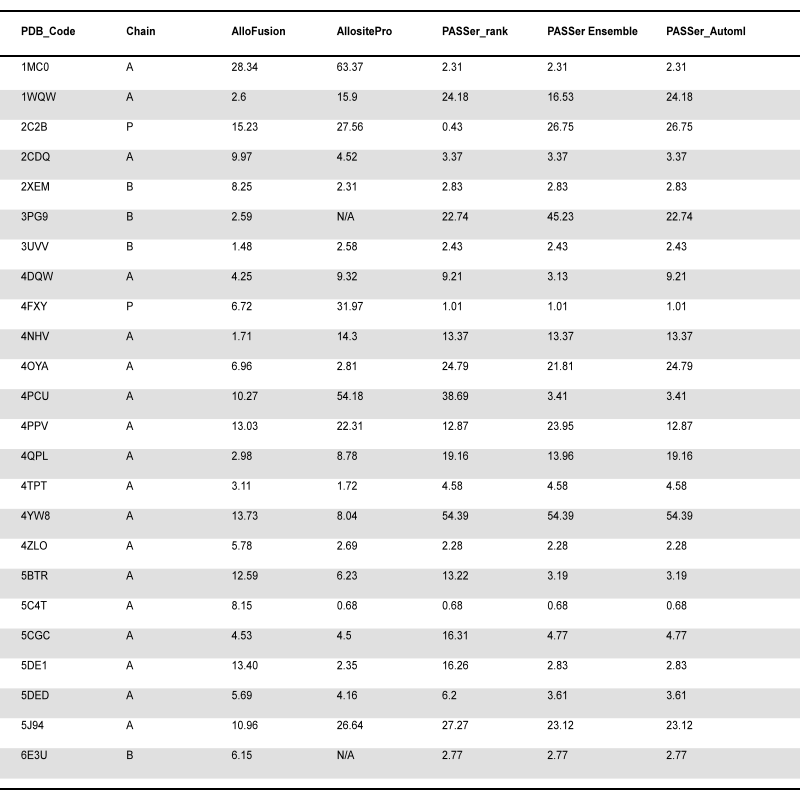
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Protein Name** | **PDB ID** | **Chain** | **Residue Num.** |
| 1 | 3',5'-cyclic nucleotide phosphodiesterase 2A | 1MC0 | A | 369 |
| 2 | Adenylate cyclase type 10 | 4OYA | A | 470 |
| 3 | Aspartokinase 1, chloroplastic | 2CDQ | A | 510 |
| 4 | BirA(Pyrococcus horikoshii) | 1WQW | A | 235 |
| 5 | Cathepsin K | 5J94 | A | 223 |
| 6 | Cystathionine beta-synthase | 4PCU | A | 549 |
| 7 | Dihydropteroate synthase | 4NHV | A | 297 |
| 8 | Inosine-5'-monophosp hate dehydrogenase | 4DQW | A | 509 |
| 9 | LIM domain kinase 2 | 4TPT | A | 303 |
| 10 | Neurolysin, mitochondrial | 4FXY | P | 693 |
| 11 | Heterodimeric HIF-2 | 6E3U | B | 642 |
| 12 | Chorismate mutase 1, chloroplastic | 4PPV | A | 280 |
| 13 | NAD-dependent protein deacetylase sirtuin-1 | 5BTR | A | 397 |
| 14 | Threonine synthase 1, chloroplastic | 2C2B | A | 486 |
| 15 | Ubiquitin-conjugating enzyme E2 D1 | 4QPL | A | 155 |
| 16 | Serine/threonine-prote in kinase PAK 1 | 4ZLO | A | 298 |
| 17 | Metabotropic glutamate receptor 5,Endolysin,Metabotr opic glutamate receptor 5 | 5CGC | A | 444 |
| 18 | Phosphoenolpyruvate carboxykinase, cytosolic [GTP] | 4YW8 | A | 624 |
| 19 | Isocitrate dehydrogenase [NADP] cytoplasmic | 5DE1 | A | 413 |
| 20 | Nuclear receptor ROR-gamma | 5C4T | A | 241 |
| 21 | Thyroid hormone receptor alpha | 3UVV | B | 244 |
| 22 | GTP pyrophosphokinase YjbM | 5DED | A | 218 |
| 23 | DynE7 | 2XEM | B | 150 |
| 24 | TmaDAH7PS | 3PG9 | B | 338 |





**Figure S1.** 3D representations of protein AFRs predicted by different models on the D24 dataset. Each row, from left to right, displays the experimentally determined AFRs, followed by predictions from AlloFusion, AllositePro, PASSer Rank, PASSer Ensemble, and PASSer Automl. Green indicates the actual allosteric sites composed of AFRs; red, cyan, orange, light yellow, and light blue represent the predicted AFRs by AlloFusion, AllositePro, PASSer Rank, PASSer Ensemble, and PASSer Automl, respectively. Gray represents the cartoon structure of the protein. "N/A" denotes cases where no allosteric site is predicted by the corresponding method.

**Table S2.** Distances between the predicted allosteric site centroids and the actual binding site on D24 using different methods.



**Table S3.** Key parameters set in LoRAConfig for fine-tuning the ProtT5 model.

|  |  |  |
| --- | --- | --- |
| **Parameters** | **Settings** | **Rationale for Selection** |
| lora\_rank | 4 | Reduces the number of trainable parameters to ensure a balance between performance and computational efficiency. |
| lora\_init\_scale | 0.01 | Small-scale initialization prevents gradient explosion and maintains training stability. |
| lora\_modules | .\*SelfAttention|.\*EncDecAttention | Focuses fine-tuning on attention layers to better capture long-range dependencies relevant to allosteric site recognition. |
| lora\_layers | q|k|v|o | Covers all components of the attention mechanism, enhancing the model’s ability to learn both local and global information. |
| trainable\_param\_names | .\*layer\_norm.\*|.\*lora\_[ab].\* | Restricts fine-tuning to LoRA and LayerNorm layers, reducing computational overhead. |
| lora\_scaling\_rank | 1 | Maintains stable parameter updates and accommodates small-scale LoRA adjustments. |