



CHALMERS
UNIVERSITY OF TECHNOLOGY

CAPIM: Catalytic activity and site prediction and analysis tool in multimer proteins

Downloaded from: <https://research.chalmers.se>, 2026-04-19 05:12 UTC

Citation for the original published paper (version of record):

Özsari, G., Garcia Soriano, D., Parate, S. et al (2025). CAPIM: Catalytic activity and site prediction and analysis tool in multimer proteins. *Protein Science*, 34(11). <http://dx.doi.org/10.1002/pro.70347>

N.B. When citing this work, cite the original published paper.

CAPIM: Catalytic activity and site prediction and analysis tool in multimer proteins

Gökhan Özsari^{1,2} | Daniela A. García-Soriano¹ | Shraddha Parate³ |
Amar el Issaoui³ | Pernilla Wittung-Stafshede^{3,4} 

¹E-Commons, Chalmers University of Technology, Gothenburg, Sweden

²Computer Engineering Department, Middle East Technical University, Ankara, Turkey

³Life Sciences Department, Chalmers University of Technology, Gothenburg, Sweden

⁴Chemistry Department, Rice University, Houston, Texas, USA

Correspondence

Pernilla Wittung-Stafshede, Chemistry Department, Rice University, Houston, TX, USA.

Email: pernilla.wittung@rice.edu

Funding information

Cancerfonden; Vetenskapsrådet; Knut och Alice Wallenbergs Stiftelse

Review Editor: Nir Ben-Tal

Abstract

Enzymes play a fundamental role in living organisms by catalyzing vital chemical reactions. While much is known about enzyme function, a substantial portion of the proteome remains uncharacterized. Computational tools have become indispensable in this field, yet most focus exclusively on either enzymatic activity prediction or active site detection, creating a gap between residue-level annotation and functional characterization. To bridge this gap, we present Catalytic Activity and Site Prediction and Analysis Tool In Multimer Proteins (CAPIM)—an integrative computational pipeline that combines binding pocket identification and catalytic site annotation with enzymatic activities, along with functional validation via enzyme–substrate docking. CAPIM unifies the capabilities of three established tools: P2Rank, GASS, and AutoDock Vina. P2Rank uses a machine learning-based approach to predict binding pockets, while genetic active site search (GASS) identifies catalytically active residues and annotates them with Enzyme Commission numbers. These outputs are merged to generate residue-level activity profiles within predicted pockets. Functional validation is then performed using AutoDock Vina, enabling substrate docking simulations for user-defined ligands. CAPIM supports any number of peptide chains in the protein complex—which may be crucial for enzymatic functions dependent on quaternary and/or polymeric (e.g., amyloid) structures. The utility of CAPIM is demonstrated through case studies involving both well-characterized enzymes and unannotated multi-chain targets. By delivering residue-level predictions and docking analyses in a unified framework, CAPIM offers a powerful resource with broad applications in drug discovery and protein engineering. CAPIM is available both as a standalone application at <https://git.chalmers.se/ozsari/capim-app> and as a hosted web service at <https://capim-app.serve.scilifelab.se>.

KEYWORDS

activity prediction, catalytic site, enzyme activity, protein structure, software

1 | INTRODUCTION

Enzymes are essential biocatalysts, and their mechanistic understanding is fundamental for applications in

Gökhan Özsari and Daniela A. García-Soriano shared first authorship.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *Protein Science* published by Wiley Periodicals LLC on behalf of The Protein Society.

drug discovery, protein engineering, and sustainable chemistry. Key to this understanding is the identification of catalytic activity, that is, what chemical reactions are facilitated, and catalytic sites, that is, the specific regions in the enzyme where substrate binding and chemical transformation occur. Computational tools have become valuable assets for addressing this task, offering fast and cost-effective aid to experimental methods.

Many computational tools address the catalytic activity of proteins by assigning Enzyme Commission (EC) numbers, often relying on sequence-based machine learning methods. EC numbers are a standardized system for classifying enzymes based on the reactions they catalyze. Each EC number consists of four numbers separated by periods (e.g., EC 2.7.1.1), where the first indicates the general class of the enzyme (such as oxidoreductases or transferases), and the following numbers provide increasing levels of specificity about the type of reaction and substrates involved. This system helps organize and identify enzymes by their function in a consistent way. Tools such as *ECPred* (Dalkiran et al., 2018), *DeepEC* (Ryu et al., 2019), and *CLEAN* (Yu et al., 2023) utilize evolutionary conservation, motifs, and functional annotations to infer catalytic roles (EC numbers) from sequence data. While effective for high-throughput annotation, these approaches lack structural context and are limited in resolving residue-level mechanistic detail.

To complement this, structure-based methods, like *GASS* (Izidoro et al., 2015; Moraes et al., 2017), search for given active site 3D templates providing EC number predictions. *P2Rank* (Jakubec et al., 2022; Krivák & Hoksza, 2018), *Fpocket* (Le Guilloux et al., 2009), *CASTp* (Tian et al., 2018), and *SITEHOUND* (Hernandez et al., 2009) identify geometric features such as pockets and cavities, helping localize potential binding sites without assigning function or identifying catalytically relevant residues. Recent advances in machine learning have led to hybrid tools like *PUResNet* (Kandel et al., 2021) and *BindWeb* (Xia et al., 2022), which integrate both structural and sequence features to enhance binding site predictions. For example, *BindWeb* combines a graph neural network (GNN) (GraphBind) and a convolutional neural networks (CNN)-long short-term memory (LSTM) model (DELIA) (Xia et al., 2020) to predict ligand-binding residues and pockets. While powerful, these tools focus on either activity classification or binding site prediction and often fall short of connecting both.

The lack of integration between catalytic residue/site identification and functional annotation remains a critical limitation in computational approaches to enzyme research. Although computational tools for the separate tasks have advanced significantly, persistent challenges remain:

1. *Fragmented capabilities*: Tools for enzymatic activity prediction typically focus on assigning high-level

functions, such as EC numbers, without providing residue-level annotations. Conversely, tools that identify active or binding sites often fail to specify the enzymatic activity occurring at these sites.

2. *Reliance on sequence only*: Many enzymatic prediction tools depend heavily on sequence-based information, leveraging evolutionary conservation, sequence motifs, and alignment-based features. While effective for identifying conserved regions, these methods neglect structural contexts, which may be essential for mechanisms and substrate specificity.
3. *Single polypeptide chain limitation*: Structure-based tools frequently restrict their input to single protein chains, or have a maximum chain limit, which prevents accurate modeling of multimers. As a result, available tools fail to capture the complexity of multi-domain enzymes and polymeric protein assemblies.

Computational tools capable of bridging the gap between catalytic residue identification and functional annotation will be instrumental in advancing our understanding of existing enzymes and how to design new ones. To address this gap, we present CAPIM, an integrative computational tool designed to unify enzymatic activity prediction with active site identification. CAPIM combines structure-based binding site identification and enzymatic activity prediction by integrating P2Rank (Jakubec et al., 2022; Krivák & Hoksza, 2018) and GASS (Izidoro et al., 2015; Moraes et al., 2017) together. By connecting the predictive algorithms with a final docking step, CAPIM enables residue-level identification of active sites directly coupled to functional annotation, followed by substrate interaction analysis, in a user-friendly pipeline. There are many docking tools available, such as *AutoDock Vina* (Eberhardt et al., 2021; Trott & Olson, 2010), *DiffDock* (Corso et al., 2023), and *GAA-Bind* (Tan et al., 2024). In CAPIM, we include *AutoDock Vina* for docking, but users may take CAPIM results to any favorite module. Crucially, CAPIM has no limitation in number of peptide chains included in the analysis, making it suitable for larger (even polymeric) protein structures.

Importantly, our aim is not to outperform existing specialized EC predictors (e.g., *DeepEC*, *CLEAN*), but to provide residue-level functional annotation and binding site validation in an integrated and accessible framework that complements such methods. CAPIM is thus designed for hypothesis generation by experimentalists rather than for benchmarking EC prediction performance.

Below, we first provide a section (Section 2) that details the design and workflow of CAPIM. It is followed by Section 3, where we demonstrate the use of CAPIM on several selected enzymes, then discussion (Section 4), and finally conclusions (Section 5).

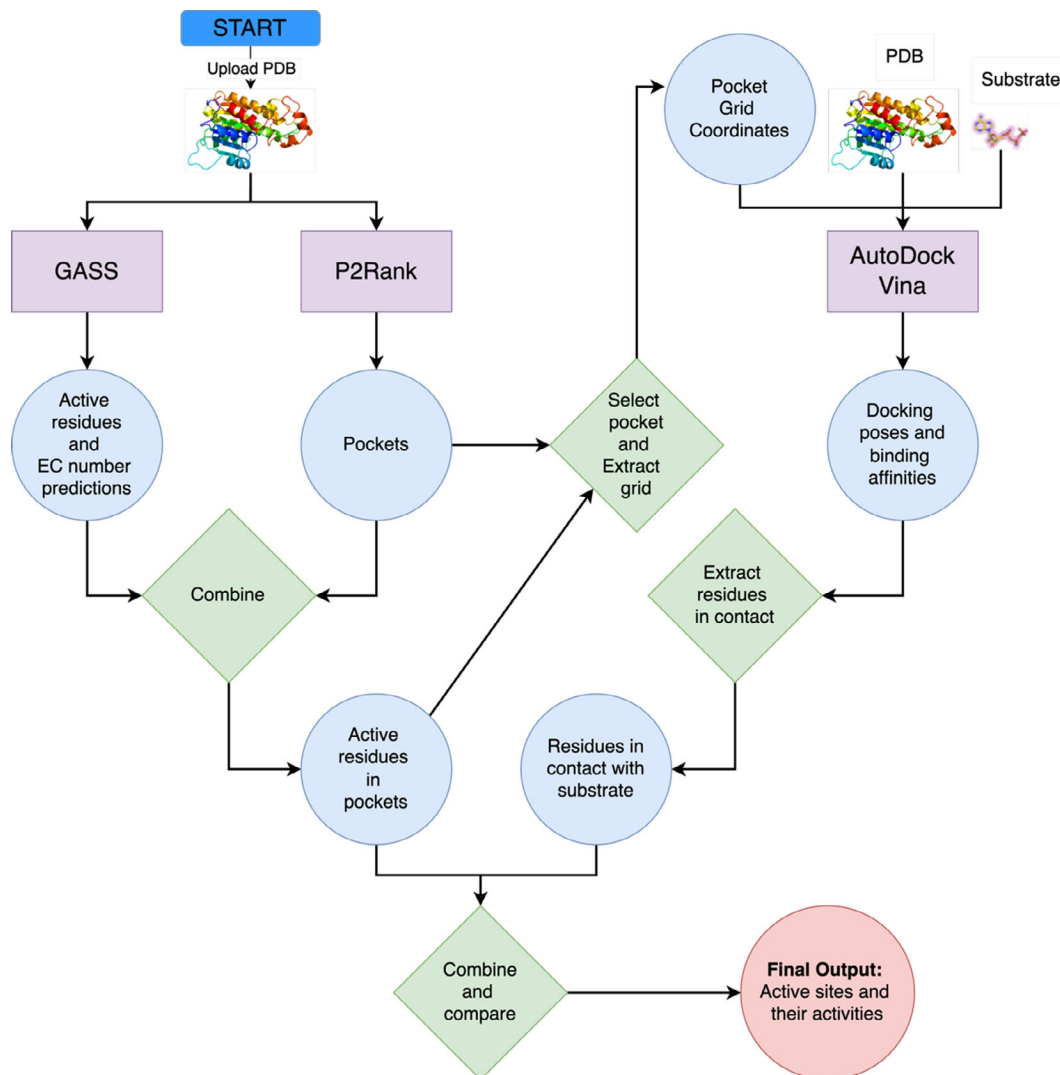


FIGURE 1 Workflow of the CAPIM tool for catalytic-site identification and analysis. The diagram illustrates the process employed by CAPIM, starting with combining pocket prediction using P2Rank and active residue identification with GASS. These results of the two are integrated (filtered) to identify active residues in binding pockets. The workflow proceeds with the possibility to validate substrate-binding sites through AutoDock Vina. The residues found to be in contact with specific substrates can then be compared to the predicted active-site residues in the pocket. The final output from CAPIM includes predicted active-site residues, their enzymatic activities, and (if desired) substrate docking poses.

2 | METHODOLOGY

2.1 | Overview of components

CAPIM combines predictive algorithms and docking simulations to address the goals of identifying catalytic residues, annotating their enzymatic activities, and validating their functional roles through enzyme–substrate interactions. The tool employs a structure-based computational approach to achieve two primary objectives: (1) identifying catalytic sites along with their enzymatic activities and (2) performing a comprehensive analysis of the interaction sites through enzyme–substrate docking. By combining predictive algorithms and docking simulations, the workflow integrates existing tools and custom analyses into a unified pipeline to explore

enzymatic activity in detail. To accomplish this, the CAPIM tool integrates three established tools: P2Rank, GASS, and AutoDock Vina, each contributing unique functionalities to the workflow as depicted in Figure 1. Below, we introduce these tools and their roles within the pipeline.

P2Rank is an innovative machine learning-based tool designed for high-accuracy prediction of ligand-binding pockets on enzyme structures. Unlike template-based methods, P2Rank operates independently of structural templates, making it highly suited for automated pipelines and large-scale analyses. The tool uses a Random Forest classifier trained on physico-chemical, geometric, and statistical features to evaluate ligandability at specific points on the enzyme's solvent-accessible surface. The workflow involves generating

solvent-accessible points, calculating feature descriptors, and clustering high-scoring points to predict binding sites efficiently. This approach enhances the precision of binding site identification by leveraging local chemical neighborhood information. Within the CAPIM pipeline, P2Rank provides the foundational binding pocket predictions used for further catalytic residue identification and is used as the reference grid in docking analysis.

The *Genetic Active Site Search (GASS)* method is a heuristic tool leveraging genetic algorithms to predict enzyme active sites, including catalytic and substrate-binding sites, based on structural templates. GASS outperforms several existing methods by allowing non-exact amino acid matches and identifying residues across different protein chains without size restrictions on active sites. It processes 3D structural data from protein databases like Protein Data Bank (PDB), employs fitness functions based on distance metrics, and handles conservative mutations through a substitution matrix. GASS has been validated against the Catalytic Site Atlas (CSA) and demonstrated high accuracy, correctly identifying over 90% of catalytic sites in multiple datasets. Additionally, it ranked fourth among 18 methods in the CASP10 substrate-binding site competition, highlighting its effectiveness and adaptability in protein function prediction tasks. GASS contributes to CAPIM by annotating catalytically active residues and assigning EC numbers to provide functional insights.

AutoDock Vina employs an energy-based docking approach to predict the binding pose and affinity of ligands to their respective receptors. Its scoring function estimates binding energy by accounting for key molecular interactions, such as hydrogen bonding, hydrophobic contacts, and van der Waals forces. The software supports flexible ligand conformations and allows partial flexibility of the protein receptor, providing a balance between computational efficiency and biological realism. Moreover, its multi-threading capability enables researchers to leverage modern computing power for high-throughput virtual screening. Within CAPIM, AutoDock Vina can be used to validate predicted catalytic sites by assessing their ability to interact with substrates, providing quantitative measures of binding affinity and spatial compatibility.

We selected P2Rank for its template-free, machine learning approach that reliably identifies ligand-binding pockets, and GASS for its ability to annotate catalytic residues across chains with EC numbers. Their complementary strengths—spatial precision from P2Rank and functional annotation from GASS—make them particularly suitable for integration within CAPIM. For the docking step, we chose AutoDock Vina because it is lightweight and central processing unit-efficient, allows definition of a region of interest via grid boxes around

predicted binding pockets, and is widely validated and accessible. These features make it a practical choice for inclusion in CAPIM, where ease of use and reproducibility are prioritized.

2.2 | Identification of catalytic sites and enzymatic activity

The identification of catalytic sites and their associated enzymatic activities is achieved through the integration of P2Rank and GASS (Figure 1). The user input is a PDB file of the protein structure of interest. As part of its pre-processing, CAPIM automatically removes ligands and solvent molecules from the structure. The workflow begins with P2Rank, which identifies potential binding pockets in the enzyme structure. These pockets represent spatial regions where substrates are likely to bind, serving as the foundation for downstream catalytic site predictions. Next, GASS is used to predict catalytically active residues and assign EC numbers. The user can select EC levels of interest at any level of detail. By integrating these two tools, CAPIM ensures a dual-layered analysis: spatial relevance from P2Rank's pocket predictions and functional importance from GASS's residue-level EC predictions. The results from P2Rank and GASS are combined to identify catalytic residues that reside specifically within the predicted binding pockets. This refinement step filters out residues that lack spatial or functional significance, thus improving the accuracy of the prediction and minimizing false positives. The resulting predictions are ranked and can be sorted based on P2Rank's pocket probability scores or GASS's fitness scores, allowing users to prioritize residues and pockets with the highest predicted relevance. By narrowing the analysis to catalytically active residues within binding pockets, the pipeline focuses on sites most likely to participate in substrate interactions and enzymatic reactions.

2.3 | Additional substrate docking module

Following the identification of catalytic residues in relevant pockets, the next module in CAPIM allows for detailed analysis of enzyme–substrate interactions. Input files here can be standard PDB files (that CAPIM will preprocess) or preprocessed Protein Data Bank, Partial Charge (Q), and Atom Type (T) files prepared by the user. The user selects what substrate to use based on the identified type of activity. The identified pockets, as defined by P2Rank, are utilized as reference grids to narrow the docking region to catalytically relevant residues directly or indirectly involved in catalysis.

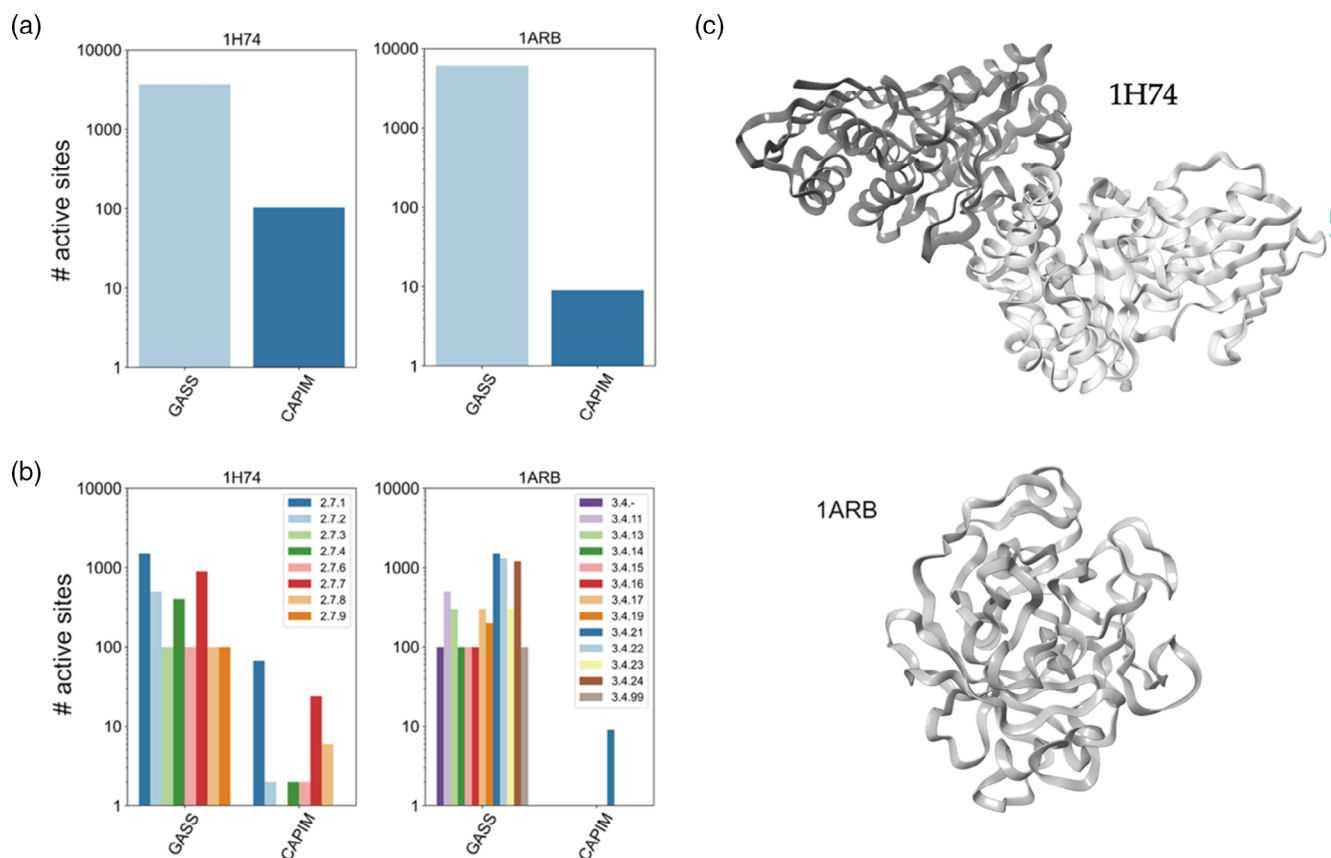


FIGURE 2 Comparison of predicted active sites from GASS alone versus CAPIM for **1ARB** and **1H74**. (a) Bar plots comparing the number of predicted active sites assigned to each protein at the second Enzyme Commission (EC) level, as identified by GASS alone versus CAPIM (EC 2.7 for **1H74** and EC 3.4 for **1ARB**). (b) Number of predicted active sites at the third EC level, contrasting predictions from GASS alone with those from CAPIM. (c) Three-dimensional structure of **1H74** (dimer, top) and **1ARB** (monomer, bottom).

Docking simulations are performed using AutoDock Vina and involve the evaluation of substrate binding poses and the calculation of binding affinities.

The final output of the CAPIM pipeline includes (Figure 1): a list of catalytic residues identified in the protein structure, their associated catalytic activities (EC numbers), as well as substrate docking poses and binding affinities. All files generated by CAPIM are available as downloadable files. By integrating P2Rank, GASS, and docking simulations into a single pipeline, CAPIM offers a robust and precise platform for studying enzyme catalytic activity in one unified user-friendly tool.

3 | RESULTS—CASE STUDIES

3.1 | Monomer (**1ARB**) and dimer (**1H74**) with known activity

To evaluate whether CAPIM can improve the prediction of enzymatic activity in proteins in general, we first selected two proteins previously reported in the GASS dataset (Morales et al., 2017): **1H74** and **1ARB**. **1ARB**

is a protease I (EC 3.4.21.50) from *Achromobacter lyticus* which hydrolyzes lysyl peptide bonds (Tsunasawa et al., 1989). It is a monomer with a chain length of 268. **1H74** is a homoserine kinase (EC: 2.7.1.39) (Krishna et al., 2001) that acts in the aspartate pathway of amino acid biosynthesis. It is a dimer of two chains of 296 amino acids. When the two proteins were analyzed previously by GASS only (Izidoro et al., 2015; Moraes et al., 2017), the experimentally determined active site was ranked relatively low in **1H74**, and in **1ARB**, the experimentally determined active site was mixed with many others. In CAPIM, we limited the GASS search on each protein to the second EC number, that is, for **1H74**, we used EC: 2.7 and for **1ARB**, we used EC: 3.4 as input, to reduce and focus the number of predictions. It is important to mention that since CAPIM can save the results from GASS alone, we did not have to run GASS on the webserver to make comparisons.

The first striking observation when applying CAPIM (the first combined P2Rank and GASS module) to these proteins was a notable reduction in the number of predicted sites as compared to GASS-only predictions (Figure 2a). Specifically, for **1H74**, GASS alone provided 3700 sites within EC 2.7, while CAPIM reported only

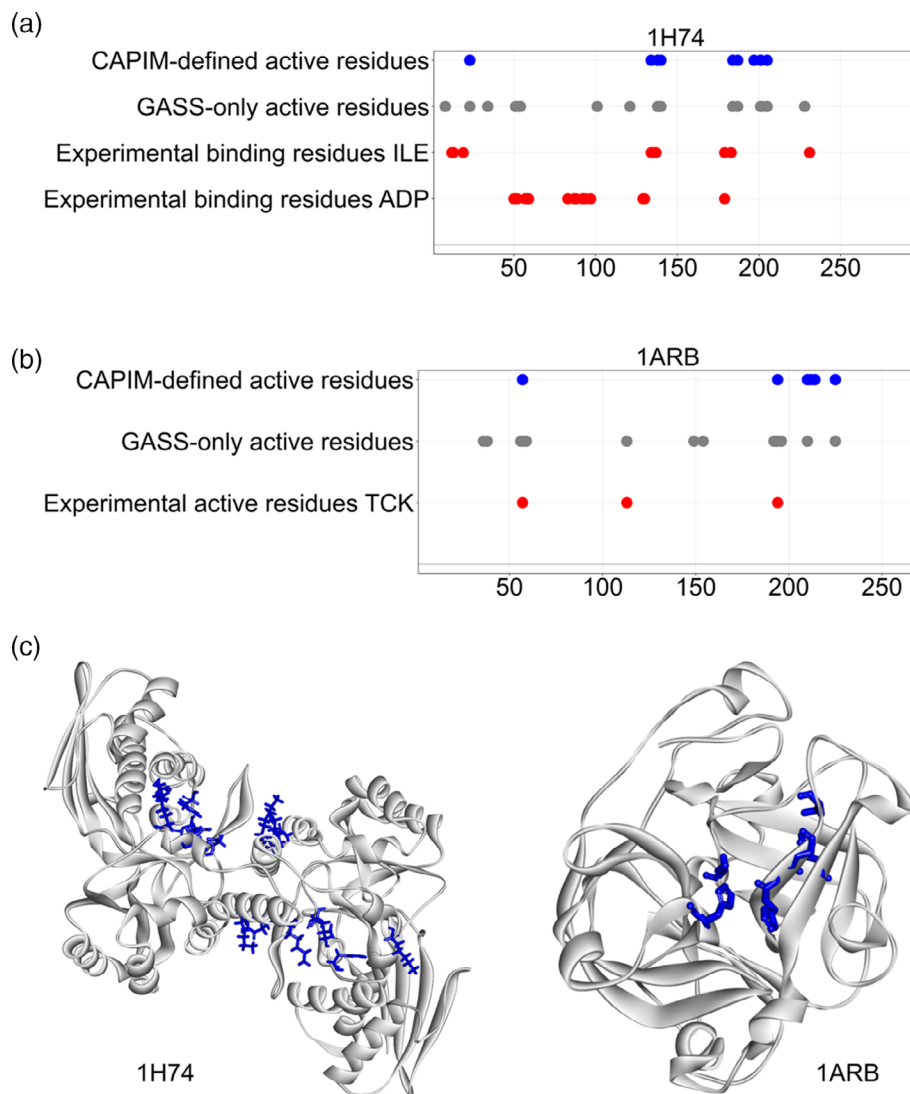


FIGURE 3 Comparison of predicted versus experimentally validated active-site residues. Dot plots of residues part of predicted and experimental active sites (a, **1H74**; b, **1ARB**). Red dots mark experimentally verified active-site residues: For **1H74**, residues represent binding site of adenosine diphosphate (ADP) and Ile, respectively (per monomer); for **1ARB**, residues represent reported single active site (taken from UniProt). Gray dots show residues predicted by GASS alone (including the top 10 predictions after ranking based on fitness score from GASS). Blue dots represent residues predicted by CAPIM as part of the active sites: For **1H74**, the residues in the top 10-ranked CAPIM-defined active sites are merged; for **1ARB**, the residues in all the CAPIM-defined active sites are merged. (c) Protein structures with the CAPIM identified active-site residues highlighted in blue (two sites for **1H74** as dimer).

103 sites. For **1ARB**, GASS alone provided 6100 sites and CAPIM only 9 at the second EC level. When looking at the results at the third EC level (Figure 2b), for **1H74** we find 40% of predicted active sites by GASS only (1500 of 3700) to have the correct EC number: EC 2.7.1. This percentage rose to 65% (67 of 103) when using CAPIM. In the case of **1ARB**, 24% (1500 of 6100) corresponds to the right EC number at the third level (3.4.21) for GASS only, while 100% of the predictions from CAPIM give EC 3.4.21. Thus, CAPIM reduces the number of hits in a way that retains the correct answer.

Next, we visually compared the results obtained from CAPIM with the experimentally determined active-site residues and GASS alone. In Figure 3, we show—along the protein sequence—the residues belonging to the CAPIM-selected top 10 active sites (all nine for **1ARB**), along with the residues belonging to the top 10 predictions when using GASS only.

Notably, many of the CAPIM predicted active sites are minor variations of the same set of residues in

different combinations. Even if the residues appear spread out in sequence, they all cluster near each other in the three-dimensional structure (Figure 3c). Thus, the CAPIM predicted active-site residues together define one region on each protein monomer. Several of the experimentally verified active-site residues (data taken from UniProt (The UniProt Consortium, 2018)) are identical to the CAPIM-identified residues (Figure 3a,b). When comparing experimental data to predictions, one must check how active-site residues were identified (mutation, interaction with ligand in crystal, what ligand was used etc.) before giving too much weight to the discrepancy of details. Experimental active sites may be limited to what was tested and do not exclude additional residues of importance. For example, for **1H74**, the reported binding site depends on the ligand assessed (Figure 3a). The important point to make from this comparison is that CAPIM indeed finds one active site region on each protein that agrees with experimentally determined active-site residues.

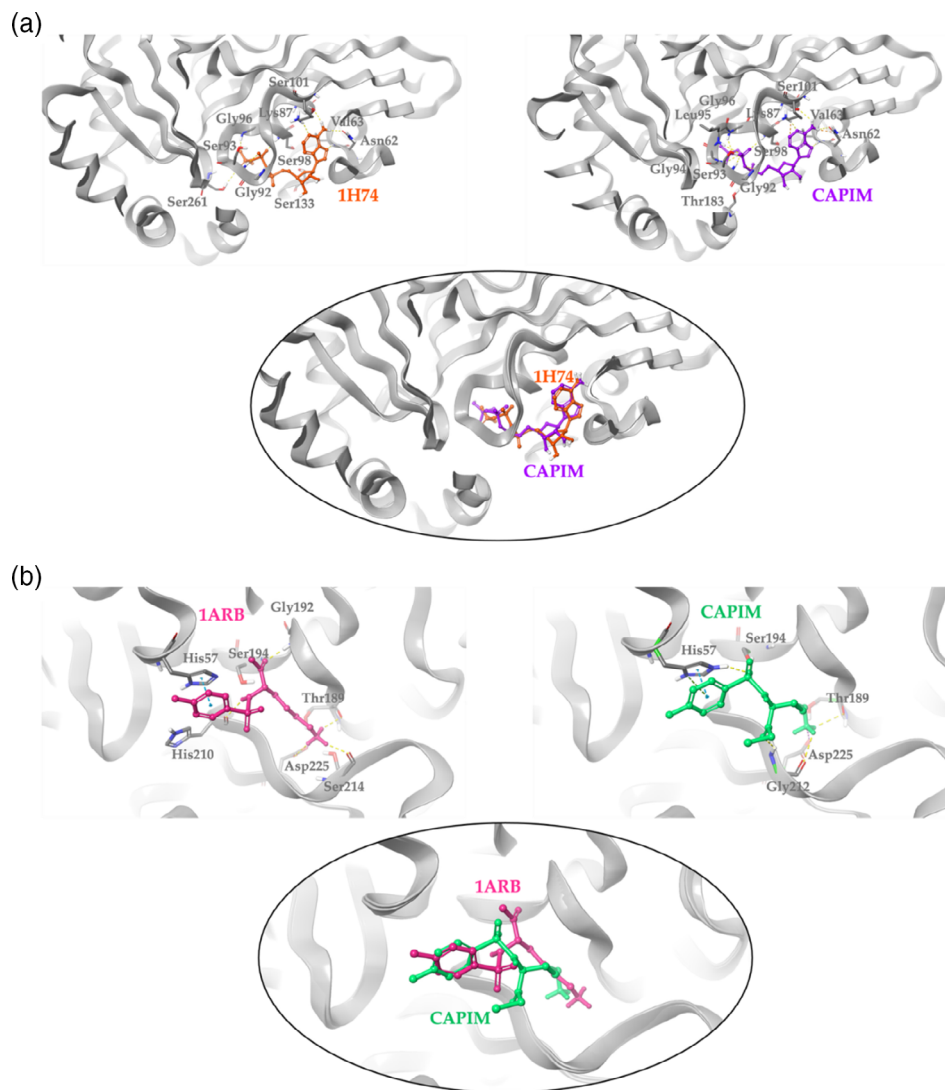


FIGURE 4 Structural comparison of experimentally determined and CAPIM-docked complexes. Visual representations of substrate sites in (a) **1H74** monomer using adenosine diphosphate ligand and (b) **1ARB** using Tosyl-Lysyl Chloromethylketone ligand. Each panel includes *Left*: The 3D protein structure from PDB with its co-crystallized ligand. *Right*: The same PDB structure after docking using the CAPIM approach, displaying the predicted ligand pose. *Bottom*: A superposition of experimental (PDB) and CAPIM-docked ligand-bound structures, illustrating their spatial agreement.

The last step in the CAPIM pipeline is the option to dock substrates into predicted pockets. This module can be used to test unknown substrates or, as here, assess the interactions of known substrates. For the model proteins analyzed above, the type of enzymatic reactions they perform is established, and we thus docked appropriate ligands (for each, binding mode known from crystal structure) to the pockets containing the top CAPIM-identified active sites. For **1ARB**, we used the inhibitor Tosyl-Lysyl Chloromethylketone (TCK; a modified lysine residue) and, for **1H74**, the substrate adenosine diphosphate (ADP). The affinity values for the docking of each substrate to each predicted pocket in CAPIM revealed high affinity: for **1H74** and ADP, -9.3 kcal/mol; for **1ARB** and TCK, -5.2 kcal/mol. We then compared the docking results

with the substrate binding poses observed in the crystal structures (Figure 4). The significant similarity between docked and experimental substrate positions for both systems clearly demonstrates that CAPIM (1) finds the right binding site and (2) adds the substrate in an orientation that agrees with experimental data.

3.2 | Exploring EC numbers from the first level (1BP4)

In the above case studies, we limited the initial CAPIM search to a known second EC level. However, in many scenarios, little may be known about a protein's function, making it necessary to explore EC numbers more broadly. To explore this, we chose papain (**1BP4**), a

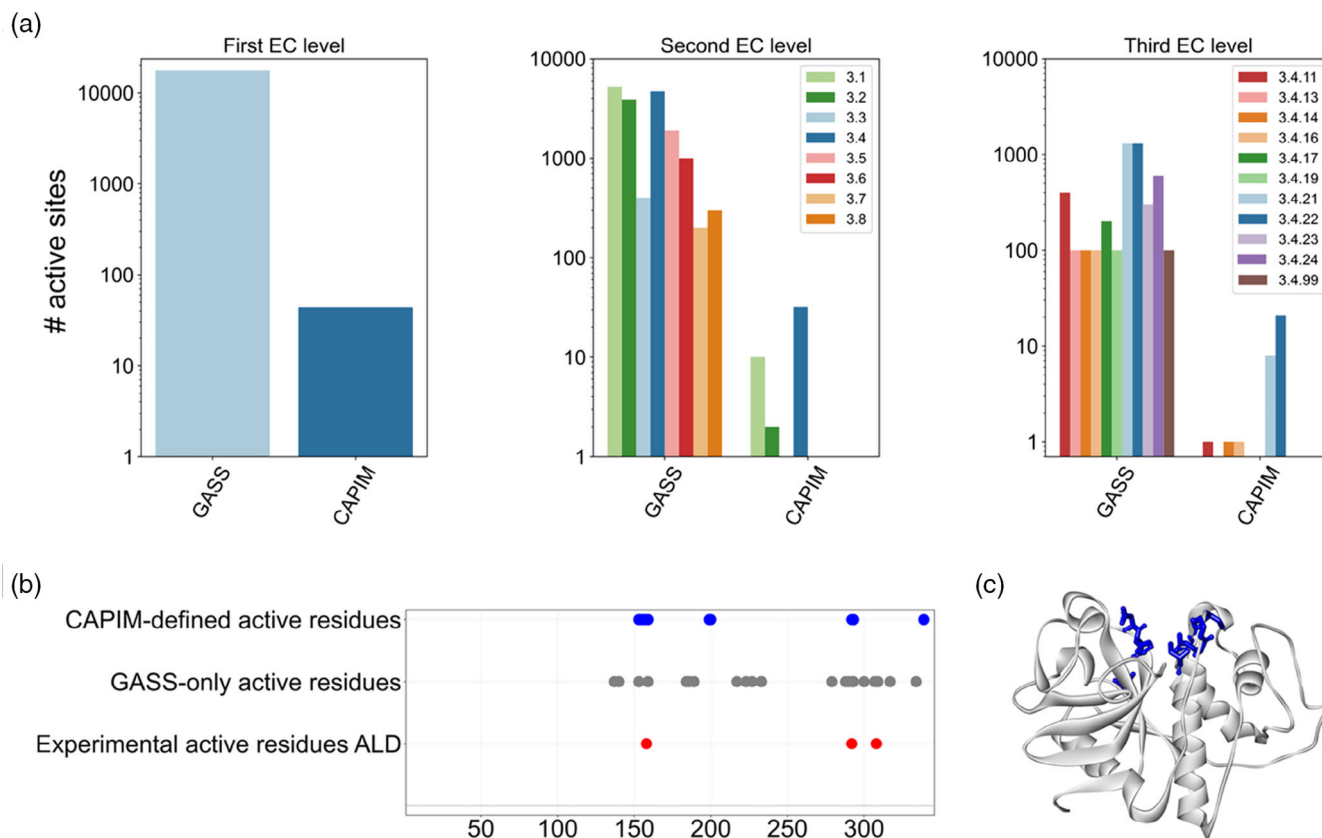


FIGURE 5 Comparison of predicted active sites from CAPIM, GASS alone, and experiment. (a) Bar plots showing a comparison of number of active sites for a specific Enzyme Commission (EC) number at the first (EC 3), second (EC 3.x), and third (EC 3.4.x) level predicted from GASS alone versus CAPIM for **1BP4** (note, 3.4.22 is correct EC number for **1BP4**). (b) Dot plots of residues part of predicted and experimentally verified active sites in **1BP4**. Color code is the same as Figure 3. Residues in the top 10 CAPIM-defined active sites for 3.4.22 and the top 10 GASS-only active sites at the first EC level are plotted (ranked by the fitness value given by GASS). (c) Protein structure with CAPIM-identified active-site residues (blue) highlighted.

212-residue cysteine protease of the papain-like C1 family that catalyzes broad-specificity peptide hydrolysis (EC 3.4.22) (LaLonde et al., 1998). Here we started the CAPIM analysis using only the first EC level as input, that is, EC 3 (hydrolases) and investigated the output at the second and third EC levels as compared to GASS only.

From the results (Figure 5), it is evident that CAPIM (again) delivers fewer hits than GASS only (44 vs. 17,600 for EC 3), and most of those are defined as EC 3.4: 32 (73% of total) and 4700 (27% of total) for CAPIM and GASS, respectively. When we investigated the third EC level within EC 3.4, we find that CAPIM again performs better than GASS: CAPIM reports 21 of 32 hits with the correct EC 3.4.22 (66%), whereas GASS reports 1300 of 4700 hits (28%) as EC 3.4.22 (Figure 5). Thus, even when CAPIM starts from a single-digit EC class, the cascade of P2Rank spatial filtering followed by GASS ranking greatly concentrates true positives, mirroring the trend seen for the earlier case studies.

Many times, the first EC number is not known, as one may want to explore enzyme activity widely

for an unknown protein. CAPIM can analyze proteins without selecting an EC number. Then the program analyzes all EC numbers at once, and one can analyze the results on different levels. In Figure 6, we show the output results at the first, second, and third EC level when we explored **1BP4** again using all EC numbers in CAPIM. Also, in this type of search is the true positive (experimentally verified activity 3.4.22) singled out by CAPIM. The data in Figure 5 is a subset of the data in Figure 6. Thus, regardless of the starting point, the CAPIM output result is the same.

3.3 | A polymer with unknown functions: α -synuclein amyloid (6A6B)

Finally, to test CAPIM on a multimeric structure with unknown catalytic annotation, we turned to α -synuclein amyloids. Recent in vitro work has shown that these amyloids can catalyze dephosphorylation and hydrolysis of ester bonds in model substrates

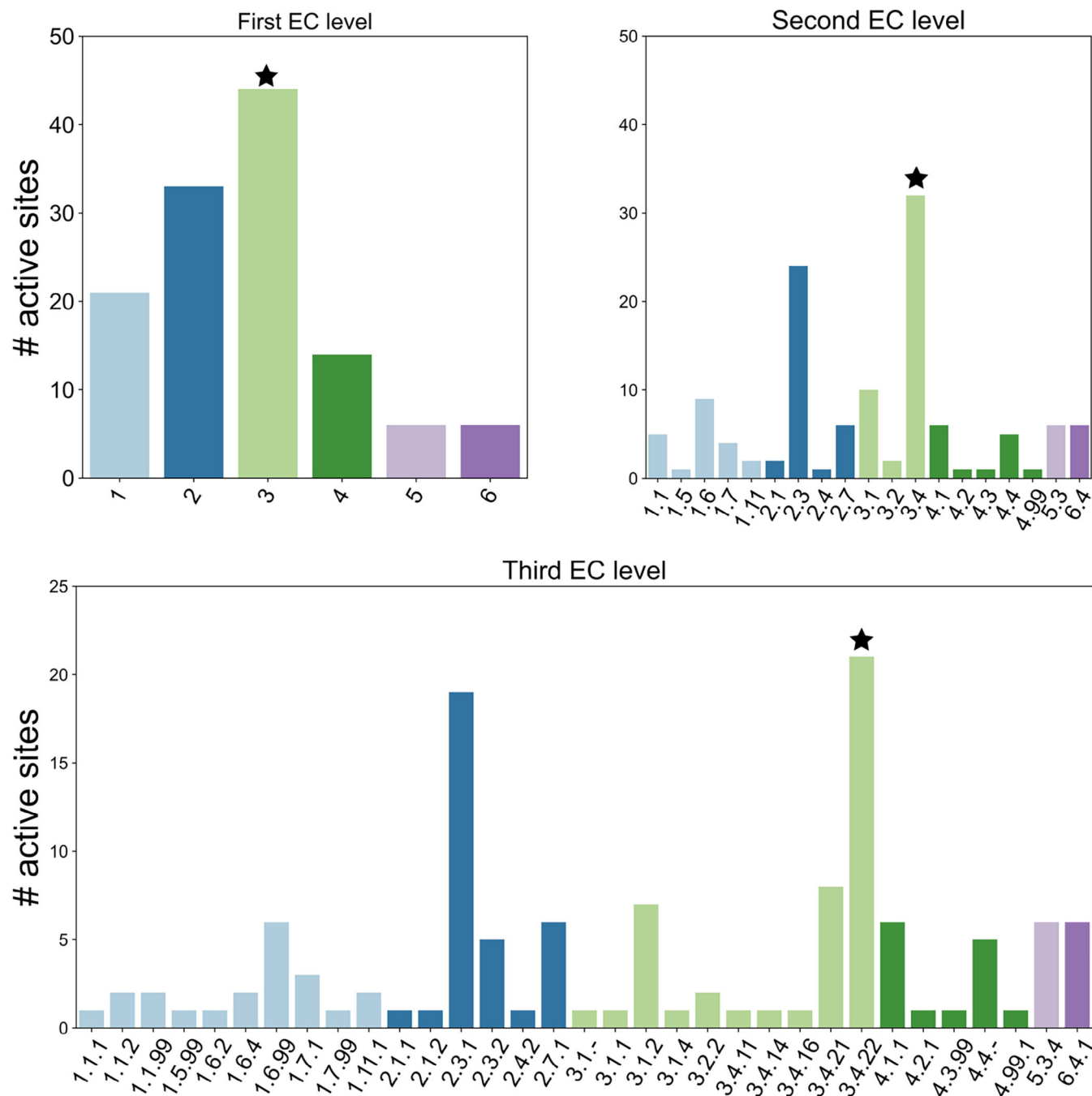


FIGURE 6 Prediction of enzymatic activity using all Enzyme Commission (EC) numbers at three different levels. Bar plots illustrate the predicted enzymatic activity across all EC numbers at three hierarchical levels. Bars represent the number of predicted active sites assigned to each EC subclass at that level. The EC subclass with the highest number of predicted active sites—which also corresponds to the experimentally validated enzyme class—is marked with a star symbol in each panel.

(Horvath & Wittung-Stafshede, 2023). However, the full range of catalytic activities these amyloids may harbor, and the identity of their catalytic sites remain unknown. Since the experimental work was conducted under physiological conditions where wild-type α -synuclein typically adopts a Type-1A fold (Frey et al., 2024), we selected a PDB structure consistent with that fold: 6A6B. This structure is a multimer

composed of 12 chains (6 per protofilament). As noted, CAPIM has no restriction on the number of protein chains within a structure to be analyzed. To explore the enzymatic potential of 6A6B, we systematically applied CAPIM across all EC levels.

At EC level 1, CAPIM predictions showed a dominance of hydrolases (EC 3), followed by oxidoreductases (EC 1) and lyases (EC 4) (Figure 7a). We then

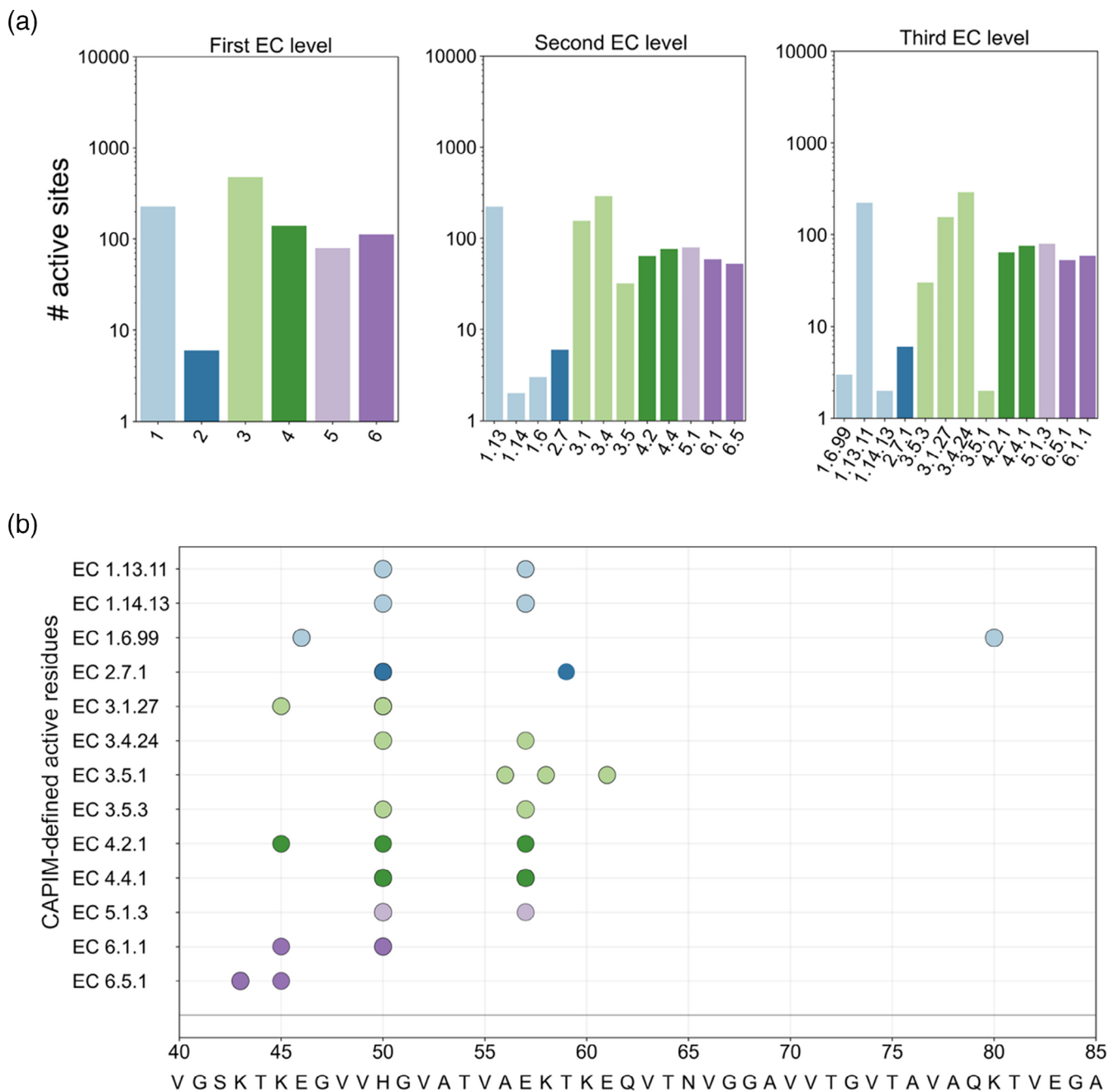


FIGURE 7 CAPIM exploration of α -synuclein amyloid structure. (a) Bar plots showing the predicted activities for 6A6B at the first, second, and third Enzyme Commission (EC) level for 6A6B exploring all EC numbers. (b) CAPIM-defined active-site residues for the EC activities found at the third EC level (here, only residues in the active site with the best fitness was plotted per EC activity). Chain identity was not considered here, all identified residues are shown on the same peptide sequence X-axis. However, active sites may contain multiple identical residues on different chains and/or residues from different chains.

delved into EC level 2, which specifies the types of bonds or functional groups involved. At EC level 2, the most frequent subclasses included EC 3.4 (peptidases), EC 3.1 (hydrolases acting on ester bonds), and EC 1.13 (oxidoreductases). Importantly, as EC 3.1 includes both phosphatases and esterases, it directly aligns with previous *in vitro* data (Horvath & Wittung-Stafshede, 2023).

At the detailed EC level 3, CAPIM predicted a range of specific activities, including EC 3.1.27 (endoribonucleases), EC 3.4.24 (metallopeptidases), and EC 1.13.11 (dioxygenases) (Figure 7a). EC 3.1.27 represents a subclass of hydrolases that cleave phosphodiester bonds. This prediction supports prior experimental data implying that α -synuclein amyloids cleaved phosphoester bonds in DNA under

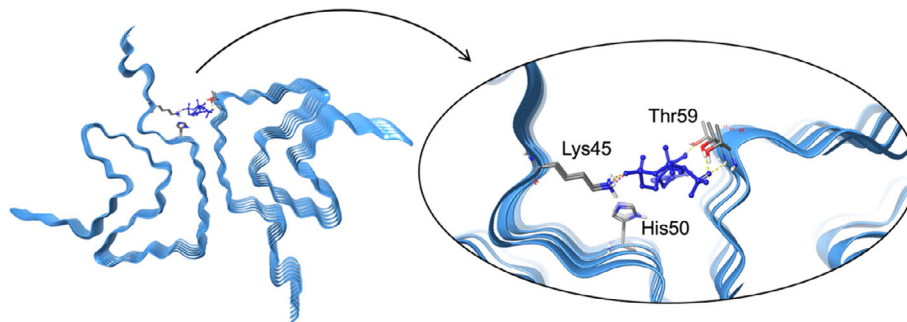


FIGURE 8 Substrate docking to CAPIM-identified binding site in amyloid. The 6A6B amyloid structure with the pNPP substrate (blue) docked to the pocket predicted by CAPIM to harbor most Enzyme Commission (EC) active sites, including EC 3.1 activity (e.g., dephosphorylation). The zoom-in (right) shows that pNPP interacts with three side chains (in stick) Lys45, His50 and Thr59. The first two residues are predicted to be active residues in several identified EC activities (see Figure 7b).

physiological conditions *in vitro* (Horvath et al., 2025). Furthermore, the prediction of metallopeptidase activity (EC 3.4.24) suggests that amyloid-mediated peptidase activity should be explored in future experiments as a function of metal ions. Earlier experimental findings have shown that copper ions can become incorporated into α -synuclein amyloids during aggregation and they can also bind to pre-form such amyloids (Walke et al., 2024). The predictions also provide other catalytic activities that may be assessed in the future (within EC 4, EC 5, and EC 6 classes).

The CAPIM result of EC 3.1, indicating phosphate bond manipulation including dephosphorylation, links experiments and predictions. Therefore, we docked the model phosphoester substrate para-nitrophenyl phosphate (pNPP), used in the *in vitro* experiments, to the 6A6B amyloids using CAPIM. We docked the substrate to the pocket identified by CAPIM to harbor EC 3.1 activity. Notably, it is clear from Figure 7b that most predicted catalytic activities involve His50 and a few other residues within the same cavity (Lys45 and Glu57) in different combinations. For example, EC 3.1.27 is predicted to involve His50 and Lys45, while EC 3.4.24 is predicted to involve His50 and Glu57. These residues (and Lys43 and Thr59 included in one active site each, Figure 7b) are all found in a cavity at the interface between the two protofilaments of the amyloid. Importantly, the same ligand binding site was identified in a recent computational study of 6A6B amyloids using a totally different approach (Parate et al., 2025).

Docking of pNPP using AutoDock in CAPIM identified pNPP interactions with His50, Lys45, and Thr59 (Figure 8) with a binding affinity of -8.3 kcal/mol. Notably, His50 was identified as essential for amyloid phosphorylation activity *in vitro* (Horvath & Wittung-Stafshede, 2023). Even if the predictions provide multiple possible catalytic activities, there appears to be a common active site (the cavity between the two protofilaments that is lined with His50, Lys45, and Asp57) in the amyloid structure that mediates activity. Taken

together, CAPIM not only captures the known dephosphorylation activity of α -synuclein amyloids, but also proposes novel enzymatic reactivities, such as peptidase activity, that can now be experimentally tested using model substrates.

4 | DISCUSSION

CAPIM introduces a comprehensive and unified approach for enzymatic activity and catalytic site analysis by integrating three established tools—P2Rank (Jakubec et al., 2022; Krivák & Hoksza, 2018), GASS (Izidoro et al., 2015; Moraes et al., 2017), and AutoDock Vina (Eberhardt et al., 2021; Trott & Olson, 2010)—into a streamlined pipeline. Unlike conventional tools that focus exclusively on either functional annotation (Dalkiran et al., 2018) or spatial prediction (Jakubec et al., 2022; Le Guilloux et al., 2009), CAPIM bridges and complements these domains, enabling users to identify catalytically active residues with their EC numbers and binding pockets and perform enzyme–substrate docking within the same framework. This makes CAPIM not only a predictive tool but also a complete exploratory platform for functional enzymology.

A key innovation in CAPIM is the combination of P2Rank and GASS, which allows it to prioritize catalytically relevant residues based on both spatial context and functional annotation. By narrowing residue selection to those residing within high-confidence P2Rank-predicted pockets, the pipeline substantially reduces the number of false positives typically produced by the predictions of GASS alone. For example, in the case studies we observed CAPIM predicts active sites in the order of a couple of dozens while GASS alone predicts in the order of more than a thousand. The case of papain (1BP4) further highlights CAPIM's robustness under minimal prior annotation: even when supplied only with the top-level EC class, CAPIM successfully

refined its prediction down to EC 3.4.22, correctly identifying key active sites with high confidence.

Another feature of CAPIM is its compatibility with large multimeric and polymeric proteins. GASS limits the analysis to eight chains, missing inter-chain active sites that may be crucial in larger enzymes. Here, CAPIM successfully analyzed the amyloid structure **6A6B** that includes 12 chains in the PDB file.

In addition to its technical strengths, CAPIM was designed with accessibility in mind. The tool includes a graphical user interface (GUI) built with Streamlit (Streamlit Inc., 2025), which requires no programming skills or machine learning expertise to operate. Users can simply upload a protein structure and ligands, and run the entire analysis pipeline through an intuitive, interactive interface. This accessibility greatly lowers the barrier to entry for researchers from diverse backgrounds—including biochemists, structural biologists, and experimental enzymologists—making CAPIM a practical choice for routine use in both academic and applied research settings.

CAPIM also enhances interpretability through its integrated visualization capabilities. Residue-level catalytic annotations, pocket maps, and docking results are presented in both sequence-level and structure-space, providing researchers with clear insights into the spatial distribution of enzymatic activity. This interactive presentation is particularly valuable for identifying functionally critical regions and designing follow-up mutational or biochemical studies.

The docking functionality in CAPIM is guided by structural prediction. By using P2Rank-generated pocket coordinates as docking grids, CAPIM ensures that substrates are evaluated within biologically relevant regions of the protein. For our case study proteins **1ARB** and **1H74**, we were able to prove that the CAPIM predicted substrate pose matched the experimentally proven one.

Despite its strengths, CAPIM's current limitations are rooted in the dependency on the GASS template database. Since GASS relies on a finite and predefined set of structural templates derived from known enzymes, the tool may underpredict activities associated with less characterized or novel folds. Future improvements could integrate contrastive learning-based EC predictors to better predict function in regions where annotations are sparse or inconsistent. Additionally, future versions of CAPIM could benefit from incorporating structure-based GNNs to directly learn catalytic patterns from the spatial and physicochemical context of residues. Representing protein structures as graphs—where nodes correspond to atoms or residues and edges capture spatial or functional proximity—would allow the model to identify subtle, fold-dependent features that template-based methods might miss. Beyond these predictive components, CAPIM's docking module also offers opportunities for expansion.

While in this study we focused on demonstrating docking with known substrates, future applications could include docking into predicted protein structures, designer proteins, testing non-cognate ligands, or comparing docking results for ligands corresponding to several top predicted functions. Such approaches may provide additional cues for discriminating between competing functional predictions and further enhance CAPIM's utility as a hypothesis-generating tool.

Another limitation is during the docking step. For PDB files that contain structural issues such as missing atoms, we recommend pre-processing to avoid errors. In cases where the removal of extra molecules is not enough to allow sufficient file preparation for AutoDock, we recommend that the user download the predicted pocket grids and use other available docking software. Future versions could include various docking options to allow wider coverage of input protein structures.

In summary, CAPIM offers a comprehensive and accessible solution for protein catalytic activity analysis. By reducing the candidate space, improving annotation precision, supporting large and complex protein structures, providing intuitive visualizations, and enabling docking-based validation—all within a user-friendly interface—CAPIM stands as a powerful tool for advancing enzymology, protein engineering, and drug discovery.

5 | CONCLUSION

We have created a tool that will help researchers more efficiently explore enzyme activity in a streamlined fashion. CAPIM combines three tools into one merged user-friendly engine. This approach helps narrow down the search for active sites in proteins with respect to both location and activity. We show using three known examples that the correct answer (as determined by experiments) is retained while false positives are significantly reduced. For an unknown case, α -synuclein amyloids, we show how CAPIM can be used to identify putative new activities that can act as the basis for exploration by strategic experimental assays *in vitro*.

AUTHOR CONTRIBUTIONS

Gökhan Özari: Conceptualization; investigation; writing – original draft; writing – review and editing; methodology; software. **Daniela A. Garcia-Soriano:** Conceptualization; investigation; writing – original draft; writing – review and editing; visualization; methodology; software. **Shraddha Parate:** Validation; visualization; writing – review and editing. **Amar el Issaoui:** Validation; visualization; writing – review and editing. **Pernilla Wittung-Stafshede:** Conceptualization; investigation; writing – original draft; writing – review and editing; methodology; supervision; funding acquisition.

ACKNOWLEDGMENTS

We thank Istvan Horvath for helpful contributions in the early stage. Chalmers e-Commons, the Swedish Research Council, Knut and Alice Wallenberg Foundation, and the Swedish Cancer Foundation are acknowledged for funding. We also thank SciLifeLab Serve for providing hosting of the CAPIM application.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Pernilla Wittung-Stafshede  <https://orcid.org/0000-0003-1058-1964>

REFERENCES

- Corso G, Stärk H, Jing B, Barzilay R, Jaakkola T. DiffDock: diffusion steps, twists, and turns for molecular docking. 2023 <https://doi.org/10.48550/arXiv.2210.01776>
- Dalkiran A, Rifaioglu AS, Martin MJ, Cetin-Atalay R, Atalay V, Doğan T. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinform.* 2018;19(1):334. <https://doi.org/10.1186/s12859-018-2368-y>
- Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: new docking methods, expanded force field, and python bindings. *J Chem Inf Model.* 2021;61(8):3891–8. <https://doi.org/10.1021/acs.jcim.1c00203>
- Frey L, Ghosh D, Qureshi BM, Rhyner D, Guerrero-Ferreira R, Pokhama A, et al. On the pH-dependence of α -synuclein amyloid polymorphism and the role of secondary nucleation in seed-based amyloid propagation. *Elife.* 2024;12:RP93562.
- Hernandez M, Ghersi D, Sanchez R. SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res.* 2009;37(Suppl_2):W413–6.
- Horvath I, Aning OA, Sriram KK, Rehnberg N, Chawla S, Molin M, et al. Biological amyloids chemically damage DNA. *ACS Chem Neurosci.* 2025;16:355–64.
- Horvath I, Wittung-Stafshede P. Amyloid fibers of α -synuclein catalyze chemical reactions. *ACS Chem Neurosci.* 2023;14(4):603–8. <https://doi.org/10.1021/acschemneuro.2c00799>
- Izidoro SC, De Melo-Minardi RC, Pappa GL. GASS: identifying enzyme active sites with genetic algorithms. *Bioinformatics.* 2015;31(6):864–70. <https://doi.org/10.1093/bioinformatics/btu746>
- Jakubec D, Skoda P, Krivak R, Novotny M, Hoksza D. PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures. *Nucleic Acids Res.* 2022;50(W1):W593–7. <https://doi.org/10.1093/nar/gkac389>
- Kandel J, Tayara H, Chong KT. PURESNet: prediction of protein-ligand binding sites using deep residual neural network. *J Chem.* 2021;13(1):65. <https://doi.org/10.1186/s13321-021-00547-7>
- Krishna SS, Zhou T, Daugherty M, Osterman A, Zhang H. Structural basis for the catalysis and substrate specificity of homoserine kinase. *Biochemistry.* 2001;40(36):10810–8.
- Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J Chem.* 2018;10(1):39. <https://doi.org/10.1186/s13321-018-0285-8>
- LaLonde JM, Zhao B, Smith WW, Janson CA, DesJarlais RL, Tomaszek TA, et al. Use of papain as a model for the structure-based design of cathepsin K inhibitors: crystal structures of two papain-inhibitor complexes demonstrate binding to S'-subsites. *J Med Chem.* 1998;41(23):4567–76.
- Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinform.* 2009;10(1):168.
- Moraes JPA, Pappa GL, Pires DEV, Izidoro SC. GASS-WEB: a web server for identifying enzyme active sites based on genetic algorithms. *Nucleic Acids Res.* 2017;45(W1):W315–9. <https://doi.org/10.1093/nar/gkx337>
- Parate S, Buratti F, Eriksson LA, Wittung-Stafshede P. In silico identification of substrate binding sites in type-1A α -synuclein amyloids. *Biophys J.* 2025;124:2418–27.
- Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc Natl Acad Sci U S A.* 2019;116(28):13996–4001. <https://doi.org/10.1073/pnas.1821905116>
- Streamlit Inc. Streamlit: An app framework for Machine Learning and Data Science. <https://streamlit.io/> (2025). Accessed 2007
- Tan H, Wang Z, Hu G. GAABind: a geometry-aware attention-based network for accurate protein-ligand binding pose and binding affinity prediction. *Brief Bioinform.* 2024;25(1):bbad462.
- Tian W, Chen C, Lei X, Zhao J, Liang J. CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Res.* 2018;46(W1):W363–7.
- Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem.* 2010;31(2):455–61. <https://doi.org/10.1002/jcc.21334>
- Tsunasawa S, Masaki T, Hirose M, Soejima M, Sakiyama F. The primary structure and structural characteristics of *Achromobacter lyticus* protease I, a lysine-specific serine protease. *J Biol Chem.* 1989;264(7):3832–9.
- The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 2018;46(5):2699.
- Walke G, Kumar R, Wittung-Stafshede P. Copper ion incorporation in α -synuclein amyloids. *Protein Sci.* 2024;33(4):e4956.
- Xia C-Q, Pan X, Shen H-B. Protein-ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data. *Bioinformatics.* 2020;36(10):3018–27.
- Xia Y, Xia C, Pan X, Shen H. BindWeb: a web server for ligand binding residue and pocket prediction from protein structures. *Protein Sci.* 2022;31(12):e4462. <https://doi.org/10.1002/pro.4462>
- Yu T, Cui H, Li JC, Luo Y, Jiang G, Zhao H. Enzyme function prediction using contrastive learning. *Science.* 2023;379:1358–63.

How to cite this article: Özsari G, García-Soriano DA, Parate S, el Issaoui A, Wittung-Stafshede P. CAPIM: Catalytic activity and site prediction and analysis tool in multimer proteins. *Protein Science.* 2025;34(11):e70347. <https://doi.org/10.1002/pro.70347>