

CANED: A PyMOL Plugin for Computer-Aided De Novo Enzyme Design

Pinyang Zhang^{1*}, Yan Ming¹

¹*College of Biotechnology and Pharmaceutical Engineering, Nanjing Tech University, Nanjing, China*

**Corresponding Author. Email: pinyangzhang1@gmail.com*

Abstract: CANED is a PyMOL plugin designed to streamline de novo enzyme design by integrating advanced computational tools into a user-friendly interface. It provides menu bar navigation and click-based operations, with standardized file output formats. Combining Rosetta's physical modeling and LigandMPNN's sequence prediction capabilities, CANED organizes the enzyme design workflow into three modular stages: File Preparation, Backbone Design, and Sequence Design. It supports preparation of multiple file formats, provides Rosetta's matching methods to generate enzyme backbones, and offers diverse sequence design methods. Dynamic table-based analysis and real-time 3D visualization provide a more intuitive way to select high-quality design results. By reducing the need for computational expertise and command-line operations, CANED makes enzyme design accessible to a broader range of researchers, advancing applications in pharmaceuticals, biocatalysis, and sustainable chemistry.

Keywords: PyMOL Plugin, De Novo Enzyme Design, Visualization, Rosetta, LigandMPNN

1. Introduction

Novel enzyme design drives innovation in computational biology, enabling tailored enzymes for pharmaceuticals, biocatalysis, and environmental applications [1, 2]. Computational approaches, such as Rosetta's physical modeling of protein folding and catalysis and deep learning models like LigandMPNN and ProteinMPNN for sequence prediction, have transformed enzyme engineering [3-5]. Rosetta's workflow defines theoretical enzymes through constraint files, matches catalytic scaffolds, and optimizes sequences, demonstrating notable success in creating catalytically active enzymes, such as those for retro-aldol reactions and Kemp elimination catalysis, which showcase its ability to design functional enzymes from scratch [6-8]. Similarly, deep learning tools like LigandMPNN have excelled in designing proteins with high affinity for small molecules like rocuronium and cholic acid, outperforming other methods in engineering residues critical for small molecule interactions, thus broadening the scope of enzyme design for precise molecular recognition [4].

Protein design tools predominantly rely on command-line interfaces. Rosetta executes tasks through XML scripting or PyRosetta programming [9]. Its workflow requires preparing various input files, involving steps such as defining catalytic constraints or parameterizing ligands [9]. Rosetta's output evaluation involves parsing scoring files with metrics like total energy or binding energy using command-line tools, such as sort or awk [10]. Other enzyme design methods, such as deep learning

models, similarly operate via command-line interfaces, analyzing outputs like sequences and confidence scores with command-line scripts or external software [11]. These operational characteristics may pose challenges for users without specialized expertise and can contribute to a fragmented workflow.

Recent studies highlight the need for integrated, accessible tools to streamline enzyme design [3,12]. Visualization platforms like PyMOL [13], widely used for molecular graphics, offer a promising foundation for such tools due to their robust visualization capabilities and flexible plugin architecture [12]. By integrating computational design methods into a unified, visualization-driven interface, it is possible to reduce technical barriers and enhance the efficiency of novel enzyme design.

2. Interface and functionality

CANED is a PyMOL plugin for computer-aided de novo enzyme design. The plugin features a visualization interface, integrating computational tools into PyMOL's molecular visualization environment. CANED employs a modular design, organizing the enzyme design workflow into distinct functional pages, each targeting a specific stage: File Preparation, Structural Design, and Sequence Design. A menu-driven navigation system facilitates seamless transitions between pages, while standardized file formats ensure compatibility across stages, allowing outputs from one step to be directly used in the next. File generation processes are enhanced with pop-up windows for real-time feedback. All generated files across the plugin include download buttons for easy access and modification.

2.1. File Preparation page

The File Preparation page (Figure 1 a) simplifies the creation of input files required for Rosetta-based enzyme design, including theozyme definitions, substrate parameter files, enzyme scaffold PDB files, and active site files, which are essential for computational enzyme design workflows [6, 9]. These tools are accessible through an intuitive, click-based interface, making file setup accessible to users without extensive computational expertise.

Theozymes encapsulate the geometric and chemical constraints of residues critical to substrate interactions, defining the ideal catalytic geometry for enzyme activity [6]. CANED includes a tool, based on the `match_ligand_binding_sites` codebase, that generates geometric constraints from protein-substrate complex structures, streamlining theozyme definition [14]. If chemical constraint information needs to be defined, it requires manual input based on external data, such as literature or experimental insights [6].

Substrate Parameter File: This file defines the chemical properties of small molecule substrates, including bond connectivity, for Rosetta compatibility [9]. CANED leverages OpenBabel to generate rotational isomers and Rosetta's `molfile_to_params.py` script to produce parameter files [15]. OpenBabel ensures precise chemical handling, while `molfile_to_params.py` converts substrates into Rosetta-compatible formats [15].

Enzyme Scaffold PDB File: In Rosetta enzyme design, scaffold files provide the protein backbone structure that serves as the foundation for catalytic site design, requiring standardization to ensure compatibility with Rosetta's algorithms [6, 9]. CANED uses Rosetta's `clean_pdb.py` script to process PDB files, removing non-standard amino acid residues and generating standardized scaffold files ready for downstream structural and sequence design [15].

Active Site File: Active site files in Rosetta define the specific residues that interact with the substrate, guiding the placement of catalytic sites within the scaffold to optimize substrate binding and catalysis [6]. CANED's tool processes the scaffold PDB files and substrate PDB files to generate active site files, ensuring accurate configuration of the enzyme's catalytic site [14, 15].

2.2. Structural Design page

The Structural Design page focuses (Figure 1 b) on identifying protein scaffolds that meet the theozyme's catalytic requirements, providing the initial backbone structure for enzyme design. CANED integrates Rosetta's Match application, a computational tool designed to identify protein scaffolds that satisfy the theozyme's geometric constraints by matching catalytic site geometries to existing protein structures [6, 16]. The interface simplifies input selection through a click-driven design, allowing users to easily specify the required files. Users input the constraint file, substrate parameter file, active site file, and scaffold PDB file generated from the File Preparation stage. This process produces candidate scaffold backbones that align with the catalytic conformation.

2.3. Sequence Design page

The Sequence Design stage consists of two distinct pages—Rosetta enzyme design page (Figure 1 c) and LigandMPNN design page (Figure 1 d)—each employing specialized computational tools to optimize enzyme sequences. These pages leverage files from previous stages in an integrated workflow, where results from the Match page can be used in the Enzyme Design page after selecting a substrate model, and outputs from the Enzyme Design page can be directly utilized in the LigandMPNN and FastRelax page, ensuring a cohesive and efficient design process.

The Rosetta Design application uses an energy-based scoring function to optimize protein sequences and all-atom coordinates by minimizing the energy of the protein-substrate complex [2, 15]. The Enzyme Design page integrates this application to optimize the active site's residue types and side-chain conformations for desired catalytic properties. The process generates multiple design outputs, accompanied by a single scoring file (score.sc) that lists the structural stability scores for all outputs based on Rosetta's energy function.

The LigandMPNN model employs a machine learning-based approach to predict protein sequences with high substrate affinity by modeling interactions with non-protein components, such as small molecules [4]. The FastRelax application refines protein structures by minimizing energy through iterative packing and minimization, optimizing side-chain and backbone conformations [17, 18]. The LigandMPNN design page integrates these tools to design enzyme sequences via LigandMPNN and generate corresponding refined structures via FastRelax for optimal catalytic performance. The interface allows specification of conserved residues, such as catalytic residues from the constraint file, to ensure their retention in the predicted sequences, thereby preserving critical functional sites for enzymatic activity. For structures derived from CANED's Match or Enzyme Design steps, conserved residues are automatically assigned, streamlining the workflow. For external structures, Match constraints must be manually incorporated into the structure file. The interface supports selection of model parameters, chain specifications, and filtering options, with thresholds configurable to streamline candidate selection. Outputs include sequence variants designed by LigandMPNN, accompanied by a sequence file containing confidence scores and other metrics, as well as structures refined by FastRelax and a scoring file (combined_scores.sc) with metrics such as CST_score.

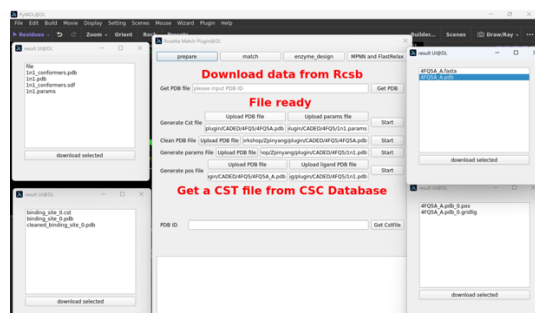
2.4. Result Evaluation functional area

Result Evaluation is embedded within the Rosetta Enzyme Design page (Figure 1 c) and LigandMPNN design page (Figure 1 d), offering intuitive tools to analyze and select high-quality designs. Unlike methods that rely on command-line tools to parse Rosetta and LigandMPNN scoring files, CANED provides a visualization-driven alternative.

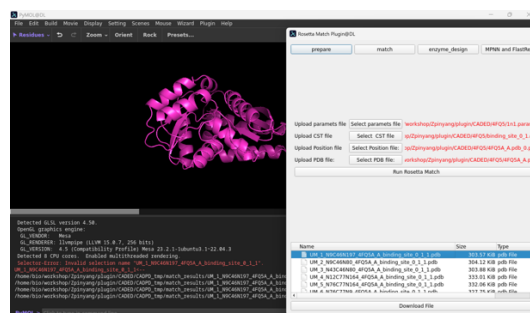
Dynamic Table-Based Analysis: The system parses scoring files, including the score.sc file from Rosetta Design, the sequence file from LigandMPNN containing confidence scores and metrics, and

the combined_scores.sc file from FastRelax with metrics such as CST_score, into dynamic tables. Clickable column headers, such as total_score or CST_score, enable rapid sorting to identify top-performing designs.

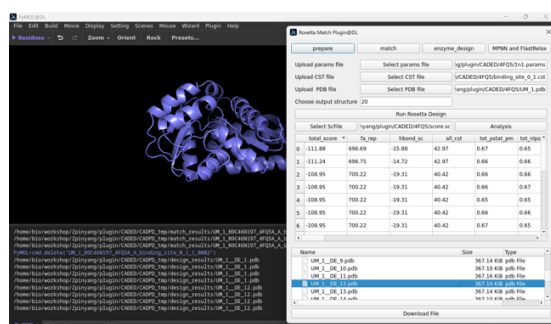
Real-Time 3D Visualization: The interface instantly visualizes selected structures within PyMOL's 3D environment, eliminating the need for manual PDB imports. This functionality enables examination of structural features, such as active site conformations, directly within the interface.



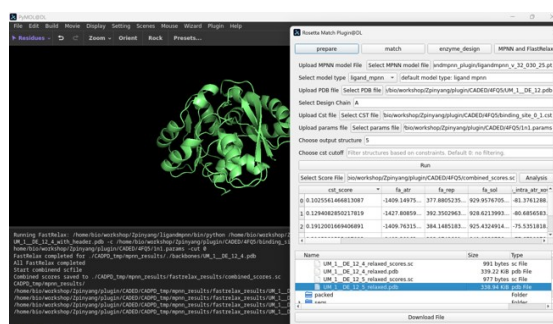
(a) File preparation page



(b) Structural design page



(C) Rosetta enzyme design page



(C) LigandMPNN design page

Figure 1: CANED interface

3. Discussion

CANED helps mitigate challenges in de novo enzyme design by integrating Rosetta and LigandMPNN into PyMOL's visualization environment, providing a cohesive and accessible platform for researchers. CANED's modular design and click-driven interface facilitate a streamlined workflow—from file preparation to result evaluation. The plugin's ability to automatically assign conserved residues for CANED-generated structures and provide real-time 3D visualization reduces manual effort and supports efficient design iteration. Dynamic table-based analysis of scoring files simplifies the identification of high-quality designs, reducing reliance on command-line utilities. However, Manual input of chemical constraints and Match constraints for external structures requires external data or expertise, potentially limiting accessibility for novices. Additionally, the computational demands of Rosetta and LigandMPNN may pose challenges for users with limited hardware resources. Compared to standalone tools, CANED's integration within PyMOL ensures a visualization-driven experience but may restrict compatibility with non-PyMOL workflows.

Future enhancements could involve automating chemical constraint generation using the PLIP tool to analyze protein-ligand interactions [19], incorporating additional design tools like ProteinMPNN, and improving computational efficiency to enhance accessibility. Supporting cloud-based computing might further broaden access to the plugin. By helping to balance computational capabilities with

user-friendliness, CANED could facilitate advancements in enzyme engineering, potentially contributing to applications in drug development, biocatalysis, and environmental sustainability.

4. Code availability

Project Repository: <https://github.com/OrangeCatzhang/CANED>

References

- [1] Huang, P.-S., Boyken, S. E., & Baker, D. (2016). The coming of age of de novo protein design. *Nature*, 537(7620), 320–327. <https://doi.org/10.1038/nature19946>
- [2] Kuhlman, B., & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences*, 97(19), 10383–10388. <https://doi.org/10.1073/pnas.97.19.10383>
- [3] Leman, J. K., Weitzner, B. D., Lewis, S. M., Adler, J., Dunbrack, R. L., & Kuhlman, B. (2020). Macromolecular modeling and design in Rosetta: Recent methods and frameworks. *Nature Methods*, 17(7), 665–680. <https://doi.org/10.1038/s41592-020-0848-2>
- [4] Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., ... Baker, D. (2025). Atomic context-conditioned protein sequence design using LigandMPNN. *Nature Methods*, 22(2), 139–147. <https://doi.org/10.1038/s41592-025-02626-1>
- [5] Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., ... Baker, D. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615), 49–56. <https://doi.org/10.1126/science.abb2187>
- [6] Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S., & Baker, D. (2011). De novo enzyme design using Rosetta3. *PLoS ONE*, 6(5), e19230. <https://doi.org/10.1371/journal.pone.0019230>
- [7] Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas III, C. F., Hilvert, D., Houk, K. N., Stoddard, B. L., & Baker, D. (2008). De novo computational design of retro-aldol enzymes. *Science*, 319(5868), 1387–1391. <https://doi.org/10.1126/science.1152692>
- [8] Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S., & Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature*, 453(7192), 190–195. <https://doi.org/10.1038/nature06879>
- [9] Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H., & Meiler, J. (2010). Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You. *Biochemistry*, 49(14), 2987–2998. <https://doi.org/10.1021/bi902153g>
- [10] Ford, A. S., Weitzner, B. D., & Bahl, C. D. (2019). Integration of the Rosetta suite with the python software stack via reproducible packaging and core programming interfaces for distributed simulation. *Protein Science*, 29(1), 43–51. <https://doi.org/10.1002/pro.3721>
- [11] Feehan, R., Montezano, D., & Slusky, J. S. G. (2021). Machine learning for enzyme engineering, selection and design. *Protein Engineering, Design and Selection*, 34, gzab019. <https://doi.org/10.1093/protein/gzab019>
- [12] Sequeiros-Borja, C. E., Surpeta, B., & Brezovsky, J. (2021). Recent advances in user-friendly computational tools to engineer protein function. *Briefings in Bioinformatics*, 22(3), bbaa150. <https://doi.org/10.1093/bib/bbaa150>
- [13] PyMOL Community. (2024). PyMOL Wiki. Retrieved from <https://pymolwiki.org>
- [14] Pan, X., & Kortemme, T. (2021). De novo protein fold families expand the designable ligand binding site space. *PLoS Computational Biology*, 17(11), e1009620. <https://doi.org/10.1371/journal.pcbi.1009620>
- [15] Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., ... & Baker, D. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, 13(6), 3031–3048. <https://doi.org/10.1021/acs.jctc.7b00125>
- [16] Zanghellini, A., Jiang, L., Wollacott, A. M., Cheng, G., Meiler, J., Althoff, E. A., Röthlisberger, D., & Baker, D. (2006). New Algorithms for Computational Enzyme Design. *Journal of Molecular Biology*, 362(4), 787–802. <https://doi.org/10.1016/j.jmb.2006.07.046>
- [17] Nivón, L. G., Moretti, R., & Baker, D. (2013). A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLoS ONE*, 8(4), e59004. <https://doi.org/10.1371/journal.pone.0059004>
- [18] Conway, P., Tyka, M., DiMaio, F., Konerding, D. E., & Baker, D. (2014). Relaxation of Backbone Bond Geometry Improves Protein Energy Landscape Modeling. *Protein Science*, 23(1), 47–55. <https://doi.org/10.1002/pro.2389>
- [19] Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F., & Schroeder, M. (2015). PLIP: Fully automated protein–ligand interaction profiler. *Nucleic Acids Research*, 43(W1), W443–W447. <https://doi.org/10.1093/nar/gkv315>