AI-Optimized Enzyme Design for Biomass Breakdown

Ahmed Omar Bahaj
College of Computer Science and Engineering, University of Jeddah
Jeddah, Saudi Arabia
ahmedomarbahaj@gmail.com

Mohamad Alim Alamine
Faculty of Computer Science and Information Systems
Medina, Saudi Arabia
alaminerca@gmail.com

Nabil Abdulkafi Alhamawi
Faculty of Computer Studies, Arab Open University
Jeddah, Saudi Arabia
nabil_ha@yahoo.com

Supervised by: Fares Fourati

Table of Contents

Table of Contents....2
Acknowledgments....3
Executive Summary....4
Objectives....5
Group Members and Their Contributions....5
Milestones Achieved....6
Methodology....7
Enzyme Candidate Generation....7
Ligand Preparation....7
Structure Prediction & Quality Control....7
Molecular Docking....7
Pipeline Architecture....8
Preliminary Results....9
Structural Confidence....9
Docking Analysis....10
Top Candidate....11
Challenges....12
Future Work....12
Resources and Tools....12
References....13

أكاديمية كاوست
KAUST ACADEMY

Acknowledgments

We would like to express our sincere gratitude to the organizers of this training program, **King Abdullah University of Science and Technology (KAUST)** and the **KAUST Academy**, for providing this invaluable opportunity to delve into the advanced fields of Bioinformatics and Artificial Intelligence. We extend our thanks to our host institution, **King Khalid University,** for providing the facilities and a conducive learning environment. We are especially grateful to our project mentor, Dr. Fares Fourati, whose guidance, expertise, and insightful feedback were instrumental in navigating the complexities of our project and ensuring its successful completion.

أكاديمية كاوست
KAUST ACADEMY

Executive Summary

This report documents the successful completion of the capstone project undertaken as part of the KAUST Academy Artificial Intelligence Program. The project directly addressed the program's core mission of applying cutting-edge AI and computational science to solve high-impact, real-world problems.

Our team focused on a critical challenge in the sustainable energy sector: the inefficiency of converting algal biomass into biofuels. Leveraging the advanced skills in bioinformatics and AI acquired during this program, we designed and executed a novel, high-throughput computational pipeline. This pipeline integrates a finetuned generative AI model (Progen2) for novel enzyme design, the state-of-the-art structure prediction capabilities of ESMFold, and large-scale molecular docking simulations (AutoDock Vina) to rapidly screen for high-potential enzyme candidates.

The project successfully moved from a broad scientific challenge to the identification of tons specific, high-efficacy enzyme designs which show exceptional promise for degrading key components of algal cell walls. This work serves as a powerful demonstration of how AI-driven discovery can dramatically accelerate the initial phases of enzyme engineering, providing a direct, data-driven pathway for future experimental validation in the development of economically viable biofuels.

The following pages provide a detailed account of our methodology, the challenges overcome, the final results, and the key learnings from this intensive and highly rewarding training experience.

أكاديمية كاوست
KAUST ACADEMY

Objectives

The primary objective of this project is to enhance the efficiency of converting algal and aquatic plant biomass into biofuels through the design and computational validation of novel enzymes. Specific goals include:

1. Generate novel enzyme amino acid sequences using a finetuned AI model, targeting key families (GH10, GH11, GH5, GH48, CE1, PL1 and PL7) for breaking down cellulose, hemicellulose, and algal polysaccharides.
2. Develop a computational pipeline to screen enzyme candidates via structure prediction and molecular docking, identifying those with high binding affinity to biomass substrates.
3. Create "super-charged" enzyme designs to digest tough plant and algal material, enabling cheaper, more sustainable biofuels.
4. Deliver a set of validated, high-priority enzyme designs for experimental validation.

Group Members and Their Contributions

1. **Ahmed Omar Bahaj** (University of Jeddah, College of Computer Science and Engineering): AI Model Management & Sequence Generation. Responsible for fine tuning the Progen2 model, generating candidate enzyme sequences, and initial data curation.
2. **Mouhamad Alim Alamine** (Islamic University of Madinah, Faculty of Computer Science and Information Systems): Pipeline Development & Docking Simulation. Developed the high-throughput Python pipeline, managed structure prediction with ColabFold, and ran many-to-many molecular docking simulations with AutoDock Vina.
3. **Nabil Abdulkafi Alhamwi** (Arab Open University, Faculty of Computer Studies): Data Analysis & Reporting. Conducted analysis of ESMFold quality metrics and docking affinity results, generated data visualizations, and led report compilation.

Milestones Achieved

The computational discovery phase has been successfully completed, meeting all initial objectives. Key milestones include:

1. **AI Model Finetuning**: Finetuned the Progen2 model on a curated dataset of (GH10, GH11, GH5, GH48, CE1, PL1 and PL7) enzyme families.
2. **Candidate Generation**: Generated 700 enzyme sequences (100 per target family) for analysis.
3. **Pipeline Development**: Built a fully automated, high-throughput modular Python pipeline for processing multiple protein structures against multiple ligands.
4. **Structure Prediction**: Predicted 3D structures for all eight enzyme candidates using ESMFold.
5. **Quality Control**: Applied a rigorous filter based on ESMFold pLDDT scores, retaining only high-confidence structural models.
6. **High-Throughput Docking**: Conducted a many-to-many docking screen, simulating 365 unique protein-ligand interactions across seven biomass-derived substrates.
7. **Lead Candidate Identification**: Identified +30 standout enzyme candidates with exceptional binding affinities.

أكاديمية كاوست
KAUST ACADEMY

Methodology

The methodology integrates generative AI with structural bioinformatics to create a discovery funnel, narrowing down from many potential sequences to high-value targets.

Enzyme Candidate Generation

A 151M finetuned Progen2-Small model generated sequences targeting seven enzyme families: (GH10, GH11, GH5, GH48, CE1, PL1 and PL7), crucial for degrading algal and plant biomass components.

Ligand Preparation

Prepared a panel of seven substrates representing cellulose (Cellobiose, Glucose), hemicellulose (Xylobiose, Ferulic Acid), pectin (Galacturonic acid), and algal polysaccharides (Mannuronic acid, Laminaribiose) using RDKit and Open Babel for docking.

Structure Prediction & Quality Control

Used ESMFold by Meta AI to predict 3D structures of the all AI-generated sequences. Evaluated models using mean pLDDT scores, retaining only those with high confidence (typically >70) for docking.

Molecular Docking

Screened validated protein structures against the seven ligands using AutoDock Vina in a blind docking approach, with a search grid covering the entire protein surface. Binding affinity (ΔGbind in kcal/mol) was the primary evaluation metric.

KAUST ACADEMY

Pipeline Architecture

1. **Input**: Curated enzyme sequence dataset for Progen2 finetuning.
2. **Sequence Generation**: Progen2 generates candidate sequences for (GH10, GH11, GH5, GH48, CE1, PL1 and PL7) families.
3. **Structure Prediction**: ESMFold predicts 3D structures, outputting pLDDT scores for quality assessment.
4. **Quality Filter**: Retain structures with pLDDT > 70 for docking.
5. **Docking Simulation**: AutoDock Vina performs many-to-many docking against seven substrates.
6. **Output**: Binding affinity heatmap and ranked list of top candidates.

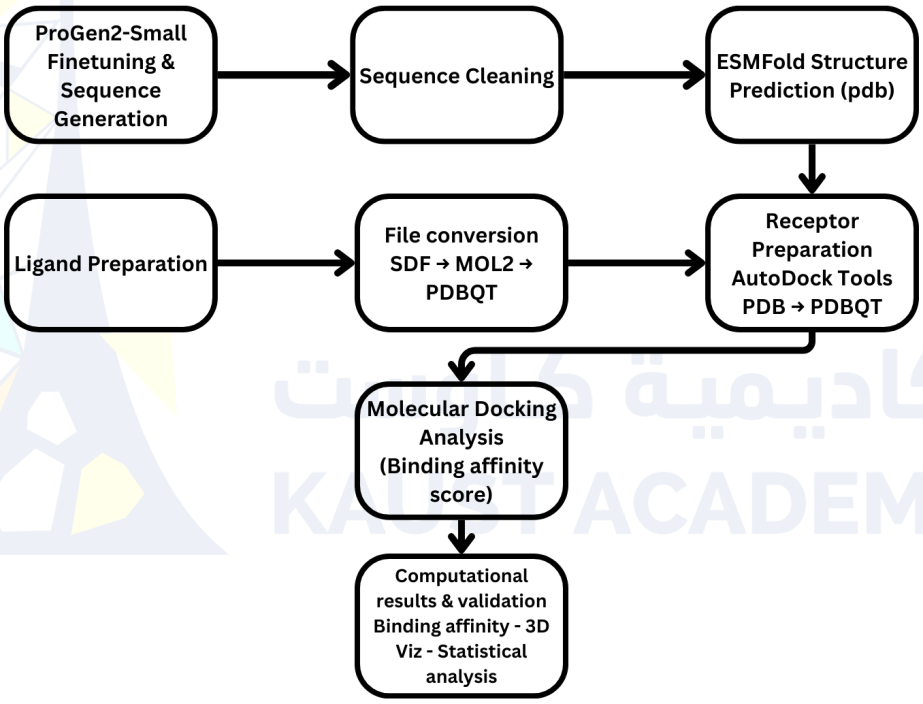


Figure 1. Overall pipeline architecture

Preliminary Results

Structural Confidence

We evaluated the average pLDDT of the generated enzyme variants across the seven targeted families (GH10, GH11, GH5, GH48, CE1, PL1 and PL7).

Figure 2. shows that GH11 (48.9) and CE1 (46.9) had the highest average pLDDT values, suggesting more reliable structural predictions. In contrast, PL7 (35.9) and GH10 (37.2) yielded the lowest scores, indicating weaker confidence.

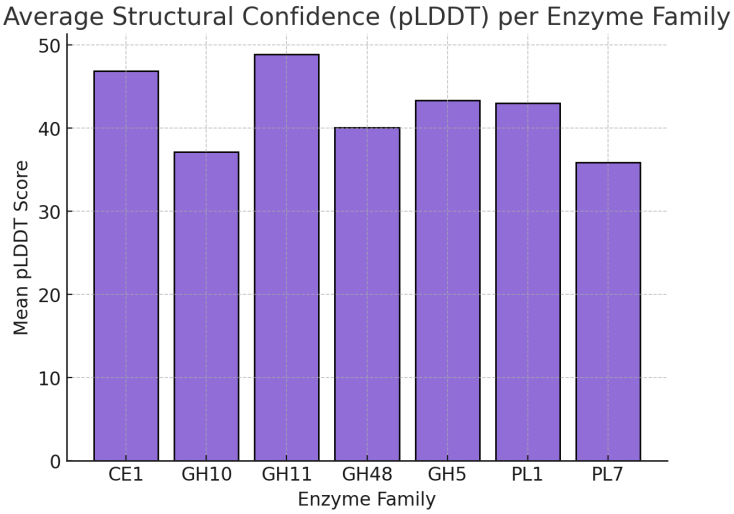


Figure 2. Average structural confidence (mean pLDDT) per enzyme family

This highlights that different enzyme families vary in how well AI-generated sequences fold into stable structures. To explore whether sequence size influences structural stability, we compared sequence length against mean pLDDT (**Figure 3**.)

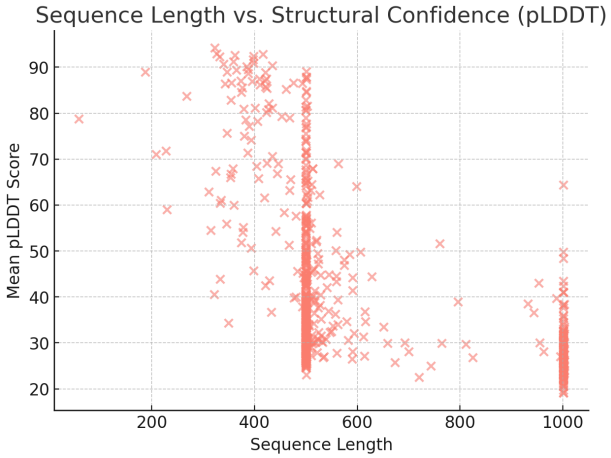


Figure 3. Scatter plot of sequence length versus structural confidence

The scatter plot shows **no strong correlation** between sequence length and structural confidence. Both short and long sequences produce a wide range of pLDDT values, from low to high. However, we notice there is a tendency for sequences of length ~400 to score higher pLDDT.

Docking Analysis

The many-to-many docking screen produced 365 protein-ligand interactions, visualized in the affinity heatmap below. Candidates with binding affinities ≤ -7.0 kcal/mol were prioritized for their strong interactions.

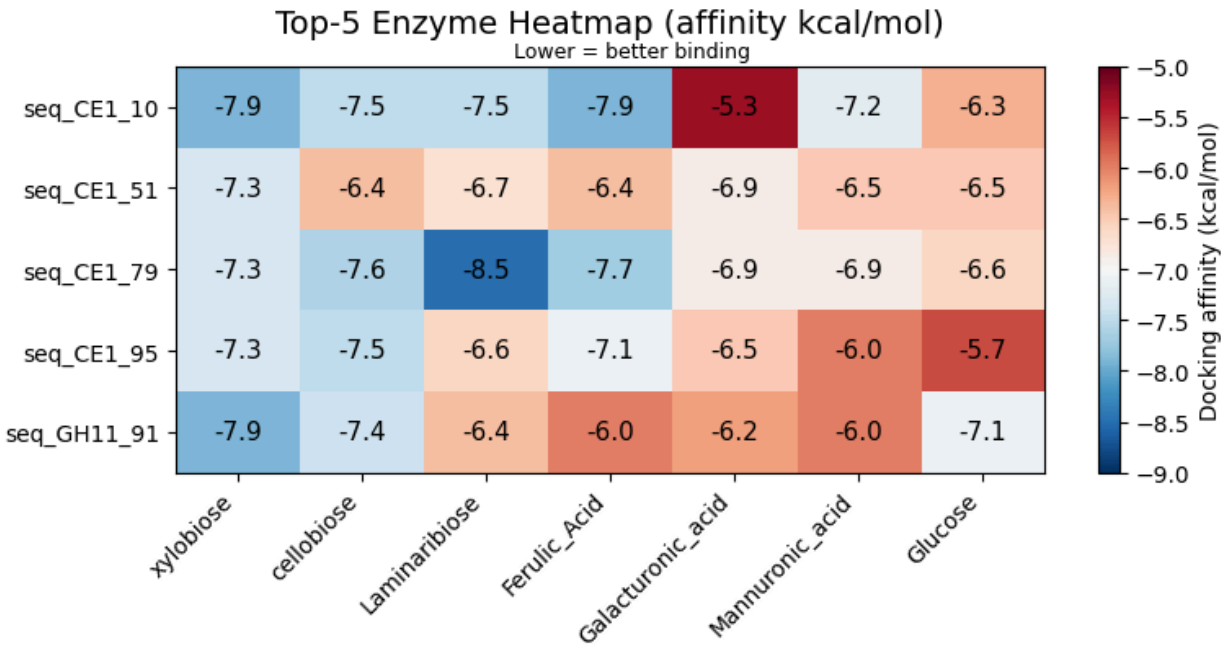


Figure 4. affinity heatmap of top 5 candidates

Top Candidate

Among all 700 generated sequences, **seq_CE1_79** stood out as the most reliable candidate, achieving a mean pLDDT score of 92.78. This indicates a very high-confidence structure prediction, suggesting that the enzyme is likely to fold into a stable 3D structure.

In addition to its structural reliability, docking simulations showed that seq_CE1_79 achieved strong binding affinities across multiple ligands. Notable, it demonstrated:

- Laminaribiose: -8.5 kcal/mol
- Ferulic Acid: -7.7 kcal/mol

These results place seq_CE1_79 among the top-performing candidates, combining both exceptional predicted structural stability and favorable substrate binding.

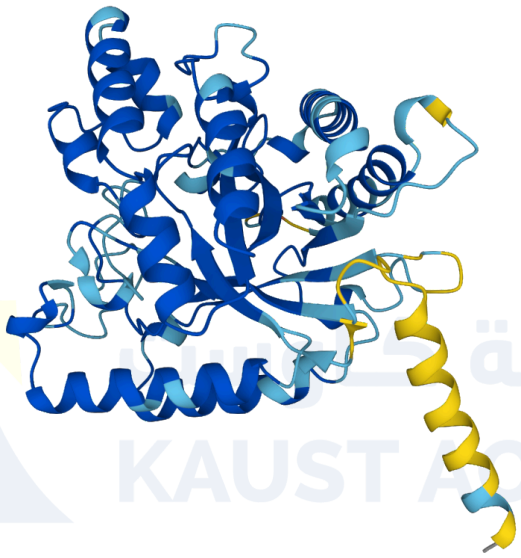


Figure 4. 3D Visualization of seq_CE1_79

Challenges

1. **AI Generation Reliability**: Initial AI-generated sequences showed variable structural integrity, requiring a strict pLDDT-based filter to ensure reliable candidates.
2. **Pipeline Robustness**: Early pipeline versions faced issues with file paths, API inconsistencies, and batch processing, resolved through iterative development.

Future Work

1. **Gene Synthesis and Protein Expression**: Synthesize genes for top candidates, expressing them in a suitable host (e.g., *E. coli* or *Pichia pastoris*).
2. **Enzymatic Assays**: Measure catalytic activity and efficiency on target substrates (e.g., ferulic acid-linked xylan for seq_CE1_79)
3. **Synergistic Testing**: Evaluate enzyme combinations for synergistic effects to enhance overall biomass breakdown.
4. **Structural Validation**: Pursue protein crystallization to validate ESMFold predictions and gain mechanistic insights.
5. **Computational Refinement:** Integrate molecular dynamics simulations to assess enzyme stability under industrial conditions and optimize binding interactions.
6. **Industrial Deployment:** Explore application of these AI-designed enzymes in Saudi Aramco's biofuel initiatives, supporting Saudi Arabia's Vision 2030 goals for renewable energy and economic diversification.

أكاديمية كاوست
KAUST ACADEMY

Resources and Tools

- **Computing Units**: KAUST Ibex Cluster
- **AI Sequence Generation**: Finetuned Progen2-Small Model
- **Structure Prediction**: ESMFold
- **Molecular Docking**: AutoDock Vina
- **Cheminformatics**: RDKit, Open Babel
- **Data Analysis & Scripting**: Python 3, Pandas, Biopython, Matplotlib, RDKit

References

- Hrbáň, H., & Hoksza, D. (2024). Protein family sequence generation through ProGen2 fine-tuning. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 7037–7039). IEEE. https://doi.org/10.1109/BIBM62325.2024.10821712
- Trott, O., & Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry, 31*(2), 455–461. https://doi.org/10.1002/jcc.21334
- Nijkamp, E., Ruffolo, J., Weinstein, E. N., Naik, N., & Madani, A. (2022). *ProGen2: Exploring the boundaries of protein language models*. arXiv:2206.13517. https://arxiv.org/abs/2206.13517
- Yang, J., Li, F.-Z., & Arnold, F. H. (2024). Opportunities and challenges for machine learning–assisted enzyme engineering. *ACS Central Science, 10*(2), 226–241. https://doi.org/10.1021/acscentsci.3c01275
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2022). Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*. https://doi.org/10.1101/2022.07.20.500902
- https://news.engr.psu.edu/2024/molecular-roadblocks-slow-cellulose breakdown.aspx
- https://www.frontiersin.org/journals/energyresearch/articles/10.3389/fenrg.2016.00036/fulSupporting Documents
- Source Code: https://github.com/ahmedbahaj/EnzymesBreakthrough

أكاديمية كاوست
KAUST ACADEMY