# AI-Optimized Enzyme Design for Biomass Breakdown

Accelerating Biofuel Production


ABSTRACT

This work demonstrates the power of integrating generative AI with structural bioinformatics to solve critical challenges in sustainable energy.



**Authors**:

Ahmed Omar Bahaj,
University of Djeddah – CCSE

Mouhamad Alim Alamine,
Islamic University of Madinah – FCSIS

Nabil Abdulkafi Alhamwi,
Arab Open University – FCS

**Project Mentor:** Dr. Fares Fourati,
King Abdullah University of Science and Technology

Table of Contents

## 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.

## 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 AlphaFold2, and large-scale molecular docking simulations to rapidly screen for high-potential enzyme candidates.

The project successfully moved from a broad scientific challenge to the identification of two specific, high-efficacy enzyme designs—GH11_sample1 and PL7_sample2—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.

## Project Objectives and Goals

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 (GH6, GH11, CE1, 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 finetuning 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 AlphaFold quality metrics and docking affinity results, generated data visualizations, and led report compilation.

## Current Progress and 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 GH6, GH11, CE1, and PL7 enzyme families.
2. **Candidate Generation**: Generated eight novel enzyme sequences (two per target family) for analysis.

3. **Pipeline Development**: Built a fully automated, high-throughput Python pipeline for processing multiple protein structures against multiple ligands.
4. **Structure Prediction**: Predicted 3D structures for all eight enzyme candidates using ColabFold.
5. **Quality Control**: Applied a rigorous filter based on AlphaFold pLDDT scores, retaining only high-confidence structural models.
6. **High-Throughput Docking**: Conducted a many-to-many docking screen, simulating 56 unique protein-ligand interactions across seven biomass-derived substrates.
7. **Lead Candidate Identification**: Identified two standout enzyme candidates (GH11_sample1 and PL7_sample2) with exceptional binding affinities.

## Methodology and Approach

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 finetuned Progen2 model generated sequences targeting four enzyme families: GH6 (cellulase), GH11 (xylanase), CE1 (esterase), and PL7 (alginate lyase), 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 AlphaFold2 via ColabFold to predict 3D structures of the eight AI-generated sequences. Evaluated models using mean pLDDT scores, retaining only those with high confidence (typically >80) 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.

### Pipeline Architecture:

1. **Input**: Curated enzyme sequence dataset for Progen2 finetuning.

2. **Sequence Generation**: Progen2 generates candidate sequences for GH6, GH11, CE1, and PL7 families.
3. **Structure Prediction**: ColabFold predicts 3D structures, outputting pLDDT scores for quality assessment.
4. **Quality Filter**: Retain structures with pLDDT > 80 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.

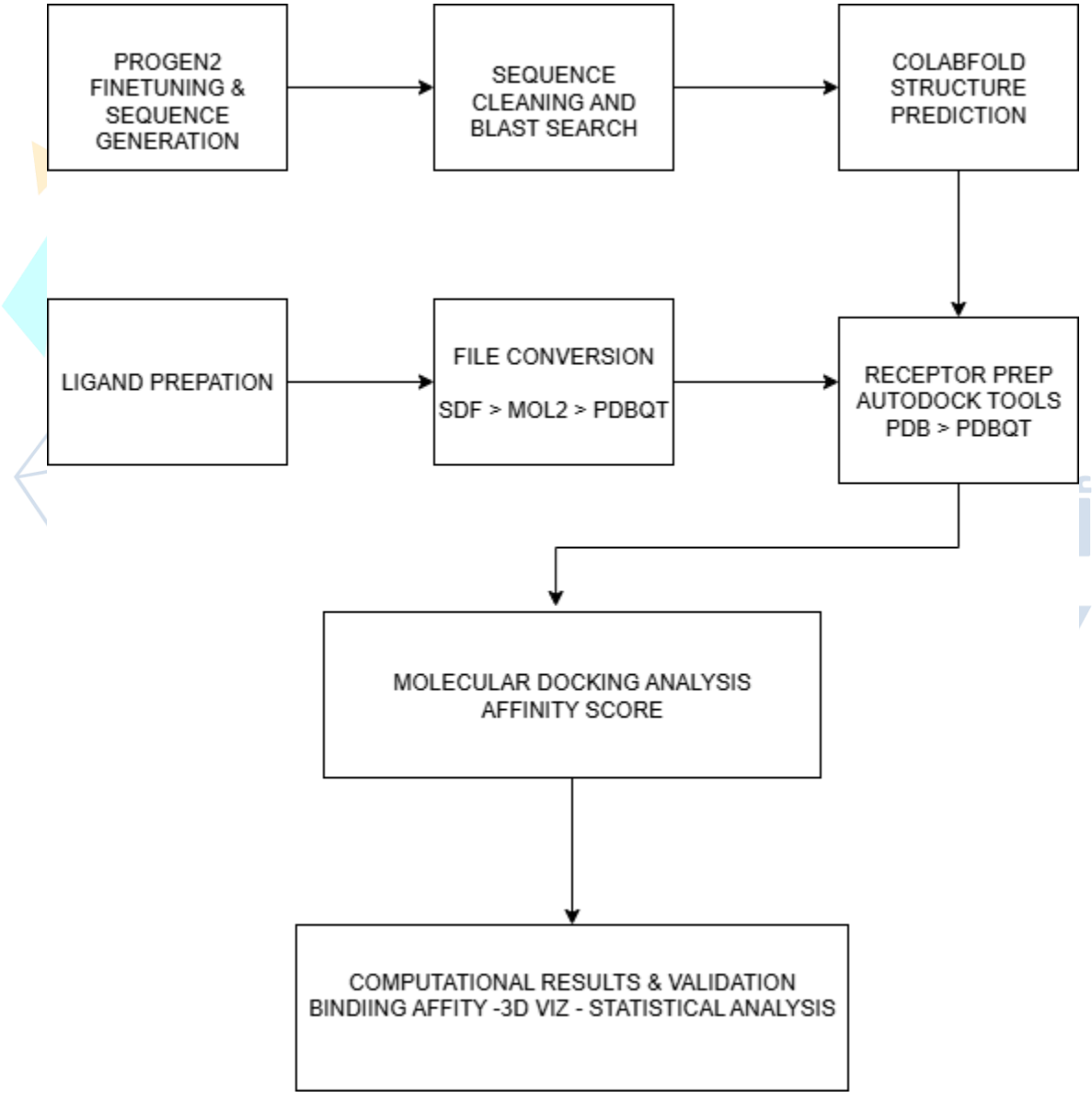


Figure 1. Project overall architecture

## Preliminary Results or Findings

The pipeline yielded robust results, identifying two high-priority enzyme candidates.

### Structural Confidence:

AlphaFold quality assessment confirmed high-confidence models for GH11_sample1 (mean pLDDT ≈ 89.8) and PL7_sample2, while low-confidence models (e.g., GH6_sample1, mean pLDDT ≈ 32.4) were excluded to ensure reliable docking results.

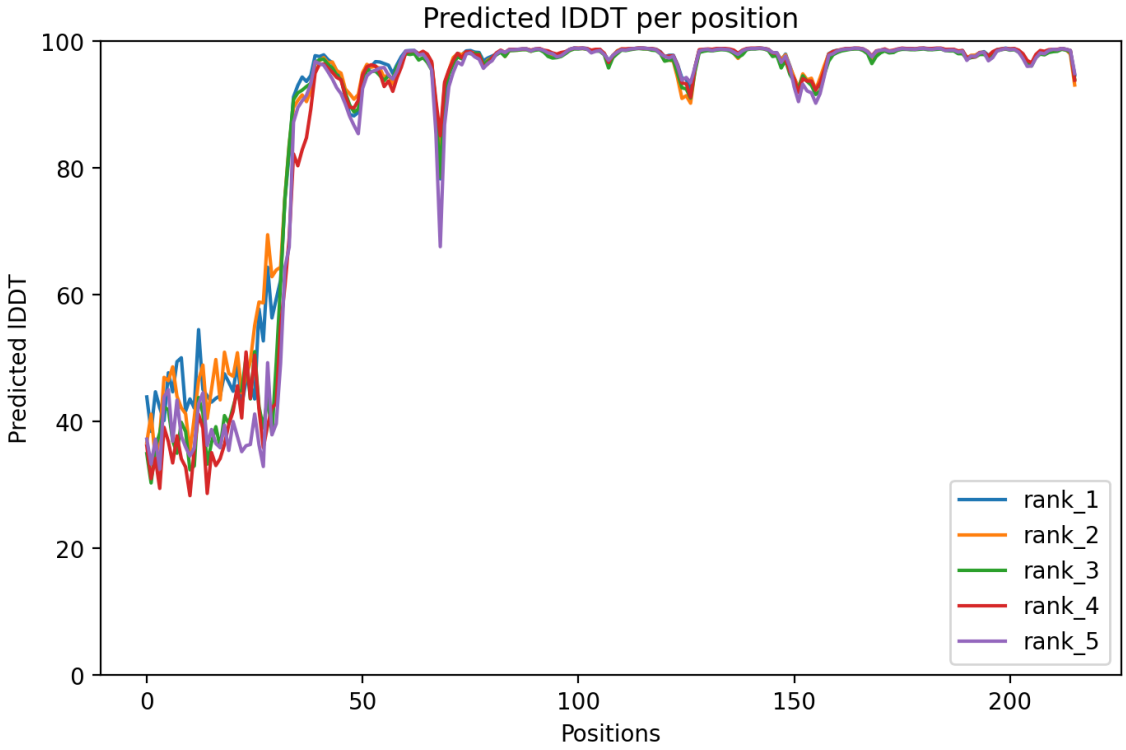


Figure 2. GH11_sample2 plddt score

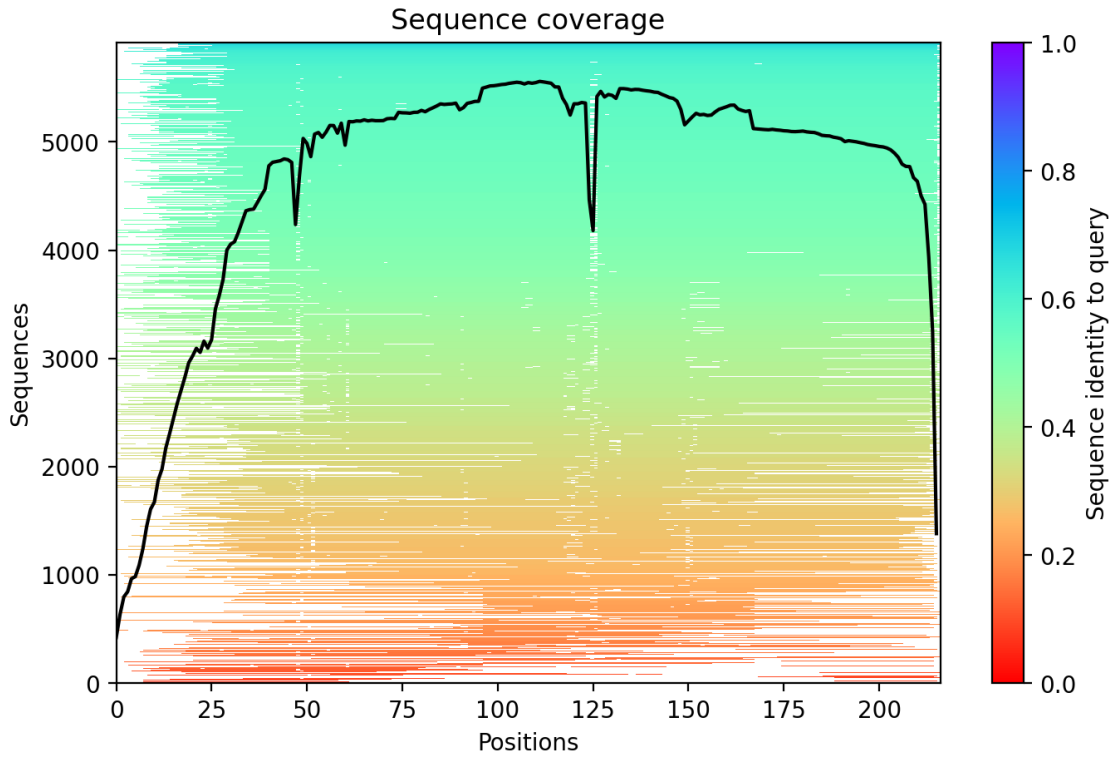


Figure 3. GH11_sample1 coverage

## Docking Analysis:

The many-to-many docking screen produced 56 protein-ligand interactions, visualized in the affinity heatmap below. Candidates with binding affinities ≤ -6.0 kcal/mol were prioritized for their strong interactions.

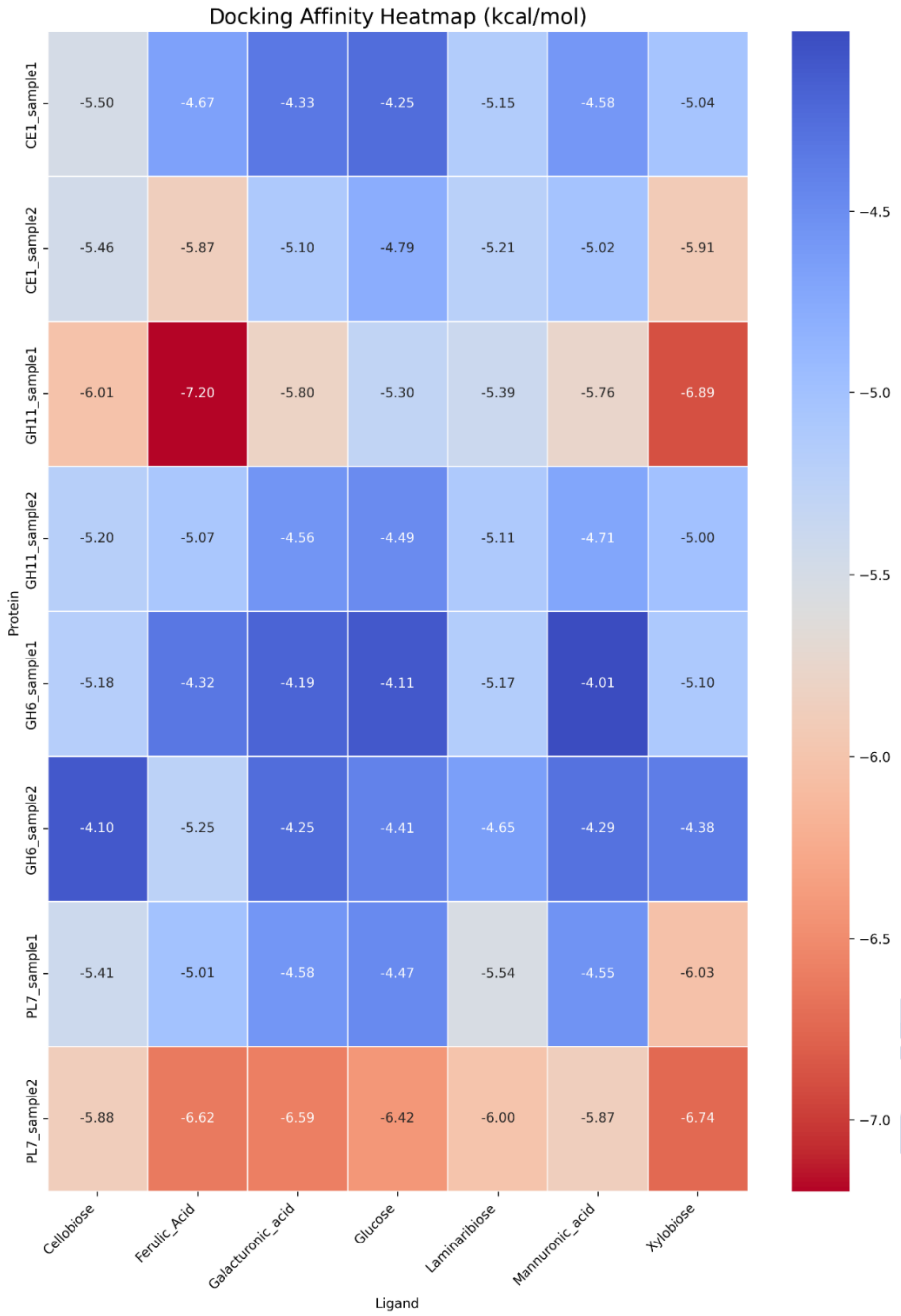


Figure 4. affinity heatmap

## Candidate 1: GH11_sample1 - The High-Affinity Specialist:

- Binding affinities: -7.20 kcal/mol (Ferulic Acid), -6.89 kcal/mol (Xylobiose).
- Demonstrates dual-action capability to cleave xylan and break ester cross-links, addressing a key rate-limiting step in biomass breakdown.

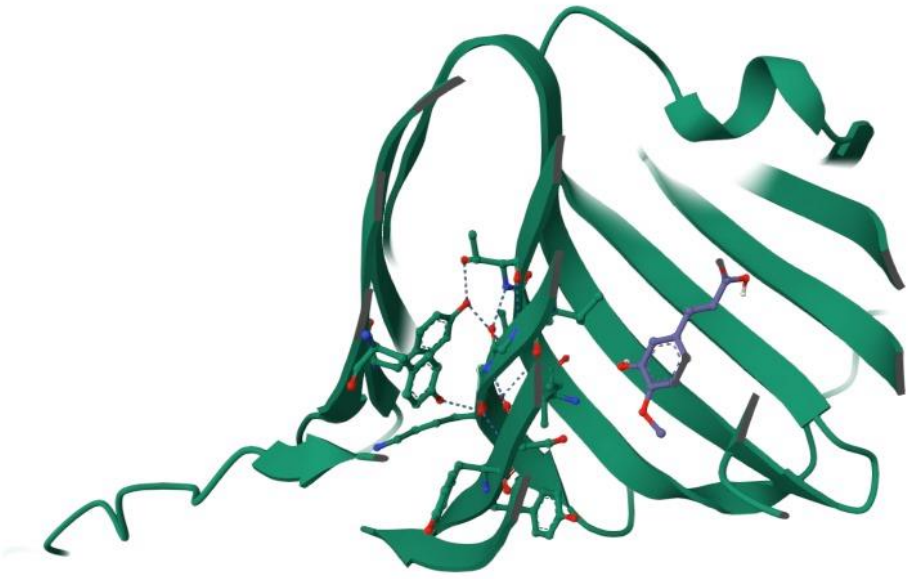

Figure 5. 3D relationship between GH11_sample1 and the docked Ferulic_Acid

Candidate 2: PL7_sample2 - The Versatile Algae Degrader:

- Binding affinities: -6.74 kcal/mol (Xylobiose), -6.62 kcal/mol (Ferulic Acid), -6.59 kcal/mol (Galacturonic acid), -6.42 kcal/mol (Glucose).
- Shows versatility across multiple substrates, ideal for broad-spectrum enzymatic cocktails.

| Ligand | Affinity (kcal/mol) |
|---|---|
| Xylobiose | -6.74 |
| Ferulic Acid | -6.62 |
| Galacturonic Acid | -6.59 |
| Glucose | -6.42 |

Figure 6. PL7_sample2: Versatile Powerhouse, Multi-substrate enzyme ideal for broad-spectrum applications

## Timeline and Schedule

- **Weeks 1-2**: Dataset curation and Progen2 finetuning.
- **Week 3**: Candidate sequence generation and pipeline development.
- **Week 4**: Structure prediction and quality control.
- **week 5**: Molecular docking and data analysis.
- **Week 6 - 7**: Report compilation and candidate prioritization.

## Challenges and Obstacles Encountered

### AI Generation Reliability:

Initial AI-generated sequences showed variable structural integrity, requiring a strict pLDDT-based filter to ensure reliable candidates.

### Computational Resource Management:

ColabFold server queue delays necessitated careful scheduling for structure prediction.

### Pipeline Robustness:

Early pipeline versions faced issues with file paths, API inconsistencies, and batch processing, resolved through iterative development.



## Next Steps and Future Work

### Gene Synthesis and Protein Expression:

Synthesize genes for GH11_sample1 and PL7_sample2, expressing them in a suitable host (e.g., *E. coli* or *Pichia pastoris*).

### Enzymatic Assays:

Measure catalytic activity and efficiency on target substrates (e.g., ferulic acid-linked xylan for GH11_sample1, mixed algal polysaccharides for PL7_sample2).

Synergistic Testing: Evaluate enzyme combinations for synergistic effects to enhance overall biomass breakdown.

Structural Validation: Pursue protein crystallization to validate AlphaFold predictions and gain mechanistic insights.

## Resources and Tools Used

- **Cloud Computing**: Google Colaboratory
- **AI Sequence Generation**: Finetuned Progen2 Model
- **Structure Prediction**: ColabFold (AlphaFold2 & MMseqs2)
- **Molecular Docking**: AutoDock Vina
- **Cheminformatics**: RDKit, Open Babel
- **Data Analysis & Scripting**: Python 3, Pandas, Biopython, Seaborn, Matplotlib

## References:


1. Hugo et al. (202400. Protein Family Sequence Generation through ProGen2 Fine-Tuning, Arxiv 2 Jumper, J., et al. (2021).
2. Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589. 3 Mirdita, M., et al. (2022).
3. ColabFold: making protein structure prediction accessible to all. Nature Methods, 19(6), 679-682. 4 Trott, O., & Olson, A. J. (2010).
4. 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.
5. https://news.engr.psu.edu/2024/molecular-roadblocks-slow-cellulose breakdown.aspx
6. https://www.frontiersin.org/journals/energyresearch/articles/10.3389/fenrg.2016.00036/fulSupporting Documents
7. Gemini AI – Google, Claude AI - Anthropic, and ChatGPT – OpenAI, Copilot – Microsoft