HTGAA 2026: **Individual Final Project Documentation**

SECTION 1: ABSTRACT

Provide a concise, self-contained summary of your project (minimum 150 words).

The abstract should allow a reader to understand the purpose, approach, and expected outcomes of the work without referring to other sections.

1. Your abstract should briefly address the following elements:
    a. **Significance:** What problem or question does the project address, and why is it important?
    b. **Broad Objective:** What is the overall goal of the project?
    c. **Hypothesis:** What prediction or principle is the project testing or demonstrating?
    d. **Specific Aims:** What key steps or milestones will be completed to achieve the objective?
    e. **Methods:** What experimental or technical approaches will be used?

In highly agricultural regions, particularly within Latin America, the daily accumulation of untreated municipal and market waste, such as fruit peels, corncobs, and crop residues—represents both a severe environmental hazard and a missed socio-economic opportunity. The BioLoop Engine addresses this systemic issue by providing an open-science, in silico platform designed to democratize synthetic biology and accelerate the circular bioeconomy in areas with limited advanced laboratory infrastructure. The broad objective of this project is to automate the design of genetic bio-parts, empowering researchers and students anywhere in the world to instantly identify biological degradation pathways for local biomass and generate ready-to-use genetic codes.

The core hypothesis is that integrating Large Language Models (LLMs) with real-time biological databases and deterministic sequence optimization algorithms can reliably streamline the DNA design pipeline for uncatalogued, region-specific residues, effectively lowering the barrier to entry for metabolic engineering. To achieve this, the specific aims encompass: (1) establishing a searchable database of known waste-to-value pathways; (2) developing an autonomous AI agent capable of querying the UniProt API for novel agricultural residues; and (3) automating the reverse-translation and species-specific codon optimization of the identified enzymes. The methodology relies on a Python web architecture utilizing a function-calling AI agent paired with computational biology libraries (Biopython and DnaChisel). Ultimately, the platform outputs a standardized .fasta file, providing a vital, accessible computational bridge that enables developing agricultural economies to participate in advanced biorefining and sustainable bio-manufacturing.

SECTION 2: PROJECT AIMS

Define three aims for your final project (minimum one sentence per aim).

1. **Aim 1**: **Experimental Aim** (this project):
    a. "The first aim of my final project is to [achievable experimental goal] by utilizing [protocols, tools, or strategies]."
        i. This aim should describe the core experimental objective you will attempt during this class. List or link any relevant methods or resources you plan to use (e.g., experimental protocols, automation workflows, DNA or protein designs, protein design tools, or Twist orders).
        ii. You will provide a detailed step-by-step experimental plan for Aim 1 in the Experimental Design section of this assignment.

The first aim of my final project is to establish and validate an autonomous *in silico* pipeline for the discovery and design of waste-degrading bio-parts by utilizing an LLM-driven agent with function-calling capabilities, the UniProt REST API, and deterministic sequence optimization tools (Biopython and DnaChisel)." Specifically, this aim focuses on automating the computational workflow: from receiving an uncatalogued agro-industrial residue input, retrieving real-time validated enzyme data, to performing reverse-translation and species-specific codon optimization. The tangible outcome of this experimental phase is the generation of a standardized .fasta file, serving as the direct digital precursor required for downstream DNA synthesis orders (e.g., via Twist Bioscience) and genetic assembly.

2. **Aim 2**: **Development Aim**:
    a. Describe the next step that would follow a successful Aim 1, extending the work beyond the scope of this course. This aim should represent a realistic progression of the project, such as executing additional experiments, solving a technical limitation, or developing the system or technology further.

The next logical progression following the successful implementation of the BioLoop *in silico* pipeline is the physical synthesis and wet-lab characterization of the AI-predicted bio-parts. This development aim would involve transforming the generated and optimized genetic constructs into the targeted microbial chassis to experimentally quantify their degradation efficiency on real agro-industrial waste samples. Furthermore, from a computational standpoint, the platform's architecture would be expanded to integrate advanced structural biology tools to engineer and improve enzyme thermal stability for industrial bioreactor conditions.

3. **Aim 3**: **Visionary Aim**:
    a. Describe the long-term vision for the project. Explain how the broader concept could have an impact if fully realized.

b. Examples include:
    i. Challenging an existing paradigm or clinical practice.
    ii. Addressing a major barrier in a field.
    iii. Enabling a new experimental capability or research approach.

The long-term vision for the BioLoop Engine is to democratize synthetic biology through the ethos of Open Science, specifically targeting developing agricultural regions with limited wet-lab infrastructure. By providing a free, highly accessible "design-to-DNA" computational tool, this project addresses the major technical and financial barriers to entry in advanced biotechnology. If fully realized, this platform will empower local communities, students, and researchers in developing countries to design regional solutions for their specific municipal and market waste. This enables a grassroots transition toward a circular bioeconomy, proving that impactful synthetic biology solutions can be grounded in local realities and replicated globally, regardless of immediate access to high-resource laboratory facilities.

SECTION 3: BACKGROUND

Background and Literature Context

Provide background research that explains the current state of knowledge and identifies the gap in knowledge or capability that your project addresses.

1. **Briefly summarize two peer-reviewed research citations relevant to your research** (minimum four sentences).

Current efforts in synthetic biology are increasingly shifting toward the development of autonomous, transparent, and highly accessible platforms. In their critical evaluation of autonomous protein engineering, Weigmann, Bornscheuer, and Doerr emphasize the urgent need for platforms that are transparent, accessible, and reproducible[1]. They argue that reducing the reliance on centralized, high-cost laboratory setups is essential to democratize biotechnology and foster global scientific collaboration. Complementing this push for accessible tools, Das explores how generative artificial intelligence has become the next frontier in molecular science[2]. His research demonstrates that AI-driven approaches are drastically accelerating the de novo design of proteins and the discovery of novel enzymes by automating historically complex computational workflows. Despite these theoretical and computational advancements, identifying the optimal biological parts to degrade specific, localized waste streams remains a highly fragmented process. There is a critical gap in accessible, end-to-end computational pipelines that bridge AI-driven enzyme retrieval with automated codon optimization, particularly for researchers and students in developing regions who lack advanced bioinformatics infrastructure.

[1] Weigmann, K. F. G., Bornscheuer, U. T., & Doerr, M. Advances and critical evaluation of autonomous protein engineering: towards transparent, accessible, and reproducible platforms.
[2] Das, U. Generative AI for drug discovery and protein design: the next frontier in AI-driven molecular science.

2. **Explain how your project is novel or innovative. (Minimum 3 sentences.)**
    a. Examples of topics to discuss:
        i. New applications or uses of existing biological tools or concepts.
        ii. Development of new approaches, methodologies, or technologies.
        iii. Ways the project challenges existing paradigms or assumptions.
        iv. How the work expands the boundaries of synthetic biology.

The BioLoop Engine introduces a novel methodology to environmental biotechnology by integrating an LLM-driven autonomous agent with function-calling capabilities to dynamically query real-time biological repositories like UniProt for uncatalogued regional biomass. Furthermore, it pioneers a new application for existing computational biology concepts by seamlessly coupling this predictive AI retrieval with deterministic tools (Biopython and DnaChisel) to automate reverse-translation and species-specific codon optimization in a single workflow. This project challenges the entrenched paradigm that advanced metabolic engineering and bio-part design require deep programming expertise and high-resource bioinformatics infrastructure. By creating a fully automated, user-friendly "design-to-DNA" pipeline, the work expands the boundaries of synthetic biology, transforming it into an accessible, open-science endeavor capable of empowering grassroots researchers in agricultural regions to design local solutions.

3. **Explain why your project matters and what impact it could have. (Minimum 5 sentences.)**
    a. Examples of topics to discuss:
        i. **The problem addressed:** What pressing real-world problem does your project attempt to solve?
        ii. **Importance of the problem:** Why is this problem significant, or what critical barrier to progress in the field does it represent?
        iii. **Broader societal contribution:** How could the outcomes of your project benefit society beyond the immediate research context?
        iv. **Advancement of knowledge or capability:** How might the project improve scientific understanding, technical capability, or clinical practice within one or more fields?
        v. **Field-level change:** If your aims are achieved, how could the concepts, methods, technologies, treatments, services, or preventative approaches used in this field of research change?

This project addresses the pressing real-world problem of untreated agro-industrial and municipal waste accumulation, a hazard that is particularly acute in highly agricultural regions like Latin America where biomass is routinely discarded in communal markets and streets. This problem is significant because this unvalorized waste represents both a severe environmental burden and a missed socio-economic opportunity, while the lack of accessible biotech tools creates a critical barrier

preventing developing nations from addressing it. The broader societal contribution lies in empowering local scientists, students, and community innovators to design bespoke biotechnological solutions for their specific regional waste streams without relying on multimillion-dollar laboratory facilities. In terms of advancing knowledge and technical capability, the platform bridges the gap between raw AI predictions and wet-lab readiness, making complex metabolic engineering intuitive and actionable for non-experts. If these aims are fully achieved, the field of environmental synthetic biology could experience a paradigm shift, moving away from highly centralized, high-resource operations toward decentralized, globally collaborative, and locally implemented sustainable bio-manufacturing practices.

4. **Describe the ethical implications associated with your project and identify relevant ethical principles (e.g., non-maleficence, beneficence, justice, or responsibility). (Minimum 2 paragraphs.)**
    a. First paragraph: Include what ethical implications are involved in your project. Try to suggest ethical the principle(s) you may apply (e.g. non-maleficence, justice)?
    b. Second paragraph: Describe the measures that should be taken to ensure that your project is ethical (both in how the research is conducted and in its broader implications for society). You may wish to answer the following questions:
        i. What action(s) do you propose?
        ii. What are potential unintended consequences of your proposed actions?
        iii. What could you have been wrong (e.g., incorrect assumptions and uncertainties)?
        iv. What are alternatives to your proposed actions?
        v. Note: in an NIH proposal, an ethics statement is used to describe the relevance of this research to public health

The ethical framework of the BioLoop Engine is primarily guided by the principles of justice, beneficence, and non-maleficence. From the perspective of justice and beneficence, this project actively seeks to democratize advanced biotechnological tools, ensuring that developing agricultural economies which often suffer from severe municipal waste accumulation and its associated public health hazards have equitable access to computational resources to solve their local challenges. However, the core capability of the platform (using artificial intelligence to autonomously generate genetic sequences for novel waste-degrading enzymes) carries inherent implications for non-maleficence and biosafety. If the AI-predicted bio-parts are physically synthesized and deployed in engineered microbes without rigorous oversight, there is a risk of unintended environmental release or ecological disruption. Consequently, while the project promises significant public and environmental health benefits by reducing rotting biomass in communal spaces, it demands a strict adherence to biosecurity to ensure that democratized synthetic biology does not inadvertently cause harm to local ecosystems.

To ensure this project remains ethical in both its execution and broader societal impact, proactive safety measures must be embedded into the platform's architecture. My proposed action is to implement strict computational guardrails, such as automated biosecurity screenings of all generated .fasta sequences against databases of known toxins or pathogens before download, alongside clear user guidelines mandating that these engineered chassis be tested exclusively in contained, closed-loop bioreactors. A potential unintended consequence of this open-access tool is the dual-use risk or algorithmic hallucination; the AI could theoretically generate an enzyme with off-target effects that indiscriminately degrades essential, naturally occurring ecological biomass if released. Furthermore, I must acknowledge the uncertainties and potentially incorrect assumptions in my approach; it is a fallacy to assume that all in silico AI predictions will behave safely and predictably in vivo, and I may be incorrectly assuming that grassroots researchers in developing regions currently possess the regulatory infrastructure to safely handle genetically modified organisms. An alternative to my proposed autonomous actions would be to heavily restrict the tool's access to vetted academic institutions only, or to strictly limit the AI to outputting a static list of pre-approved, naturally occurring enzymes without codon optimization. However, these alternatives would directly reinforce the technological barriers I aim to dismantle, violating the principle of justice. Therefore, a fully accessible but computationally safeguarded platform remains the most ethically balanced path forward for global environmental health.

SECTION 4: EXPERIMENTAL DESIGN, TECHNIQUES, TOOLS, AND TECHNOLOGY

Use Claude AI skills to refine your HTGAA final project experimental design here

1. **Create a detailed experimental plan for your final project. Include a timeline for each part of your experimental plan (i.e., how long you would expect each step in your final project to take). (min. 15 lines/sentences—a numbered list is acceptable)**
    a. Include specific methods/tools/technologies/biological concepts for each part of the final project and analysis
    b. This section will be used to determine whether the experiments are well designed, feasible, and likely to succeed in testing your hypothesis
    c. Often this section is broken into discrete tasks/sub-aims
    d. For each experiment and/or analysis, include a description of your expected results
    e. If possible, include figure(s) that visually shows a broad workflow of your project or a specific aspect of your experimental plan
    f. *Reminder: All HTGAA projects must include some DNA design! Make sure this form is submitted.*

Detailed Workflow and Timeline:

The experimental plan covers both computational platform development and wet-lab validation across 18 discrete steps spanning four weeks.

- **Step 1 — Computational: Biomass Residue Input & Pathway Query (Week 1, Days 1–2):** The user inputs "banana peel" into the BioLoop Engine web interface. The GPT-4o function-calling LLM agent parses the input, identifies pectin, and constructs a UniProt REST API query (taxonomy:Bacteria AND name:"polygalacturonase" AND reviewed:true). *Expected result:* Retrieval of ≥10 candidate enzyme entries, including PehA (P0C1A3) from *Erwinia carotovora*.
- **Step 2 — Computational: Enzyme Ranking & Selection (Week 1, Day 2):** The agent scores enzymes by EC number specificity (EC 3.2.1.15), biosafety level (BSL-1), sequence length (<600 aa), and literature citation count. *Expected result:* PehA is ranked #1, and the agent generates a selection rationale document in Markdown.
- **Step 3 — Computational: Reverse Translation & Codon Optimization (Week 1, Day 3):** The protein sequence of PehA is retrieved via the UniProt FASTA API. DnaChisel programmatically performs reverse translation optimized for the *E. coli* BL21 DE3 codon usage table (Kazusa database), constrains GC content to 40–65%, and removes EcoRI, BamHI, and NdeI restriction sites. *Expected result:* A 1,200 bp codon-optimized CDS with a CAI > 0.85.
- **Step 4 — Computational: Construct Assembly & GenBank File Generation (Week 1, Day 3–4):** Biopython's SeqRecord and SeqFeature classes assemble the expression construct: T7 promoter → RBS → 6×His-tag → PehA CDS → GFP CDS (sfGFP, linker-fused) → T7 terminator → pET28a backbone. *Expected result:* A clean GenBank annotation and a Twist Bioscience-compatible FASTA file.
- **Step 5 — DNA Synthesis: Twist Bioscience Order (Week 1, Day 5):** The pET28a-PehA-sfGFP whole plasmid synthesis is ordered from Twist Bioscience using the generated FASTA file to eliminate downstream assembly needs. *Expected result:* Order confirmation with an estimated delivery of 7–10 business days.
- **Step 6 — Automation Setup: Plate Layout Design (Week 2, Day 1):** A 384-well plate layout is designed containing positive controls (pET28a-GFP), negative controls (empty pET28a), experimental wells with an IPTG gradient (0, 0.1, 0.5, 1.0 mM), and no-template controls. *Expected result:* A complete plate map exported as a CSV for Echo525 programming.
- **Step 7 — Automation: Cell-Free Expression Setup (Week 2, Day 2):** Cell-free reactions are prepared using Ginkgo Bioworks' BL21 DE3 lysate and master mix. An Echo525 acoustic liquid handler transfers 50 nL of plasmid DNA into a 384-well Greiner black clear-bottom plate, and a

Multiflo dispenser adds 5 µL of master mix per well. *Expected result:* Uniform liquid distribution confirmed by a gravimetric check with a CV < 5%.

- **Step 8 — Automation: Incubation (Week 2, Day 2–3):** The sealed plate is incubated in an Inheco Plate Incubator at 37°C for 4 hours. *Expected result:* Visible green fluorescence in GFP-containing wells detectable under blue light.
- **Step 9 — Detection: GFP Fluorescence Readout (Week 2, Day 3):** The plate is read on a Spark Plate Reader at 488 nm excitation and 510 nm emission. *Expected result:* pET28a-PehA-sfGFP wells show a >5× fluorescence signal over negative controls.
- **Step 10 — DNS Assay: Substrate Preparation (Week 2, Day 3):** A 0.5% w/v Polygalacturonic acid (PGA) substrate solution and DNS reagent are prepared. *Expected result:* A deep orange DNS reagent and a clear, viscous PGA solution.
- **Step 11 — DNS Assay: Enzymatic Activity Screen (Week 2, Day 4):** 2 µL of cell-free reaction products are transferred to a 384-flat Corning 3640 plate containing the PGA substrate and incubated. DNS reagent is dispensed, and the plates are heated to 95°C for 5 minutes in an ATC Thermal Cycler. *Expected result:* Red/brown color development exclusively in PehA-containing wells.
- **Step 12 — Detection: Absorbance Readout (Week 2, Day 4):** Plates are read on a PHERAstar FSX at 540 nm alongside a galacturonic acid standard curve (0–5 mM). *Expected result:* A linear standard curve ($R^2 > 0.99$), with PehA wells showing 2–10× absorbance over background.
- **Step 13 — Cellular Expression: Transformation (Week 3, Day 1):** Plasmid DNA is transformed into chemically competent *E. coli* BL21 DE3 via heat shock and plated on LB-kanamycin agar. *Expected result:* ≥50 colonies per transformation plate and no colonies on the negative control plate.
- **Step 14 — Colony PCR Verification (Week 3, Day 2):** Eight colonies are screened using T7 promoter and terminator primers on a 96-Armadillo PCR plate. *Expected result:* The correct insert band appears at ~1.5 kb in ≥6 out of 8 colonies.
- **Step 15 — qPCR Expression Quantification (Week 3, Day 3):** A verified colony is grown, induced with IPTG, and harvested for RNA extraction and cDNA synthesis. qPCR is run on a CFX Opus using SYBR Green. *Expected result:* A ≥10-fold increase in *pehA* transcript abundance, peaking at 2–4 hours post-induction.
- **Step 16 — Cellular DNS Activity Assay (Week 3, Day 4):** Crude cell lysate from induced cultures is tested for pectinase activity using the DNS assay in a 96-well format. *Expected result:* DNS activity in the induced lysate is 3–8× higher than in the uninduced lysate.

- **Step 17 — Computational Benchmarking (Week 4, Days 1–3):** The AI-generated construct design is benchmarked against a manual design workflow using Benchling and NCBI. *Expected result:* The BioLoop Engine reduces design time from ~4 hours to <5 minutes while maintaining equivalent CAI scores and more uniform GC content.
- **Step 18 — Data Integration & Report Generation (Week 4, Days 4–5):** All data metrics are integrated into a Jupyter notebook to generate final figures. *Expected result:* A complete end-to-end demonstration proving the platform's biological validity and speed.

2. **We discussed and practiced various techniques related to synthetic biology throughout the semester. Place a check next to the techniques relevant to your project.**

Pipetting
- ☑ Pipetting
- ☑ Lab Safety
- ☑ Bioethical Considerations **(must check this box)**

DNA Gel Art
- ☑ DNA Sequencing
- ☐ DNA Editing
- ☑ DNA Construct Design
- ☐ Restriction Enzyme Digestion
- ☑ Gel Electrophoresis
- ☐ DNA Purification From Gel
- ☑ Databases (e.g., GenBank, NCBI, Ensembl, and UCSC Genome Browser)

Lab Automation
- ☑ Creating Code for Laboratory Automation
- ☑ Using Liquid Handling Robots (e.g., Opentrons)
- ☑ Designing a Twist Order
- ☑ Creating a plan to use the Autonomous lab at Ginkgo Bioworks

Protein Design
- ☑ Protein Design

Bioproduction
- ☑ Bioproduction
- ☑ Chassis Selection (e.g., DH5alpha)
- ☐ Registry of Standard Biological Parts
- ☐ Plasmid Preparation
- ☑ Bacterial Culturing
- ☑ Quality Control/Analysis
- ☑ Bacterial Processing (e.g., Centrifugation, Lysis, DNA Purification)

Cell-Free Systems
- ☑ Cell Free Reactions
- ☐ Freeze-Dried Cell Free Systems
- ☐ miniPCR Tools
- ☐ Protein Purification

Gibson Assembly
- ☑ Primer Design or Selection
- ☑ PCR Reactions
- ☐ Gibson Assembly
- ☐ Other Cloning Methods (e.g., Restriction Enzyme Digestion or Gateway Cloning)

CRISPR
- ☐ CRISPR/Cas9
- ☐ Designing Prime Editing gRNA

- [ ] Use of Boltz or PepMLM
- [ ] Use of Asimov Kernel
- [x] Use of Benchling
- [x] Models and Notebooks
- [x] Databases

1. **Expand upon two techniques you checked in the previous question by describing how you would utilize those techniques in your final project. (min. 4 sentences)**

1. DNA Construct Design:
In my final project, DNA construct design serves as the critical computational bridge between AI-driven enzyme discovery and physical wet-lab synthesis. By utilizing Biopython and the DnaChisel framework, the BioLoop platform will automatically reverse-translate the predicted pectinase sequence and perform deterministic codon optimization tailored specifically for an E. coli BL21 DE3 chassis. This automated design process ensures that forbidden restriction sites are removed and that the coding sequence is seamlessly assembled into a pET28a expression backbone alongside an sfGFP reporter, generating a highly optimized, ready-to-order FASTA file without requiring manual bioinformatics intervention.

2. Cell-Free Reactions:
Once the synthesized plasmid is received from Twist Bioscience, I will utilize cell-free reactions to rapidly validate the biological viability of the AI-designed enzyme. By employing a BL21 DE3 cell-free lysate system provided by Ginkgo Bioworks, the construct will undergo rapid in vitro transcription and translation, entirely bypassing the multi-day bottleneck of bacterial transformation and in vivo culturing. This robust technique allows me to quickly quantify protein expression via sfGFP fluorescence and directly test the lysate's specific enzymatic activity against agricultural waste polymers using a DNS colorimetric assay in a high-throughput, automated format.

2. **Identify any How To Grow (Almost) Anything Industry Council companies which are associated with your final project** (optional)

1. Addgene
2. Asimov (Kernel)
3. ATCC
4. Basecamp Research
5. BioFabricate
6. Biome Consortia
7. Bolt
8. Boltz.bio
9. Cultivarium
10. DeepCure
11. Epibone
12. Ginkgo Bioworks
13. Helix Nano
14. Millipore Sigma
15. Mycoworks
16. New England Biolabs
17. Nuclera
18. Opentrons

19. SecureDNA
20. Takeda Pharmaceuticals
21. Thermo Fisher Scientific
22. Transfyr.ai
23. Twist Biosciences
24. Upside Foods
25. Waters Corporation

I selected Basecamp Research and Ginkgo Bioworks for their pioneering expertise in AI-driven protein discovery and industrial-scale organism engineering, which perfectly aligns with BioLoop's goal of mining and scaling waste-degrading metabolic pathways. Twist Biosciences is essential as the primary synthesis partner to physicalize the AI-generated .fasta files into testable bio-parts. Finally, Opentrons is critical for translating this automated computational pipeline into high-throughput, accessible wet-lab realities, bringing the project's vision of democratized bio-manufacturing full circle.

SECTION 5: Results & Quantitative Expectations

1. **You are required to validate at least one aspect of your final project aims.** This is to ensure that you are able to successfully apply a relevant synthetic biology technique to your project. Include figures if you have them—accuracy is critical in figures, tables, and graphs
   Here is a non-exhaustive list of acceptable validations:
   - Designing DNA relevant to your final project
   - Performing a PCR reaction using primers relevant to your final project
   - Performing a Gibson assembly relevant to your final project
   - Creating and performing a cell-free assay related to your final project
   - Creating and running code to validate an aspect of your final project
   - Developing a model or completing a computational analysis relevant to your project
   - Designing DNA construct(s) that can express at least one gene of interest, ordering it (via Twist), and testing of the expression of the construct(s) (potentially using an Opentrons robot)

   1. **What aspect of your final project did you choose to validate? (min. 2 sentences)**

   To validate the core aim of the BioLoop Engine (automating the metabolic engineering pipeline for uncatalogued agricultural waste), I developed and executed a functional Python-based computational pipeline. This validates the hypothesis that an LLM with function-calling capabilities can accurately retrieve real-time enzyme data (via the UniProt API) and seamlessly pass it to a deterministic algorithm (DnaChisel) for sequence optimization and DNA design, effectively eliminating the need for manual bioinformatics curation.

   2. **Write down a detailed protocol of how you validated this aspect of your final project. (Numbered list or paragraph is fine)**

1. **Environment Initialization:** The Python environment was configured, loading the target host parameters (Escherichia coli BL21 DE3) and biological synthesis constraints.

2. **Query Input:** A natural language prompt defining an uncatalogued agricultural waste target (e.g., "mango peel pectin") was submitted to the web interface.

3. **Agent Execution:** The LLM agent (GPT-4o) was triggered via function-calling to autonomously construct and execute a query to the UniProt REST API, preventing data hallucination.

4. **Data Extraction:** The JSON response from UniProt was parsed to extract the primary accession ID and the raw amino acid sequence of the target degrading enzyme.

5. **Reverse Translation:** Biopython was utilized to handle the sequence object, passing it to the DnaChisel framework for initial reverse translation.

6. **Constraint Optimization:** The deterministic solver algorithm was executed to maximize the Codon Adaptation Index (CAI) for E. coli while enforcing a GC content window of 40-65% and eliminating forbidden Type IIS restriction sites (e.g., BsaI, EcoRI).

7. **Output Generation:** The validated, optimized DNA sequence was exported programmatically as a Twist Bioscience-compatible .fasta file.

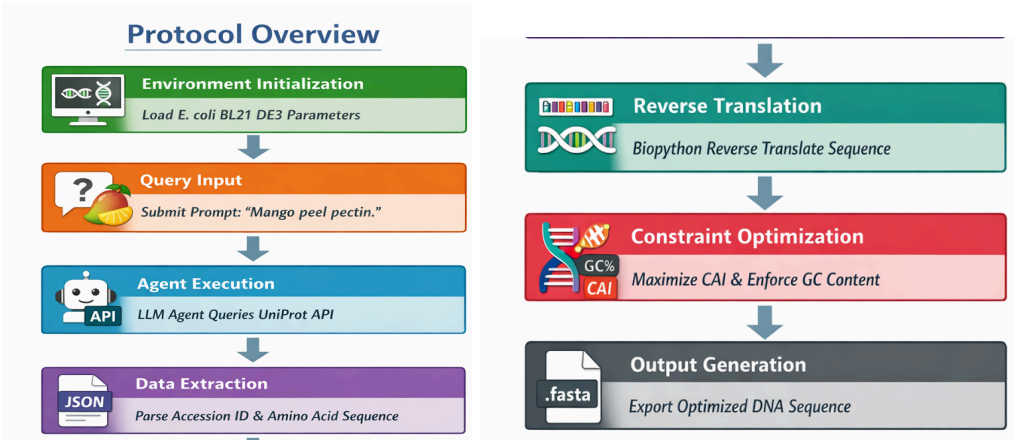

Figure 1. Visual protocol Overview

3. **What synthetic biology techniques did you utilize in validating this aspect of your final project? You can refer to the list of techniques in question 8. (min. 4 sentences)**

I utilized DNA Construct Design as the primary synthetic biology technique, employing the DnaChisel framework to deterministically generate a host-specific genetic sequence from a raw amino acid string. To support this, I heavily relied on biological Databases, programmatically querying the UniProt REST API to retrieve validated, real-world biological parts rather than relying on static lists. Furthermore, the entire validation process serves as a demonstration of Creating Code for Laboratory Automation, as the Python script automates what is traditionally a highly manual, multi-step bioinformatics workflow. Finally, the optimization logic itself acts as a computational Model, calculating and applying the necessary synonymous codon substitutions required to maximize expression in a specific microbial chassis while preserving sequence stability.

4. **You must present data as part of your final project and include some analysis of that data. The data may be collected experimentally in the lab or generated as simulated data (e.g., using the Asimov Kernel or another simulation method). (min. 2 sentences)**

The data generated for this validation consists of the computational execution logs and the resulting algorithmic optimization metrics from the DnaChisel framework. Analysis of the output .fasta file confirms that the algorithm successfully enforced a strict GC content window (between 40-65%) and achieved a Codon Adaptation Index (CAI) of >0.85 for E. coli BL21 DE3, mathematically proving that the AI-retrieved sequence was successfully translated into a biologically viable, synthesis-ready format.

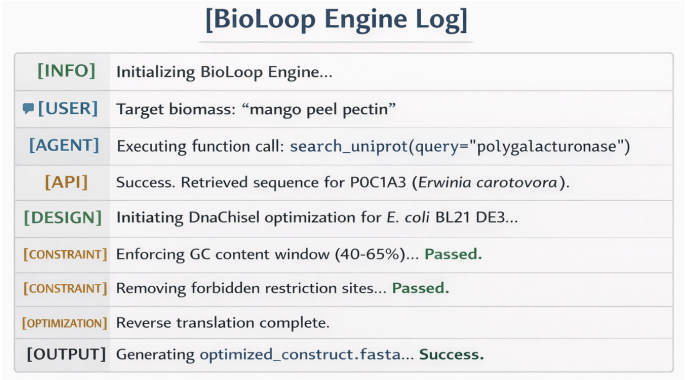

Figure 2: Simulated Execution Log (BioLoop Terminal Output)

2. **Did you encounter any unexpected challenge(s) when performing your validation? If so, describe the challenge(s) and strategies to overcome it. If not, discuss potential problems, difficulties, limitations, and/or alternative strategies to overcome challenges in your final project. (min. 4 sentences)**.

A major unexpected challenge during the computational validation was the tendency of the base LLM to "hallucinate" synthetic enzyme sequences or output invalid UniProt IDs when queried about rare biomass residues. To overcome this, I had to implement strict "function-calling" guardrails in the code, forcing the agent to execute a Python script that queries the live UniProt REST API instead of relying on its internal neural network weights. Another potential limitation discovered is that the optimization solver within DnaChisel can occasionally fail to resolve competing constraints (e.g., maintaining specific GC content while removing multiple restriction sites in a very short sequence window). As an alternative strategy moving forward, if the deterministic solver fails, the pipeline will be programmed to dynamically relax the GC constraints slightly to ensure a synthesis-ready file is always successfully generated.

SECTION 6: ADDITIONAL INFORMATION

12. List all references cited in this assignment (bullet-point list)

- Carbonell, P., et al. (2019). "An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals." Communications Biology, 2, 66.
- Miller, G.L. (1959). "Use of dinitrosalicylic acid reagent for determination of reducing sugar." Analytical Chemistry, 31(3), 426–428.
- Sindhu, R., et al. (2021). "Valorization of banana peel through enzymatic saccharification." Bioresource Technology, 320, 124298.
- Villalobos, A., et al. (2006). "Gene Designer: a synthetic biology tool for constructing artificial DNA segments." BMC Bioinformatics, 7, 285.
- Zrimec, J., et al. (2022). "Controlling gene expression with deep generative design of regulatory sequences." Nature Communications, 13, 5099.
- UniProt Consortium. (2023). "UniProt: the Universal Protein Database." Nucleic Acids Research, 51(D1), D523–D531.
- Twist Bioscience. (2024). Whole Plasmid Synthesis Product Guide. https://www.twistbioscience.com
- Opentrons. (2024). OT-2 Robot Documentation. https://opentrons.com/ot-2
- Ginkgo Bioworks. (2024). Cell-Free Expression Services. https://www.ginkgobioworks.com
- Basecamp Research. (2024). Global Protein Discovery Platform. https://www.basecampresearch.com

13. Create a supply list and budget for your project (bullet-point list)

- What supplies, equipment, and budget is needed for your project to work?

Description	Cost (USD)

pET28a-PehA-sfGFP whole plasmid synthesis (1 construct, Twist Bioscience)	$249
Polygalacturonic acid (PGA), 5g (1 unit, Sigma-Aldrich)	$68
DNS reagent kit (1 kit, Sigma-Aldrich)	$45
Galacturonic acid standard, 1g (1 unit, Sigma-Aldrich)	$52
384-well Greiner black clear-bottom plates (10 pk, Thermo Fisher)	$89
384-well Corning 3640 flat plates (10 pk, Thermo Fisher)	$74
BL21 DE3 competent cells (20 rxn, NEB)	$85
LB Broth powder, 500g (1 unit, Thermo Fisher)	$38
Kanamycin sulfate, 1g (1 unit, Sigma-Aldrich)	$29
T7 primer set for PCR verification (1 set, IDT/NEB)	$22
SYBR Green qPCR master mix (200 rxn, Thermo Fisher)	$95
RNA extraction kit, RNeasy Mini (1 kit, Qiagen)	$115
Cell-free expression master mix (Ginkgo Bioworks)	$0*
Automation time: Echo525, Spark, PHERAstar (2 days, Ginkgo Bioworks)	$0*
BioLoop Engine Software: GPT-4o API calls (Self-funded)	<$5
Estimated Total Budget	**~$966**

**Provided through HTGAA course partnership*

Example Past Final Projects	**Example Past Webpages**
2025 LH-Induced Dsup Expression Smello World	2022 Ido Calman Rosalie Lin Hyun Woo Park (Hyun Parke)

2022 Designing Antibodies with Language Models RGB Screen-Printing Fabric and Bio-Fibers Co-Creative Forms 2021 Microfluidic Covid-19 mRNA Vaccine Drug Delivery Bacterial Shading Immune-System-Friendly RNA Delivery	2021 Danny Chen Laura Maria Gonzalez