ENS210 - COMPUTATIONAL BIOLOGY

Project Documentation

Three Project Tracks Available

Option 1 — Rare Genetic Diseases
Option 2 — Common Diseases, Pharmacogenomics & Polygenic Risk Scores
Option 3 — Protein Design Competition (Anthropic-Sponsored)

Project Overview

Students may select one of three project tracks this semester. Each track targets a different domain of computational biology and varies in methodology, data sources, and expected deliverables. Read all three options carefully before choosing.

	Option 1	Option 2	Option 3
Domain	Rare Genetic Diseases	Common Diseases & PGx	Protein Design
Core Task	Predict variant pathogenicity	Build polygenic risk models	Engineer novel proteins via AI
Primary Data	ClinVar, gnomAD, HGMD	GWAS Catalog, UK Biobank	PDB, UniProt, AlphaFold
Key Methods	ML classifiers, feature engineering	PRS, pharmacogenomics	Diffusion models, docking
Unique Output	Variant pathogenicity predictor	PRS pipeline + drug response	Competition submission

OPTION 1 — Rare Genetic Diseases

1.1 Objective

Develop the best computational tool for predicting the functional effects of genetic mutations in a specific, assigned monogenic disorder. This project integrates genetics, bioinformatics, and data science to address variant interpretation in personalized medicine.

1.2 Description

Each student is assigned a monogenic disorder and its associated gene(s). The task is to design a workflow or tool that predicts the pathogenicity of novel variants in that gene, covering the following steps:

- **Literature Review:**

Comprehensive review of the assigned genetic disorder — clinical features, inheritance patterns, gene function, and existing computational prediction methods.

- **Data Curation:**

Collect known pathogenic and benign variants from established databases (ClinVar, HGMD, gnomAD). Clean and label the data.

- **Feature Engineering:**

Calculate predictive features: sequence conservation scores, biochemical properties of amino acid changes, structural information, and other annotations.

- **Tool Development/Selection:**

Implement a predictive model (rule-based, statistical, or ML classifier) or systematically evaluate existing tools (SIFT, PolyPhen-2, CADD, REVEL) on your disease-specific dataset.

- **Evaluation:**

Assess performance using accuracy, sensitivity, specificity, and Matthew's Correlation Coefficient (MCC).

1.3 Timeline & Milestones

Milestone	Description	Due
Milestone 1	Disease Selection & Project Proposal — submit a one-page proposal outlining the gene, data sources, and chosen approach.	*Week 5*
Milestone 2	Deep Dive & Data Collection — literature review, variant data download from ClinVar/HGMD/gnomAD, cleaning and labeling.	*Week 7*
Milestone 3	Feature Engineering & Implementation — calculate features, begin coding your model, run variants through existing tools.	*Week 9*
Milestone 4	Evaluation & Analysis — cross-validation, performance metrics, ROC curves, confusion matrices, critical analysis.	*Week 11*

Milestone	Description	Due
Milestone 5	Final Report & Presentation — formal scientific report and 3-minute presentation.	*Week 13*

1.4 Deliverables

- Project Proposal (Milestone 1): 1-page document — disease, gene(s), planned methodology.
- Final Report: Abstract · Introduction · Methods · Results · Discussion · References.
- Source Code: Well-commented script or GitHub repository with README.md.
- Final Presentation: 3-minute slide deck — objective, methods, key results, conclusions.

1.5 Grading Rubric

Category	Weight	Description
Code Organization	**20%**	GitHub repository structure, readability, and documentation.
Novelty in Methodology & Data Curation	**30%**	Rigor in data collection and cleaning; novelty of chosen methods.
Results & Analysis	**10%**	Correctness of evaluation; depth of interpretation; quality of figures.
Final Report & Presentation	**40%**	Clarity, organization, and scientific writing; professionalism of oral presentation.

1.6 Resources

Databases:

- ClinVar — pathogenicity reports for human variants.
- OMIM — comprehensive catalog of human genes and genetic disorders.
- HGMD — known gene lesions responsible for inherited disease.
- gnomAD — exome and genome data from a large population cohort.
- dbSNP — general catalog of short genetic variations.

Prediction Tools:

- SIFT, PolyPhen-2, CADD, REVEL, VEP (Ensembl).

Programming:

- Python: Pandas, NumPy, Scikit-learn.
- R: statistical analysis and visualization.

OPTION 2 — Common Diseases, Pharmacogenomics & Polygenic Risk Scores

2.1 Objective

Develop a computational pipeline to model complex, polygenic traits in common diseases and explore how genetic variation influences drug response. This project integrates GWAS data, polygenic risk score (PRS) methodology, and pharmacogenomics to address real-world clinical prediction challenges.

2.2 Description

Each student is assigned a common disease (e.g., Type 2 Diabetes, Coronary Artery Disease, Major Depressive Disorder) and relevant pharmacogenomic traits. The project covers:

- **Literature Review:**

Background on the disease's genetic architecture, key GWAS findings, and existing PRS methods in clinical context.

- **Data Acquisition:**

Download summary statistics from GWAS Catalog; obtain reference panels (1000 Genomes, UK Biobank proxies) for LD computation.

- **PRS Construction:**

Apply and compare PRS approaches (P+T, LDpred2, PRSice-2). Evaluate their discriminative ability using AUC and R^2.

- **Pharmacogenomics Module:**

Identify clinically relevant pharmacogenes (CYP450 family, DPYD, TPMT, etc.) for your disease. Map variant-drug interaction evidence from PharmGKB and CPIC guidelines.

- **Clinical Contextualization:**

Critically assess the clinical utility, population stratification issues, and ethical implications of your PRS model.

2.3 Timeline & Milestones

Milestone	Description	Due
Milestone 1	Disease Assignment & Proposal — one-page proposal on chosen disease, key GWAS studies, and planned PRS approach.	*Week 5*
Milestone 2	Data Collection & QC — download GWAS summary stats, apply quality control filters, document data provenance.	*Week 7*
Milestone 3	PRS Construction — implement at least two PRS methods, compute scores on target cohort, begin pharmacogenomics annotation.	*Week 9*
Milestone 4	Evaluation & Pharmacogenomics — compare PRS performance metrics, finalize drug-variant interaction analysis, create visualizations.	*Week 11*
Milestone 5	Final Report & Presentation — formal scientific report and 3-minute presentation.	*Week 13*

2.4 Deliverables

- Project Proposal (Milestone 1): 1-page document — disease, GWAS sources, PRS strategy.
- Final Report: Abstract · Introduction · Methods · Results · Discussion · References.
- Pipeline Code: Reproducible scripts or Snakemake/Nextflow workflow on GitHub with README.md.
- Final Presentation: 3-minute slide deck — objective, methods, key results, conclusions.

2.5 Grading Rubric

Category	Weight	Description
Code Organization	**20%**	GitHub repository structure; reproducible pipeline; documentation quality.
Novelty in Methodology & Data Curation	**30%**	Rigor of data QC; sophistication of PRS method selection; depth of pharmacogenomics integration.
Results & Analysis	**10%**	Accuracy of performance evaluation; depth of clinical interpretation; quality of figures.
Final Report & Presentation	**40%**	Clarity, organization, and scientific writing; professionalism of oral presentation.

2.6 Resources

GWAS & Population Data:

- GWAS Catalog (ebi.ac.uk/gwas) — curated GWAS summary statistics.
- 1000 Genomes Project — reference LD panels.
- UK Biobank — large-scale phenotype and genotype resource (proxy access).

PRS Tools:

- PRSice-2 — P+T based PRS computation.
- LDpred2 — Bayesian PRS with LD adjustment.
- PRS-CSx — cross-population PRS.

Pharmacogenomics:

- PharmGKB — variant-drug interaction evidence.
- CPIC Guidelines (cpicpgx.org) — clinical dosing recommendations.
- FDA Table of Pharmacogenomic Biomarkers.

Programming:

- Python: Pandas, NumPy, Scikit-learn, matplotlib/seaborn.
- R: ggplot2, data.table, bigsnpr (for LDpred2).

OPTION 3 — Anthropic × Adaptyv Protein Design Competition

★ **Live Competition Track** — Students selecting this option will participate in the real Anthropic × Adaptyv Bio Protein Design Competition (Sep 28 – Nov 1, 2026). Anthropic and Adaptyv are jointly sponsoring $1M+ in experimental wet-lab validation to test more than 5,000 protein designs at no cost to participants, plus $1M+ in Claude credits to selected teams.

3.1 Competition Overview

The competition runs five consecutive one-week challenges (Sep 28 – Nov 1, 2026), each focused on a different disease-relevant protein target. A new challenge opens every Monday; submissions close the following Sunday at 23:59 AoE. Selected designs are synthesized and tested by Adaptyv's automated lab, with all experimental results — including failures — published openly on Proteinbase.

Competition page: proteinbase.com/competitions/anthropic-adaptyv-2026

3.2 Challenge 1 (Active): EGFR Conditional Binder

Challenge 1 (Sep 28 – Oct 4) asks participants to design a binder against human EGFR's extracellular domain with three increasingly difficult objectives:

- **Objective 1 — Binding Affinity:**

Design a de novo binder against human EGFR (UniProt P00533-1, residues 25–645). Recommended epitope: Domain III (also targeted by cetuximab and panitumumab).

- **Objective 2 — Mouse Cross-Reactivity:**

The same design must bind both human and mouse EGFR, enabling use in preclinical mouse models without a separate surrogate binder.

- **Objective 3 — pH-Selective Binding:**

The design must bind EGFR at pH 6.5 (tumour microenvironment) but show no detectable binding at pH 7.4. This mimics tumour-selective targeting for therapeutic relevance.

Designs are ranked in this priority order: pH selectivity > mouse cross-reactivity > human affinity.

3.3 Submission Requirements

- Format: single-chain protein, nanobody, scFv, or Fab.
- Length: 10–250 amino acids.
- Designs must be de novo and zero-shot — no modification of existing binders.
- Submit as a CSV (columns: name, sequence, molecule_class) ordered by your personal ranking.
- Include as much supporting metadata as possible: pLDDT scores, ipTM/ipSAE, structure models, design method description. Claude can help write a methods summary for submission.
- Track 3 (open): up to 20 designs per challenge; self-supported. This will be your plan!

3.4 Molecule Categories

Performance is stratified across five molecule classes. Students should choose one focus:

- Protein minibinders — 40–100 amino acids.
- Large protein binders — >100 amino acids.
- Protein microbinders — <40 amino acids.
- Nanobodies — annotated by ANARCI; use standard framework regions.
- Antibodies (scFv or Fab format).

3.5 Recommended Design Workflow

- **Target Analysis:**

Study EGFR Domain III structure (PDB 6ARU). Identify key interface residues targeted by cetuximab. Review Adaptyv's EGFR competition post-analysis preprint (bioRxiv 2025.04.17).

- **Generative Design:**

Use RFdiffusion for backbone generation; ProteinMPNN for sequence design. Alternatively, use ESM-3 generative sampling. Leverage Anthropic's inference-optimized open-source models (github.com/anthropics/uplifting-biomolecular-modeling).

- **Computational Validation:**

Predict structures with ColabFold/AlphaFold2 multimer. Filter by pLDDT ≥80, PAE, and interface pTM. Run FoldX or Rosetta for stability checks.

- **pH-Selectivity Strategy:**

Incorporate histidine-switch mechanisms or pH-sensitive interface contacts. Use Claude to reason about titration of interface residues and literature precedent for pH-responsive binders.

- **Iterative Refinement:**

Use Claude to interpret failure modes, design new hypotheses, and prioritize the top candidates. Ask Claude to draft a methods paper and metadata package for submission.

- **Submission:**

Submit through proteinbase.com/competitions/anthropic-adaptyv-2026/submit. Challenges 2–5 open weekly — apply the same workflow to each new target.

3.6 Course Timeline & Milestones

Milestone	Description	Due
Milestone 1	Proposal — one-page summary of your molecule class choice, design tools, and initial EGFR strategy. Register on Proteinbase.	*Week 5*
Milestone 2	Challenge 1 Submission — submit ≥10 EGFR designs to the live competition. Document pipeline, metrics, and design rationale.	*Week 5 (Oct 4)*
Milestone 3	Challenges 2–4 — participate in at least two additional weekly challenges as they open. Maintain a running design log.	*Weeks 6–10*

Milestone	Description	Due
Milestone 4	Analysis & Refinement — collate all in silico metrics across challenges; critically assess what worked; prepare final designs for Challenge 5.	*Week 11*
Milestone 5	Final Report & Presentation — scientific report + 3-minute presentation. Include any experimental results published on Proteinbase by Dec 15.	*Week 13*

Note: The competition's experimental results are released Dec 15, 2026. If your designs are selected for wet-lab testing, you may incorporate those results into your final report.

You should participate in at least 3 challenges! Even if you miss the first challenge (EGFR), you can design your proteins for the other challenges.

3.7 Deliverables

- Project Proposal (Milestone 1): 1-page — molecule class, design tools, EGFR targeting strategy.
- Competition Submissions: CSV files submitted to Proteinbase for each challenge attempted (minimum: Challenge 1).
- Final Report: Scientific paper format — Introduction · Design Strategy · Computational Methods · Per-Challenge Results · Discussion · References. Incorporate experimental validation data if available.
- Code & Repository: GitHub repo with all scripts, input PDB/FASTA files, generated structures, filtering notebooks, and README.md.
- Final Presentation: 3-minute slide deck — competition context, design approach, in silico metrics, wet-lab results (if released).

3.8 Grading Rubric

Category	Weight	Description
Code Organization & Reproducibility	**20%**	GitHub repo quality; documented pipeline; ability to reproduce design runs.
Design Novelty & Scientific Rigor	**35%**	Originality and sophistication of the design strategy; appropriate use of AI tools; depth of computational validation; quality of pH-selectivity approach.
Competition Participation & Results	**15%**	Number of challenges attempted; quality of submitted designs; in silico metrics; experimental results if available from Proteinbase. **If at least one of your designs is selected for wet-lab testing you will get full points from novelty and results (50%)**
Final Report & Presentation	**30%**	Clarity of scientific writing; depth of self-critical analysis; professionalism of oral presentation.

3.9 Resources

Competition:

- Proteinbase competition page: proteinbase.com/competitions/anthropic-adaptyv-2026
- Challenge 1 (EGFR): proteinbase.com/competitions/anthropic-adaptyv-2026/challenges/egfr
- Submission portal: proteinbase.com/competitions/anthropic-adaptyv-2026/submit
- Proteinbase Slack community for questions and updates.

Anthropic Resources:

- Anthropic blog — protein design with Claude: anthropic.com/research/Claude-accelerates-protein-design
- Inference-optimized folding/design models: github.com/anthropics/uplifting-biomolecular-modeling
- Binder design technical report and simplified prompt (strongly recommended starting point).

AI Protein Design Tools:

- RFdiffusion — diffusion-based de novo backbone design (Baker Lab).
- ProteinMPNN — sequence design for fixed backbones.
- ESM-3 (EvolutionaryScale) — generative protein language model.
- ColabFold — fast AlphaFold2 multimer inference.
- AlphaFold3 / AF3 Server — structure prediction for protein complexes.
- ANARCI — nanobody/antibody annotation.

Structural Validation:

- Rosetta Suite — energy-based design and ddG estimation.
- FoldX — rapid stability and binding free energy calculations.

Background Reading:

- Adaptyv EGFR competition post-analysis (bioRxiv 2025.04.17.648362).
- Adaptyv blog: adaptyvbio.com/blog — past competition analyses and novelty guidance.
- Proteinbase data corpus — prior competition designs and experimental results.

Shared Policies & Expectations (All Options)

Academic Integrity

All code, analysis, and writing must be your own. You may use AI tools (including Claude) as a research and coding assistant, but must disclose their use and be able to explain every part of your submitted work. Copying code without understanding it is a violation.

GitHub Requirement

All options require a GitHub repository. Accept the assignment link provided by the instructor to create your repository. Commit regularly — a history of meaningful commits is part of the Code Organization grade.

Collaboration Policy

Projects are individual. You may discuss concepts with peers but all implementation and writing must be independent. For Option 3, small teams may be permitted — check with the instructor.

Questions & Support

Office hours (Instructor and TA) are available. For Option 3 competition-specific questions, refer to the Anthropic competition brief or contact the instructor for clarification.