Spring 2025
GENE 46100 Deep Learning in Genomics

Instructors: Ran Blekhman (blekhman@uchicago.edu), Hae Kyung (Haky) Im (haky@uchicago.edu)
Teaching Assistant: Charles Yichao Zhou (yichaozhou@uchicago.edu)

Lectures: T-Th 11:00 - 12:20 PM

Classroom: GCIS W123

Course Description

This fast-paced, hands-on course is designed for students who want to apply deep learning techniques to solve problems in genomics. The course focuses on equipping students with the computational skills needed to leverage deep learning in biological research.

Throughout the course, students will work with case studies in genomics, applying deep learning tools to real-world data. Topics will include large language models (LLMs) for DNA sequence analysis, gene expression prediction from DNA sequences (e.g., Enformer, SEI, Borzoi), single cell data analysis, and protein language models (e.g., ESM). Students will be expected to quickly become proficient in essential tools such as the Linux command line, Python, R, GitHub, VS Code, Jupyter notebooks, and popular deep learning frameworks.

The course includes four comprehensive units, each exploring a key domain of AI applications in genomics. In each unit, students will complete background reading of key papers, followed by in-class discussions and hands-on application. Throughout the unit, students will progress from training simplified models, to application using real genomics datasets, culminating in a presentation of their findings. The course concludes in a capstone project, where students will identify an open challenge in genomics and develop an AI-driven solution.

Who should take this course?

- Computational biologists who want to enhance their toolkit with modern deep learning methods.
- Molecular biologists and geneticists with programming experience who want to leverage AI methods in their research.
- Computer scientists and ML researchers interested in applying their expertise to important problems in genomics.

Prerequisites

- Computer literacy, command line familiarity, coding in python and R.
- An introductory statistics course: HGEN 47400 Introduction to Probability and Statistics for Geneticists, or STAT 24400 Statistical Theory and Methods I, or equivalent.
- An introductory course in genetics: BIOS 20187 Fundamental of Genetics or equivalent.
- Familiarity with fundamental concepts in molecular biology and genetics.

Text and Materials

Required readings will be provided for each lecture, and include foundational papers, recent publications, and preprints from arXiv and bioRxiv. Students will also work with public genomic datasets from repositories like SRA, GEO, GTEx, TCGA, ENCODE, and the UK Biobank. Code, tutorials, and additional resources for genomics and deep learning will be made available on the course webpage.

Expectations

Your final grade will be based on the following items:

- Weekly project completion (30%): Projects emphasize hands-on implementation of deep learning models using genomic data.
- Knowledge check summaries and quizzes (10%): Brief quizzes on key concepts and critical evaluations of assigned papers.
- Peer Learning Activities (10%): Collaborative coding sessions and group problem-solving.
- Class Participation (10%): Meaningful contributions to class discussions and presentations.
- Final project (40%): Final research project including idea, implementation, and class presentation.

Honor statement

Academic integrity and honesty are core principles of the University of Chicago and are always expected from its students. We ask that you add an honor statement to the end of every assignment specifying who you consulted or worked with. The statement should also include the following:

“I discussed with xx and yy for this assignment and received help from the tutor (name). In submitting this assignment, I attest that it contains only my own independent work carried out according to the directions given to me by the instructor. I understand that academic integrity and honesty is a core principle of the University of Chicago and is expected at all times from its students.”

Communication and Office Hours

Office hours

Each instructor will be available for office hours on the weeks they are teaching.

- Ran Blekhman (blekhman@uchicago.edu): 2:00 - 3:00 PM Thursday or other time by appointment (office: KCBD 918).
- Haky Im (haky@uchicago.edu): HKI will hold office hours Mondays 3-4pm. You are also encouraged to schedule an appointment at other times as needed. #154E 920 E 58thSt. CLSC (ring the bell to get access)
- Teaching assistant: Charles Zhou (yichaozhou@uchicago.edu): 1:00-2:00pm Friday at CLSC 154 or any time via slack.

Canvas

The announcements will be made through Canvas (link here)

Slack (TODO create slack)

For other types of communication, we will use Slack. You can post your questions about lectures, homework, etc. in slack, and if it's easier for you, you can form study groups and discuss using your own slack channels. To join the slack workspace of the class, use this link:https://join.slack.com/t/2025deeplearn-msn4002/shared_invite/zt-32eztw5ir-C28Yi3iP8ayBlenNEE5Tmg

Special accommodations

Student Disability Services (SDS) works to provide resources, support, and accommodations for all students with disabilities and works to remove physical and mitigate attitudinal barriers, which may prevent their full participation in the life of the university. SDS seeks to ensure that disability is included as a valued part of the institution's diversity and that accessibility is priority. More information here.

If you feel you need accommodation, the process for Biological Science students is outlined on this webpage. As noted there, Student Disability Services engages with students to determine the necessary accommodations, including those temporary in nature. There is a range of accommodations they can be helpful implementing.

Course Schedule

See schedule table here

1. **SETUP: Course overview and setup (week 1)**
    1. crash course in command line
    2. github repo/quarto blog format
    3. conda/virtual environments
    4. install vscode
    5. ssh to supercomputer with GPU (apply for accounts)
    6. Set up environment for training
    7. Obtain google colab accounts (financial aid available)
    8. Toy model training
2. **UNIT 1: natural language models minimal GPT (weeks 2-3)**
    1. Attention is all you need paper
    2. Karpathy's GPT notebook link
    3. Karpathy's video explaining GPT training
    4. Fun stuff (optional)
        1. Spreadsheet implementation of minimal GPT
        2. pico GPT in 60 lines of code with numpy
    5. dna language models (DNABERT, nucleotide transformers)
    6. Henry Raeder's DNA language model project (jupyter notebook)
3. **UNIT 2: DNA seq to epigenome (enformer, sei, BPnet) (weeks 4-5)**
    1. Avsec, Ž., Agarwal, V., Visentin, D. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196–1203 (2021). https://doi.org/10.1038/s41592-021-01252-x
    2. notebook
4. **UNIT 3: single cell methods (weeks 6-7)**
    1. Paper:
    2. Presentation by geneformer collaborator
    3. Application: do something with it
5. **UNIT 4: final project and presentation (weeks 8-9)**
    1. Students pick an approach and dataset, run an analysis, and present their results to the class
    2. Options include any of the methods discussed, or new approaches/datasets, including:
        1. Notebook protein language models ESM/genSLM () 1.
            1. Paper ESM-2 https://www.science.org/doi/10.1126/science.ade2574
        2. Puffin + interpreting promoter sequence grammar (paper puffin)
        3. Alphafold
            1. paper
            2. https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb
        4. microbiome
        5. alpha-missense
        6. deepVariant
        7. EHR, dimension reduction

- class structure
    - reading/videos before lecture
    - quiz
    - Introduction stating the problem in genomics: what are the questions?
    - minimal example jupyter notebook
    - application
    - propose improvement

Suggested reading and other material

Zou, J., Huss, M., Abid, A. *et al.* A primer on deep learning in genomics. *Nat Genet* **51**, 12–18 (2019). https://doi.org/10.1038/s41588-018-0295-5	
Let's build GPT: from scratch, in code, spelled out. "We build a Generative Pretrained Transformer (GPT), following the paper "Attention is All You Need" and OpenAI's GPT-2 / GPT-3. We talk about connections to ChatGPT, which has taken the world by storm. We watch GitHub Copilot, itself a GPT, help us write a GPT (meta :D!) . I recommend people watch the earlier makemore videos to get comfortable with the autoregressive language modeling framework and basics of tensors and PyTorch nn, which we take for granted in this video."	1h56m Karpathy's video explaining the minimal gpt code and training
https://colab.research.google.com/drive/1ZLTicmfZvoOvOcOdcDby2VMkE1rtOpMz?usp=sharing	Boxiang Liu's intro to deep learning approach
https://colab.research.google.com/github/deepmind/deepmind_research/blob/master/enformer/enformer-training.ipynb	Enformer Training notebook
https://colab.research.google.com/drive/1nsXO9A516LIHGEeA6jPdQJmca_NPcxkM?usp=sharing	Enformer usage notebook
https://playground.tensorflow.org/ https://playground.hakyimlab.org	here is a deep learning playground, may help develop intuition
https://hakyimlab.notion.site/Syllabus-of-Deep-Learning-in-Genomics-d1c308f6fc2d4c54ae9c19c492671e2a?pvs=4	Pilot course with undergrads summer 2023

https://www.alcf.anl.gov/alcf-ai-science-training-series	Argonne's Intro to AI-driven Science on Supercomputers: A Student Training Series
https://pytorch.org/tutorials/beginner/deep_learning_60min_blitz.html	Learn pytorch in 60'
https://github.com/ArcInstitute/evo2/blob/main/notebooks/brca1/brca1_zero_shot_vep.ipynb	Variant effect prediction in BRCA1 using evo2, dataset by findlay on saturation mutagenesis of BRCA1
https://www.3blue1brown.com/topics/neural-networks	Visual and intuitive explanation of neural networks including attention and LLMs. If not familiar with linear algebra, you may also watch the linear algebra series

Project ideas

- Random promoter Dream challenge. Can you beat the leaders?
  - Rafi, A.M., Nogina, D., Penzar, D. *et al.* A community effort to optimize sequence-based deep learning models of gene regulation. *Nat Biotechnol* (2024). https://doi.org/10.1038/s41587-024-02414-w
- Predict 3d conformation, paper by Jian Zhou: https://www.nature.com/articles/s41588-022-01065-4
- Present the bpnet paper. Compare to our simpler examples in class. https://colab.research.google.com/drive/1VNsNBfugPJfJ02LBgvPwj-gPK0L_djsD#scrollTo=YRuPISG391E3
- Present the puffin paper. Can you find additional patterns of transcription initiation in humans? https://pubmed.ncbi.nlm.nih.gov/38662817/

2025 Academic calendar

Spring 2025

Date	Event/Deadline
Monday, March 24	Spring Quarter Begins
Saturday, May 24 – Monday, May 26	College Reading Period
Monday, May 26	Memorial Day
Tuesday, May 27 – Friday, May 30	College Final Exams
Monday, June 2 – Friday, June 6	College Senior Week
Saturday, June 7	Convocation
Saturday, June 7	Spring Quarter Ends