CSE7850/CX4803 Machine Learning in Computational Biology Course Syllabus

Instructor Information

Instructor	Email	Office Hours & Location
Yunan Luo	yunan@gatech.edu	Mon, 2:00-3:00 PM on Zoom. (Book a slot on Canvas)

General Information

Description

This graduate-level course focuses on the exciting intersection between machine learning and computational biology. We will cover modern machine learning techniques, including supervised and unsupervised learning, feature selection, probabilistic modeling, graphical models, deep learning, and more. Students will learn the fundamental principles, underlying mathematics, and implementation details of these methods. Through reading and critiquing published research papers, students will learn the applications of machine learning methods to a variety of biological problems in genomics, single-cell analyses, structural biology, and system biology. Students will also learn to implement deep learning models using PyTorch, a popular deep learning library through in-depth programming assignments. In the final project, students will apply what they have learned to real-world data by exploring these concepts with a biological problem that they are passionate about.

This course is appropriate for graduate students or advanced undergraduate students in computer science, bioinformatics, biomedical engineering, mathematics, and statistics. Familiarity with basic linear algebra, statistics, probability, and algorithms is expected. Background knowledge in data analytics and machine learning will be helpful for this course. Students are also expected to have programming experience in Python.

Pre- &/or Co-Requisites

While there are no formal prerequisites for this course, it is intended as a graduate-level course, and as such there is recommended background necessary to keep up with the material covered. Programming skills (specifically Python) are necessary to complete the assignments. Students must have a strong mathematical background (linear algebra, calculus especially partial derivatives, and probabilities & statistics). Experience in introductory courses in Machine Learning (e.g. CS 4641/7641), Data Analytics (e.g., CX4240, CSE 6740), Data Science, or other equivalents is a plus.

Learning Objectives

- Learn how to formulate computational biological questions as machine learning problems.
- Understand fundamental methods in machine learning, including major types of models, underlying mathematical principles, and what types of problems each model is appropriate for.
- Gain experience in reading and reviewing published research papers in machine learning for computational biology.
- Gain practice in using widely-used bioinformatics and machine learning toolbox to analyze biological data.

Course Materials

Course Text
No required textbook. All course slides and reading materials will be made available on Canvas.

Additional Materials/Resources
All additional reading materials will be available on Canvas.

Course Website and Other Classroom Management Tools
Canvas will be used as the course website.

Course Schedule
A tentative schedule is given below and is subject to change depending on the enrollment size. Please see this link for the most up-to-date schedule.

Week	Date	Topic	Contents	Paper Discussion	Deadline (tentative)
1	01/08	Introduction	Introduction & Logistics		
1	01/10	Basics in computational biology	Molecular biology		
2	01/15		No class (MLK day)		
2	01/17		Sequence alignment I		
3	01/22		Sequence alignment II		
3	01/24	ML foundations	No Class (PyTorch video + exercise)		
4	01/29		Regression & Gradient descent		
4	01/31		Classification & Toolbox for Applied ML		
5	02/05		Neural networks		
5	02/07		Deep learning		
6	02/12	Learning from sequence data	Deep learning for Protein/DNA sequences		
6	02/14		Large language models (LLMs)		HW1 Due 02/16
7	02/19	Learning from high-dim data	Clustering and dimensionality reduction		
7	02/21		Generative AI		
8	02/26	Learning from network data	Network basics & ML for graphs		
8	02/28		Graph neural network		HW2 Due TBA
9	03/04	Learning from structure data	Protein structure prediction & generation (AlphaFold, diffusion models)		

9	03/06	Advanced topics: ML for sequence data	Protein language models for prediction and generation	Student presentation	
10	03/11		Disease variant prediction	Student presentation	HW3 Due TBA
10	03/13		ML for protein engineering	Student presentation	
11	03/18		No class (Spring break)		
11	03/20		No class (Spring break)		
12	03/25	Advanced topics: ML for structure data	Deep learning for structure prediction	Student presentation	
12	03/27		GNN for 3D structures	Student presentation	Project proposal due TBA
13	04/01		Deep learning for structure generation	Student presentation	
13	04/03	Advanced topics: ML for network data	Network embeddings	Student presentation	
14	04/08		ML for graph bio data	Student presentation	HW4 Due TBA
14	04/10		ML for drug discovery	Student presentation	
15	04/15	Advanced topics: ML for high-D data	Dim reduction in bio data	Student presentation	
15	04/17		ML for system bio	Student presentation	
16	04/22		ML-guided biological discovery	Student presentation	
					Project report (due on 04/29)

Course Requirements & Grading

Assignments

Assignments	Weight
Homework (3 x HWs + Kaggle)	50%
Paper presentation / Literature review	15%
Project - proposal	5%
Project - final report	25%
Class participation	5%
Total	**100%**
Bonus quizzes*	+3%

Grading scale

Your final grade will be assigned as a letter grade according to the following scale:

A	90-100%
B	80-89%
C	70-79%
D	60-69%
F	0-59%

* If you are taking this course with a S/U (Pass/Fail) grade, you need to earn at least 70% of the grades to pass this course.

Assignment Due Dates

All assignments are due at 23:59:59 PM Eastern Time (ET) on the day the assignment is due unless otherwise noted. If you are located outside of ET, Canvas will display the due dates in your local time (which can be changed by editing your personal Canvas settings). We will not accept assignments submitted late due to time zone issues, so do verify your desired settings as there are no exceptions. See "Course Expectations & Guidelines" for the policies of make-ups and late submissions.

Homework

Homework assignments will include written or programming problems. Students are expected to complete homework problems individually. All homework assignments should be submitted in PDF (together with source code if there is a programming problem) on Canvas/GradeScope.

Paper presentation and literature review

We will discuss cutting-edge research papers during the course. A paper list will be released by the instructor, and two papers will be discussed per lecture during the second half of this course ("Advanced topic" lecture series). For each paper to be discussed, there will be a team of 2 students presenting the paper in class and another team of 2 students writing a literature review for that paper. Students only need to sign up for one option for paper discussion, either the in-class paper presentation or literature review. The team size is subject to change depending on enrollment.

- **Presentation**: Students choosing this option will form teams of 2 students for paper presentations (team size subject to change depending on enrollment). Each team will present a published paper, selected from the paper list provided by the instructor, with a 15-20 min time slot in the class and answer questions by the instructor and other students. The presentation slides should be made as visual (with videos, images, and animations) and clear as possible. Students should practice their talks ahead of time to make sure they are of appropriate length -- not shorter by more than a few minutes, and certainly not longer (we will set a timer that will go off). The presentations should be well-organized and polished. Every member of a group should present (part of) the paper in the presentation. The presentation will be graded according to the following criteria:
    - Quality of Slides: How good are the design, organization, and content in the presentation slides? It is expected to include visuals, diagrams, and bullet points to effectively convey complex information and support the presentation narrative.

    - Presentation Clarity: How clearly and coherently did the presenters communicate their ideas? It includes the ability to convey complex concepts in an accessible manner, maintain a logical flow of ideas, and engage the audience.
    - Question Addressing: How effective are the responses to questions from the audience? This involves demonstrating a thorough understanding of the topic and providing clear and concise answers.
    - Ending on Time: Whether the presenters managed their presentation time effectively, ensuring that all key points were covered within the given duration without rushing or overly exceeding the allotted time.

- **Literature review**: Students choosing this option will form teams of 2 students (team size subject to change depending on enrollment) and write a 4-6 page literature review for a paper chosen from the same paper list released by the instructor. Each team is expected to complete a *comprehensive* review. The team should do a literature search using the selected paper as the "seed paper" and find related papers on the same topics (e.g., prior papers that the seed paper was built upon and follow-up papers that would not be feasible without the methods/results in the seed paper). The literature review should provide a holistic view of the topic. Note that only summarizing each paper individually does not meet the requirement. Preferably, the review needs to also compare or discuss a number of papers on the same topics using a unified framework (e.g., the methodological differences, how the literature evolved to deliver the state-of-the-art methods, different data types used in the ML model, and the key problems, solutions, and future challenges for the topic). The literature review needs to be submitted by midnight the day before the presentation date of that paper. The literature review will be graded according to the following criteria:
    - Paper Selection: Whether relevant and representative papers have been selected for the review, using the paper presented in class as the "seed paper"?
    - Review Structure: The review is expected to employ a cohesive framework for discussing the selected papers collectively, rather than offering individual summaries of each paper. It is suggested to have a comparative analysis that highlights the relationships, differences, and contributions of the papers within the context of the topic.
    - Discussion and Insights: The review should go beyond mere summaries to provide insightful analysis and discussion. High-quality reviews are expected to reflect your critical thinking, interpretations, critiques, and the potential implications of the findings in the reviewed papers.
    - Writing Quality: the clarity, coherence, and overall presentation of the review. This includes adherence to academic writing standards, proper citation practices, and the logical organization of ideas.

Course project

The course project is a group assignment comprising 3-4 members (depending on enrollment). The project group can have different members from the presentation group. The course project is meant for students to gain experience in implementing machine models and applying them to real-world biological data or have a more in-depth understanding of a particular problem by literature review. Examples of project ideas include (1) formulating a novel problem in computational biology as a machine learning problem and implementing machine learning methods you learned in this class to address it; (2) developing novel machine learning methods and applying them to an existing computational biology problem; (3) creating a benchmarking dataset for comparing the performance of different machine learning methods for a specific biology problem, which includes curating and unifying the biological data generated by different papers or

databases and implementing common machine learning methods as baselines (Examples: TDC, OGB, and ATOM3D). Students are encouraged to be creative and come up with their own project ideas. Students who need help with project ideas should talk to the TAs or the instructor. Every member of a group is expected to contribute a substantial part to the project. The contributions of each member should be clearly stated in the final project report.

Class participation

The participation points (5%) include in-class discussion points (2%) and online discussion points (1%x3). To earn the in-class discussion points, you should ask one question to at least one presenting group. The question is expected to be in-depth and preferably prompt discussion, and simple clarification-based questions do not count for the point. You should log your question in a survey form released by the teaching team to receive the points. For online discussion: before each student presentation class, the teaching team will create a post on the online discussion platform. Within that post, you need to write a review of a paper that will be presented by other students in the next class, which includes a short summary (~100 words) and your comments (pros/cons of the paper). The review should be completed by the day before the presentation date. Each review counts as 1 participation point. You can review at most 3 papers to earn the 3 online discussion points.

Bonus quiz

Students will have opportunities to complete 3 quizzes online for bonus points. Each quiz contains multiple choice questions (around 5-10 problems) related to the topics covered in lectures and counts at most 1% extra point for the final grade. The scores of the three quizzes (maximum 3% course points) can be added to a student's total points (capped at 100%).

Technology Requirements and Skills

Computer Hardware and Software

- Laptop or desktop computer with internet connection. Students do not have to have GPU-equipped computers for assignments or projects, as they can utilize free computational resources such as Google Colab to develop GPU-based deep learning models.
- This class will use Canvas to deliver course materials to online students. All course materials and quiz assessments will take place on this platform. Gradescope will be used for the submission of assignments and the project. Ed will be used as the discussion platform. Zoom will be used for remote meetings if needed.

Course Expectations & Guidelines

Communication Policy

You are responsible for knowing the following information:

1. Anything posted to this syllabus
2. Anything emailed directly to you by the teaching team (including announcements via Ed Discussion), 24 hours after receiving such an email or post.

Because Ed announcements are emailed to you as well, you need only to check your Georgia Tech email once every 24 hours to remain up to date on new information during the semester. Georgia Tech generally

recommends students check their Georgia Tech email once every 24 hours. So, if an announcement or message is time-sensitive, you will not be responsible for the contents of the announcement until 24 hours after it has been sent.

University Use of Electronic Email

A university-assigned student e-mail account is the official university means of communication with all students at Georgia Institute of Technology. Students are responsible for all information sent to them via their university-assigned e-mail account. If a student chooses to forward information in their university e-mail account, he or she is responsible for all information, including attachments, sent to any other e-mail account. To stay current with university information, students are expected to check their official university e-mail account and other electronic communications on a frequent and consistent basis. Recognizing that some communications may be time-critical, the university recommends that electronic communications be checked minimally twice a week.

Late and Make-up Work Policy

The late submission policy for homework assignments is as follows: full credit is given if submitted before the due date, 75% credit is given for submissions within 24 hours after the due date, 50% credit is given for submissions within 48 hours after the due date, and no credit is given for submissions after 48 hours past the due date.

There will be no make-up work provided for missed assignments. If you are unable to present during one of the pre-defined presentation days please contact the instructor to coordinate a solution. Of course, emergencies (illness, family emergencies) will happen. In those instances, please contact the Dean of Students office. The Dean of Students is equipped to verify emergencies and pass confirmation on to all your classes. For consistency, we ask all students to do this in the event of an emergency. Do not send any personal/medical information to the instructor or TAs; all such information should go through the Dean of Students.

Plagiarism & Academic Integrity

Georgia Tech aims to cultivate a community based on trust, academic integrity, and honor. Students are expected to act according to the highest ethical standards. All students enrolled at Georgia Tech, and all its campuses, are to perform their academic work according to standards set by faculty members, departments, schools, and colleges of the university; and cheating and plagiarism constitute fraudulent misrepresentation for which no credit can be given and for which appropriate sanctions are warranted and will be applied. For information on Georgia Tech's Academic Honor Code, please visit http://www.catalog.gatech.edu/policies/honor-code/ or http://www.catalog.gatech.edu/rules/18/.

Each student (or project group) must write their own solutions, in their own words, and must properly credit all sources. You are encouraged to discuss problems and papers with others as long as this does not involve the copying of code or solutions. After discussions, all materials that are part of a submission should be wholly your own. A good practice is to allow a 24-hour reflection period post-discussion before you begin working on your assignment independently, avoiding the use of any notes taken during the discussion for your submission. Any public material that you use to gain an understanding of the materials (open-source software, help from a textbook, or substantial help from a friend, etc.) should be acknowledged explicitly in anything you submit to us. To re-emphasize, no matter what the source you cannot copy any existing code, from other students, online, or otherwise, and all code must be wholly your own code that you wrote by yourself. If you have any doubts about whether something is legal or not, please do check with the class

Instructor or the TA. We will actively check for cheating, and any act of dishonesty will result in a Fail grade. Any student suspected of cheating or plagiarizing on a quiz, exam, or assignment will be reported to the Office of Student Integrity, who will investigate the incident and identify the appropriate penalty for violations.

Policy on the uses of AI tools

We treat AI-based assistance, such as ChatGPT and Copilot, the same way we treat collaboration with other people: you are welcome to talk about your ideas and work with other people, both inside and outside the class, as well as with AI-based assistants.

However, all work you submit must be your own. You should never include in your assignment anything that was not written directly by you without proper citation (including quotation marks and in-line citation for direct quotes).

Including anything you did not write in your assignment without proper citation will be treated as an academic misconduct case. If you are unsure where the line is between collaborating with AI and copying AI, we recommend the following heuristics:

Heuristic 1: Never hit "Copy" within your conversation with an AI assistant. You can copy your own work into your own conversation but do not copy anything from the conversation back into your assignment. Instead, use your interaction with the AI assistant as a learning experience, then let your assignment reflect your improved understanding.

Heuristic 2: Do not have your assignment and the AI agent open at the same time. Similar to the above, use your conversation with the AI as a learning experience, then close the interaction down, open your assignment, and let your assignment reflect your revised knowledge. This heuristic includes avoiding using AI directly integrated into your composition environment: just as you should not let a classmate write content or code directly into your submission, so also you should avoid using tools that directly add content to your submission.

Deviating from these heuristics does not automatically qualify as academic misconduct; however, following these heuristics essentially guarantees your collaboration will not cross the line into misconduct.

Accommodations for Students with Disabilities

If you are a student with learning needs that require special accommodation, contact the Office of Disability Services at (404)-894-2563 or http://disabilityservices.gatech.edu/, as soon as possible, to make an appointment to discuss your special needs and to obtain an accommodations letter. Please also e-mail me as soon as possible in order to set up a time to discuss your learning needs.

Collaboration & Group Work

- **Homework**: Students can discuss the homework with any other students in the class but should write their solutions individually.
- **Presentation and project**: The paper presentation and project will be completed as a group. You will sign up as a team and work together throughout the semester. Your team is welcome (and encouraged) to discuss your presentation/project with other members of the class.

Research is highly collaborative and exchanging ideas is expected. However, you must implement your project ideas, write project reports, and create paper presentations within your group.

- **External resources**: We allow and encourage any outside reading material, blog posts, related work, and the use of open-source software for use within your project. Your proposed problem, approach, experiment implementation, and project presentation slides should be the original work of your project group. Any used resources should be cited or acknowledged in your report.

- **Cite Your Resources** [Adapted from Jeff Erickson's course page]: We strongly encourage you to use any printed, online, or living resource at your disposal to help you solve homework problems, but you must cite your sources.
    - If you use an idea from a book, cite the book.
    - If you use an idea from a paper, cite the paper.
    - If you use an idea from Wikipedia, cite Wikipedia.
    - If you use an idea from CS StackExchange, cite CS StackExchange.
    - If you use an idea from your last semester's homework solutions, cite last semester's homework solutions.
    - If you use an idea from another student, cite that student.
    - If you use an idea from your conversation with ChatGPT, cite ChatGPT (note that on the upper-right corner of your ChatGPT conversation window, there is a button to create a shareable URL link that you can use as a citation in your submission)

  There are only two exceptions to this rule. You are not required to cite the following:
    - Official course materials (lecture slides and homework from this semester)
    - Sources for prerequisite material (which we assume you already know by heart)

  Submitting someone else's work without giving them proper credit is plagiarism, *even if you have the other person's explicit permission*. Citing your sources will *not* lower your homework grade. Allowing someone else to use your ideas without giving you credit is also an academic integrity violation.

- **Use Your Own Words** [Adapted from Jeff Erickson's course page]: Verbatim duplication of *any* source, even *with proper citation*, is plagiarism. In particular:
    - Copying verbatim from the lecture slides is plagiarism.
    - Copying verbatim from a reference book is plagiarism.
    - Copying verbatim from MLB homework from previous years is plagiarism.
    - Submitting work done entirely by another student is plagiarism.
    - Allowing another student to copy your work verbatim is also an academic integrity violation.

Student-Faculty Expectations Agreement

At Georgia Tech we believe that it is important to strive for an atmosphere of mutual respect, acknowledgment, and responsibility between faculty members and the student body. See http://www.catalog.gatech.edu/rules/22/ for an articulation of some basic expectations that students can have of the instructor and that the instructor has of students. In the end, simple respect for knowledge, hard work, and cordial interactions will help build the environment we seek. Therefore, the instructor encourages students to remain committed to the ideals of Georgia Tech while in this class.

Student Use of Mobile Devices in the Classroom

Use of Mobile Devices, Laptops, etc. During Class. As research on learning shows, unexpected noises, and movement automatically divert and capture people's attention, which means you are affecting everyone's learning experience if your cell phone, pager, laptop, etc. makes noise or is visually distracting during class. That said, many students find it useful to have a mobile device on hand to access course materials. With this in mind, we allow you to take notes on your laptop but request that students turn the sound off so that they do not disrupt other students' learning. In addition, if you are doing anything other than taking notes or looking at course materials on your laptop, please sit in the back row so that other students are not distracted by your screen.

Institute-Approved Absences

As per Georgia Tech policy, you are permitted to be absent from class to participate in athletic events, official field trips, and religious observances. For planning purposes, please provide me with written notice of your upcoming absence at least two weeks before the event, and ideally within the first two weeks of class. When I receive this notice, you and I will discuss opportunities to make up the work you will miss in your absence. Please see http://catalog.gatech.edu/rules/4/ for more information about receiving official notice from the Registrar about the nature and timing of your upcoming Institute-approved absence.

Subject to Change Statement

The syllabus and course schedule may be subject to change. Changes will be communicated via the Canvas announcement tool. It is the responsibility of students to check Ed Discussions, email messages, and course announcements to stay current in their online courses.