MCB128 AI in Molecular Biology (Spring 2026)

https://mcb128.org
Mon/Wed 10:30-11:45

Instructors: Dr Elena Rivas elenarivas@fas.harvard.edu

Course Description

What are convolutional neural networks (CNNs) and how are they used to predict sequence motifs in biological sequences? What is a transformer and how it is used by AlphaFold to predict protein structures?
AI/deep-learning methods are now consistently used to approach many computational questions in molecular biology such as: motif finding, homology of DNA/RNA and proteins, structure prediction for DNA/RNA and proteins, amongst others. The objective of this course is to introduce AI concepts and methods in the context of important questions in computational biology. A given question (i.e. protein folding) will be paired to a AI method (transformers), and an in-depth description of both will be provided. The goal is both to describe the fundamental algorithms as well as some important AI implementation.
This course will explore the major advances in deep learning, with a special emphasis on their applications to molecular biology and genomics. Starting from a single neurons (perceptrons), we will progress to more complex architectures such as convolutional and recurrent neural networks, transformers, and generative neural networks. The course will cover both the general principles of these methods as well as specific applications in genomics. This is a computationally rigorous course for students interested in computational biology.

Course Goals

Students taking this course will come out with a mathematical knowledge of the most successful and frequently used deep learning methods, and an understanding of their uses and best applications for important computational questions in molecular biology. The students will also acquire knowledge of essential methods in Python and PyTorch used in deep learning. After taking MCB128, students will be able to design and implement their own deep learning methods for new computational questions of interest.

Course Format

The course is divided in 6 blocks (two weeks each), starting with one foundational 0-block (one week) dedicated to the Single Artificial Neuron. Blocks are structured more or less in chronologically: from perceptrons (feed-forward neural networks), to convolutional and recurrent neural networks, transformers and beyond. Each block will describe one or more related deep-learning methods, a biological question for which they been applied, a state of the art existing approach, and a specific deep-learning implementation for that biological question.

Each block expands two weeks. Each week includes two 75 minutes lectures, plus one discussion section lead by the TFs. The lectures will be given by Elena Rivas, and there will have extensive notes with bibliography, and slides of the actual lectures for students to review at any time.
Sections will be devoted to specifics of coding the methods, and to describe in detail specific deep-learning concepts and jargon standard in the field, such as, tensor, backpropagation, masking, embedding, broadcasting, distillation, regularization, loss functions, and many others. Specific homework coding questions will also be discussed in sections. Questions brought up by the students that would help them with homework would also be addressed in the weekly discussion sections.

Assignments and grading

The course is divided in 6 blocks. There will be one homework and one (very short) in-class quiz per block. The final grade will be based on all six homework (80%) and quizzes (10%) plus participation (10%). Participation includes: attendance to class, participation in Discussion sections, Student Hours or other forums such as slack.

When is the course typically offered?

Starting Spring 2026. Expected to be offered both semesters after that.

Typical Enrollees

This course is primarily designed for undergraduates (mainly juniors and seniors) as well as early graduate students with interdisciplinary interests in computation, molecular biology and mathematics. Some basic knowledge of computing (python AM10 level) is strongly encouraged, as well as some basic knowledge of molecular biology (LS1 level) and statistics (STAT110) and algebra (MA21). Having expertise in at least two of the three areas is recommended.

Getting started with coding

We will build from scratch our deep-learning toolkit, mainly using Python and PyTorch. Some introductory course to Python (such as the FAS python informatics workshop offered in the Fall 2025 would be extremely useful. There are also the python self-taught tutorial.
For deep learning, we will use mostly PyTorch. These is two good Python and PyTorch books with deep learning in mind, that we plan to have available from the Harvard library and that you should be able to access from canvas under the tab “library reserves”,

1. “Deep learning with Python”, Francois Chollet, Manning Publications Co., 2018.

2. “Deep learning with PyTorch”, Adrian Tam, 2023.

You will need a laptop. If you need help procuring one, please let us know.
We will use **Google Colab** to write our code and homework notes. We have **Google Cloud Educational credits** to run our homework directly on their hardware. Once you enroll, we will provide you with a key to redeem your Google Cloud coupon.

What can students expect from you as an instructor?

I am a computational biologist specialized in genomes and RNA structure analysis, which means that I like biological sequences and all the secrets that they hide, for which I like to design and implements new algorithms. Many years ago I was a theoretical physicist, which means that I also like math-although mathematicians make fun of the "pragmatic" ways physicist use math.
I like to understand things from first principles (much easier than memorizing them), and I bring that to the classroom. I will go into mathematical details to a certain extent, although depending on your own background and interests, that may not be completely necessary to be successful doing the homework. I also think that coding can be a lot of fun, and I expect to show you that.
I think that in the very near future any biologist would need to know some (deep-learning) computation, and have a good grasp of statistics. I hope this class can help you to take a step in that direction.
In addition to Student Hours (held by me and the TFs), I am often available for one-on-one interactions at other times.

Course Policies

Absence

Absences will be reflected in the 10% of the grade reserved for participation in the course.

Late work

Late work will have a penalty of lowering the maximum possible grade by 10% for each day that it is late. If you can anticipate (with at least 1 week) an important reason why your homework will be turned in late, you may request an exception.
Additionally, resubmission of homework (after grading) is allowed. The max grade increase is 20% of current grade.

Academic Integrity

You are encouraged to discus your homework with your classmates, but the work you present has to be your own and be written in your own words. If someone helped you with a particular aspect, add a comment in your code explaining who helped you and what is the contribution, and you want to demonstrate that you understand it. This also applies to any materials taken from GAI, and the Internet.

AI Policies

As for using ChatGPT and other generative AI (GAI) tools to produce your homework, these are the policies:

- You can look at AI generated materials at any time during the completion of your work, but you cannot present AI generated material as your own. Any AI use must be appropriately acknowledged and cited. Any AI code has to be annotated and modified for you particular purpose.

- AI materials should be considered as a tool, often useful for you to learn a technique or to see code implementing it. But be aware that these contents (unlike those in published books with authors responsible for them), are not produced with the expectation that they are correct or accurate.

- Ultimately, you are responsible for all the work that you present as yours, and it has to be written in your own words. A straight AI generated log cannot be presented as homework. Violations of this policy will be considered academic misconduct.

We draw your attention to the fact that different classes at Harvard could implement different AI policies, and it is the student's responsibility to conform to expectations for each course.

Tentative Syllabus

Block	week	Dates 2026	Description	Due
b0	**b0-w1**	01/26,01/28 01/30	A single neuron / DNA functional classification [1, 2] section_b0-w1 (Colab,Pytorch)	hw-b0 out
b1	**b1-w1**	02/02,02/04 02/06	Multi-layer-Perceptron / Protein 2D structure [3, 4] section_b1-w1	hw_b1 out / hw_b0 due
	b1-w2	02/09,02/11 02/13	Backpropagation / [5] section_b1-w2	quiz_b1
b2	**b2-w1**	02/18 02/20	Convolutional Neural Networks / DNA sequence motifs section_b2-w1	hw_b2 out / hw_b1 due
	b2-w2	02/23,02/25 02/27	Recurrent Neural Networks / section_b2-w2	quiz_b2
b3	**b3-w1**	03/02,03/04 03/06	Self-Attention / Transformers for alignments section_b3-w1	hw_b3 out / hw_b2 due
	b3-w2	03/09,03/11 03/13	Protein folding / AlphaFold2 section_b3-w2	quiz_b3
b4	**b4-w1**	03/23,03/25 03/27	Large Language Models (LLMs) section_b4-w1	hw_b4 out / hw_b3 due
	b4-w2	03/30,04/01 04/03	LLMs for DNA/RNA proteins (DNABERT, ESM, ProGen) section_b4-w2	quiz_b4
b5	**b5-w1**	04/06,04/08 04/10	AutoEncoders / scRNA-seq section_b5-w1	hw_b5 out / hw_b4 due
	b5-w2	04/13,04/15 04/17	Variational AutoEncoders / cryoEM section_b5-w2	quiz_b5
b6	**b6-w1**	04/20,04/22 04/24	Diffusion Models / Protein design section_b6-w1	hw_b6 out / hw_b5 due
	b6-w2	04/27,04/29	Graph Neural Networks / Antibiotic discovery	
Final homework		05/06		hw_b6 due

Books

1. "Deep learning with Python", Francois Chollet, Manning Publications Co., 2018.
2. "Deep learning with PyTorch", Adrian Tam, 2023.

References

[1] David J. C. MacKay, editor. *Information Theory, Inference, and Learning algorithms*, chapter 29, pages 471–481. Cambridge University Press, 2003. URL `http://rivaslab.org/teaching/MCB128_AIMB/downloads/Mackay.pdf`.

[2] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *NAR*, 10:2997–3011, 1982. URL `http://rivaslab.org/teaching/MCB128_AIMB/downloads/Stormo83.pdf`.

[3] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psycological Review*, 65:386–408, 1956. URL `http://rivaslab.org/teaching/MCB128_AIMB/downloads/Rosenblatt1958.pdf`.

[4] N. Qian and T. J. Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, 202:865–884, 1988. URL `http://rivaslab.org/teaching/MCB128_AIMB/downloads/QianSejnowski88.pdf`.

[5] D. E. Rumelhart, G. E. Geoffrey E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. URL `http://rivaslab.org/teaching/MCB128_AIMB/downloads/Rumelhart86.pdf`.