Tentative Syllabus

Block	week	Dates 2026	Description	Due
b0	b0-w1	09/02 09/04	Intro / A single neuron / section_b0-w1	hw_b0 out
	b0-w2	09/09 09/11	A single neuron / RNA functional classification [1, 2] section_b0-w2	quiz_b0
b1	b1-w1	09/14,09/16 09/18	Multi-layer-Perceptron / Protein 2D structure [3, 4] section_b1-w1	hw_b1 out hw_b0 due
	b1-w2	09/21,09/23 09/25	Fundamental of Neural Network Training / [5, 6] section_b1-w2	quiz_b1
b2	b2-w1	09/28,09/30 10/02	Convolutional Neural Networks / DNA/RNA sequence binding motifs [7–9] section_b2-w1	hw_b2 out hw_b1 due
	b2-w2	10/05,10/07 10/09	Recurrent Neural Networks / Regulatory motif prediction [10] Splice site prediction [11] section_b2-w2	quiz_b2
b3	b3-w1	10/14 10/16	Self-Attention / Transformers for alignments [12, 13] section_b3-w1	hw_b3 out hw_b2 due
	b3-w2	10/19,10/21 10/23	Protein folding / AlphaFold2 [14] section_b3-w2	quiz_b3
b4	b4-w1	10/26,10/28 10/30	Large Language Models (LLMs) [15] section_b4-w1	hw_b4 out hw_b3 due
	b4-w2	11/02,11/04 11/06	LLMs for DNA/RNA, [16] proteins, [17–19] and genomes [20] section_b4-w2	quiz_b4
b5	b5-w1	11/09,11/11 11/13	AutoEncoders / Gene expression profiles [21] section_b5-w1	hw_b5 out hw_b4 due
	b5-w2	11/16,11/18 11/20	Variational AutoEncoders / scRNA-seq [22] section_b5-w2	quiz_b5
b6	b6-w1	11/23	Diffusion Models / Protein design [23]	
	b6-w2	11/30,12/02 12/04	Flow Matching / RNA seq&structure generation [24] section_b6-w1	hw_b5 due
Final project		due 12/18		

References

[1] David J. C. MacKay, editor. *Information Theory, Inference, and Learning algorithms. Chapter 39*, chapter 39, 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.

[6] S. J. D. Prince, editor. *Understanding Deep Learning*. The MIT Press, 2023. URL https://udlbook.github.io/udlbook/.

[7] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey. Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. *Nature Biotechnology*, 33, 2015. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/Alipanahi15.pdf.

[8] J. Zhou and O. G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods*, 12:931–934, 2015. doi: 10.1038/nmeth.3547. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/ZhouTroyanskaya15.pdf.

[9] P. K. Koo and S. R. Eddy. Representation learning of genomic sequence motifs with convolutional neural networks. *PLOS CB*, 15:e1007560, 2019. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/KooEddy19.pdf.

[10] DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *NAR*, 44:e107, 2016. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/QuangXie16.pdf.

[11] N. Singh, R. Nath, and D. V. Singh. Splice-site identification for exon prediction using bidirectional LSTM-RNN approach. *Biochemistry and Biophysics Reports*, 30, 2022. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/Singh22.pdf.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/Vaswani17.pdf.

[13] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. F. Canny, P. Abbeel, T. Sercu, and A. Rives. MSA Transformer. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8844–8856. PMLR, 18–24 Jul 2021. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/MSATransformer_Rivas.pdf.

[14] J. Jumper and et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/Jumper21.pdf.

[15] T. Bepler and B. Berger. Learning the protein language: Evolution, structure, and function. *Cell Systems*, 12: 654–669, 2021. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/BeplerBerger21.pdf.

[16] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37:2112–2120, 2021. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/Ji21.pdf.

[17] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C.L. Zitnick, J. Ma, and Fergus. R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 118:e2016239118, 2021. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/Rives_2021.pdf.

[18] Z. Lin, H. Akin, R. Rao, B Hie, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 2023. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/Lin23.pdf.

[19] E. Nijkamp, J. A. Ruffolo, E. N. Weinstein, N. Naik, and A. Madani. ProGen2: exploring the boundaries of protein language models. *Cell Systems*, 14:968–978, 2023. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/Nijkamp23.pdf.

[20] E. Nguyen, M. Poli, M. G. Durrant, B. Kang, and B. L. Hie. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386, 2024. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/Nguyen24.pdf.

[21] U.S. National Cancer Institute. Computational Resources for Cancer Research. Gene Expression Autoencoder - P1B1 (GeneEx Autoencoder). URL https://computational.cancer.gov/model/gene-expression-autoencoder-p1b1.

[22] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nat Methods*, 15:1053–1058, 2018. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/Lopez18.pdf.

[23] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. Vázquez Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, Baek M., and D. Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620:1089–1100, 2023. URL http://rivaslab.org/teaching/MCB128_AIMB/downloads/Watson23.pdf.

[24] D. Rubin, A. dos Santos Costa, M. Ponnapati, and J. Jacobson. RIBOGEN: RNA sequence and structure co-generation with equivariant multiflow. *https://doi.org/10.48550/arXiv.2503.02058*, 2025.

Books

1. "Understanding Deep Learning", Simon J.D. Prince, 2023.
2. "Deep learning with Python", Francois Chollet, Manning Publications Co., 2018.
3. "Deep learning with PyTorch", Adrian Tam, 2023.
4. "Introduction to probability", Blitzstein and Hwang, 2019,

MCB128 AI in Molecular Biology (Spring 2026)

http://rivaslab.org/teaching/MCB128_AIMB
Mon/Wed/Fri 10:30-11:45

Instructor: 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 7 blocks. 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

There will be six regular homework assignments corresponding to blocks b0-b5, plus a final homework assignment consisting of a project of your choice. The final grade will be based on all six homework plus project (70%) and quizzes (30%). Your participation is very important to create a good learning environment. Participation includes: attendance to class, asking questions both during lectures and sections, Student Hours or other forums such as slack. Participation could be used to decide when a student is at the border between two letter grades.

When is the course typically offered?

Fall 2026. Expected to be offered every Fall semester 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. "Understanding Deep Learing", Simon J.D. Prince, 2023.
2. "Deep learning with Python", Francois Chollet, Manning Publications Co., 2018.
3. "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.
Resubmission of homework (after grading) is allowed under the following schedule: for a homework due on a given Friday at 23:59, you can expect to receive the graded work with comments by the following Wednesday. Resubmissions will then be accepted until the second Friday at 23:59. 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.