

Topics in Sequence-Based Machine Learning for Bioinformatics

CS 225

Spring 2026 Section 01 In Person 3 Unit(s) 01/22/2026 to 05/11/2026 Modified 01/26/2026

Contact Information

Instructor(s): William "Bill" Andreopoulos

Office Location: Online (former MacQuarrie Hall 416)

Telephone: (408) 924 5085

Email: william.andreopoulos@sjsu.edu

Office Hours: Friday 9:00-11:00am

Class Days/Time: Monday and Wednesday, 7:30-8:45pm

Classroom: MQH 422

Course Information

Course Format

This course adopts an in-person classroom delivery format.

Faculty Web Page and MYSJSU Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found on Canvas Learning Management System course login website at <http://sjsu.instructure.com>. You are responsible for regularly checking with the course messaging system to learn of any updates. You should modify the Canvas settings for notifications of announcements and messages to be sent to you.

Course Description and Requisites

A study of recent advances in machine learning methods with applications to solving sequence analysis problems in molecular biology. The methods examined include word embeddings, vector space representations, language models, and deep learning architectures. A substantial course project is required.

Prerequisite(s): BIOL 123B and MATH 162, or CS156, or CS171, or instructor consent. Graduate standing. Allowed Declared Major: Computer Science MS, Bioinformatics MS, and Data Science MS.

Letter Graded

* Classroom Protocols

Communication with the instructor

As this is an in-person section, course-related communication should preferably be done in-person during the regular class meeting time or office hours. For online communication, students should use the course Discord channel. Rather than emailing redundant questions to the teaching staff, students should post questions on the course Discord channel where the entire class can read and benefit from the responses. The system is catered to getting students help efficiently from classmates, the TA, embedded tutor, and the instructor. *Private messages sent to the instructor's other email addresses may get lost due to the large volume of emails received.*

The professor responds primarily to the Discord channel. The professor will re-post questions that are of general interest (e.g. about homework) or discuss them in class. Students are responsible for everything said in class. It is students' responsibility to keep up with what is said in class and not re-post the same questions repeatedly.

When students use the course Discord channel, they are expected to be identifiable through their names. Anonymous postings are unacceptable. Students who use fake nicknames may be removed from the Discord channel.

The instructor does not write messages after normal business hours, on weekends or holidays.

Technical trouble-shooting should be done during the office hours.

Never email your entire code for an assignment to the instructor. Limit the code you post to 20 lines or less.

Announcements that concern everyone, such as reminders about due dates or class policy, will be posted.

Class Attendance

Attendance (in-person or via Zoom) is highly recommended. When technology allows, Zoom will be available. Students are responsible for following all material presented in class.

The polling questions in the slides are in the form of multiple-choice and true-false questions. Students should participate and follow the polling questions, either via Zoom polling or Zoom chat or ask in class.

Regrading Procedure

Grades assigned are final, unless there was an error in the grading. Special requests (e.g. grade changes or deadline extensions) should be done in-person; such special requests sent via electronic messages to the teaching staff will be disregarded, since this is an in-person section. To request a higher grade, students should first submit the Canvas "Regrade request" form so there is a record of the request, and afterwards speak with the professor. Grades may be reevaluated at anytime and may go down as a result of a regrade. There will be no regrades after the end of the semester (final exam).

At the end of the semester grade roundups (e.g. 89.95% to 90%) will only be considered if a student has pursued any extra credit opportunities offered and completed the SOTE evaluation.

Classroom Protocol

Students on Zoom should be muted when not speaking, and must be dressed appropriately when their camera is on.

Course material developed by the instructor is the intellectual property of the instructor. Students can not publicly share or upload instructor generated material for this course such as exam questions, lecture notes, hands-on exercises or homework solutions without instructor permission.

Program Information

Diversity Statement - At SJSU, it is important to create a safe learning environment where we can explore, learn, and grow together. We strive to build a diverse, equitable, inclusive culture that values, encourages, and supports students from all backgrounds and experiences.

Course Learning Outcomes (CLOs)

Upon successful completion of this course, students will be able to:

1. Use machine learning and deep learning in bioinformatics sequence analysis to answer biological questions and to generate biological hypotheses.
2. Comprehend the nature, scope and limits of using machine learning and deep learning in the field of bioinformatics.
3. Develop machine learning and deep learning solutions for sequence data.
4. Compare different machine learning algorithms and choose a solution based on suitability for a particular data set.
5. Compare biomolecular analysis with machine learning to analysis with classical bioinformatics tools.
6. Appreciate some of the most challenging problems in life sciences that use machine learning methods, possess insight into how to solve those problems.

Course Materials

Texts/Readings

We don't use a specific textbook in this class as there exists a lot of relevant material on bioinformatics found in various references. The reading material will be the slides, references and handouts.

A copy of my slides will be available to the students enrolled in the class.

Additional handouts will be provided through Canvas.

Major references:

- Data Analytics in Bioinformatics: A Machine Learning Perspective, 1st Edition (2021). by Rabinarayan Satpathy, Tanupriya Choudhury, Suneeta Satpathy, Sachi Nandan Mohanty, Xiaobo Zhang (Editors). ISBN-13: 978-1119785538.
- Haoyang Li, Shuye Tian, Yu Li, Qiming Fang, Renbo Tan, Yijie Pan, Chao Huang, Ying Xu, Xin Gao. Modern deep learning in bioinformatics. *Journal of Molecular Cell Biology*, Volume 12, Issue 11, November 2020, Pages 823–7.
- Deep learning in bioinformatics. Edited by Xin Gao, Wei Wang. Elsevier Methods. Volume 166, 15 August 2019, Pages 1-120.
- Walsh, Ian; Pollastri, Gianluca; Tosatto, Silvio C. E. (September 2016). "Correct machine learning on protein sequences: a peer-reviewing perspective". *Briefings in Bioinformatics*. 17(5): 831–840.
- Chicco, D (December 2017). "Ten quick tips for machine learning in computational biology". *BioData Mining*. 10 (35): 35.
- Yang, Yuedong; Gao, Jianzhao; Wang, Jihua; Heffernan, Rhys; Hanson, Jack; Paliwal, Kuldeep; Zhou, Yaoqi (May 2018). "Sixty-five years of the long march in protein secondary structure prediction: the final stretch?". *Briefings in Bioinformatics*. 19 (3): 482–494.
- Wang, Sheng; Peng, Jian; Ma, Jianzhu; Xu, Jinbo (January 2016). "Protein secondary structure prediction using deep convolutional neural fields". *Scientific Reports*. 6: 18962.

Other technology requirements / equipment / material

Students will use colab.research.google.com and create Jupyter notebooks in Python to ensure their work is shareable and reproducible.

Course Requirements and Assignments

SJSU classes are designed such that in order to be successful, it is expected that students will spend a minimum of forty-five hours for each unit of credit (normally three hours per unit per week), including preparing for class, participating in course activities, completing assignments, and so on.

Reading assignments: Readings will regularly be assigned for the next class (see schedule). Slides will be posted under the Canvas modules before the next class.

Hands-On Worksheets: We will have a number of hands-on worksheets. The hands-on worksheets will involve use of bioinformatics tools. The purpose of the hands-on exercises is to develop your understanding of the material and skills in using the tools.

The Hands-On worksheets will involve learning how to use machine learning and deep learning tools with the Python programming language for performing bioinformatics analysis. Students will use colab.research.google.com and create Jupyter notebooks in Python to ensure their work is shareable and reproducible.

Term Project and In-Class Presentation: There will be a term project. It is a group project. Each group consists of two students. A list of possible projects will be provided to you by the instructor.

Team Formation is due on Monday, February 2, 2026.

A Progress Report is due on Monday, April 6, 2026 (after Spring Recess).

The final project is due on Monday, May 11, 2026.

The in-class presentations will also take place from May 4-11, 2026.

A grading rubric will be provided.

All homework should be submitted on Canvas, not by e-mail.

Late policy: Late penalty is 2% per day up to 14 days. After 14 days (or after the last day of classes if it is sooner) the submission page will be closed and will not be re-opened. No submission will be accepted after the closing deadline.

Examinations

Midterm exams: There will be two Midterm exams during the semester.

Final exam: One final cumulative exam.

The exams will contain multiple choice questions, true/false and short answer questions. Exams are closed book, closed notes, and comprehensive. Exams are in-person. The exams should be done individually. No make-up exams except in case of verifiable medical reasons.

Presentation of a research paper: Each student should present an influential research paper of his/her choice, which is related to their project topic, to one of the classes. Students should sign up in the given spreadsheet for a date to present a paper. The paper, chosen by the student, should either use machine learning/deep learning towards making a biological discovery or introduce a novel tool for Natural Language Processing or text mining or bioinformatics. The presentation should last for no more than 10 minutes followed by Q&A. A grading rubric will be provided.

✓ Grading Information

The course grade is based on:

Hands-On Worksheets: 10%

Midterms: 20%

Final: 20%

Project: 40%

Presentation of a research paper: 10%

<i>Grade</i>	<i>Points</i>	<i>Percentage</i>
<i>A plus</i>	<i>960 to 1000</i>	<i>96 to 100%</i>
<i>A</i>	<i>930 to 959</i>	<i>93 to 95%</i>
<i>A minus</i>	<i>900 to 929</i>	<i>90 to 92%</i>
<i>B plus</i>	<i>860 to 899</i>	<i>86 to 89 %</i>
<i>B</i>	<i>830 to 859</i>	<i>83 to 85%</i>
<i>B minus</i>	<i>800 to 829</i>	<i>80 to 82%</i>
<i>C plus</i>	<i>760 to 799</i>	<i>76 to 79%</i>
<i>C</i>	<i>730 to 759</i>	<i>73 to 75%</i>
<i>C minus</i>	<i>700 to 729</i>	<i>70 to 72%</i>
<i>D plus</i>	<i>660 to 699</i>	<i>66 to 69%</i>

<i>Grade</i>	<i>Points</i>	<i>Percentage</i>
<i>D</i>	<i>630 to 659</i>	<i>63 to 65%</i>
<i>D minus</i>	<i>600 to 629</i>	<i>60 to 62%</i>

University Policies

Per [University Policy S16-9 \(PDF\)](http://www.sjsu.edu/senate/docs/S16-9.pdf) (<http://www.sjsu.edu/senate/docs/S16-9.pdf>), relevant university policy concerning all courses, such as student responsibilities, academic integrity, accommodations, dropping and adding, consent for recording of class, etc. and available student services (e.g. learning assistance, counseling, and other resources) are listed on the [Syllabus Information](https://www.sjsu.edu/curriculum/courses/syllabus-info.php) (<https://www.sjsu.edu/curriculum/courses/syllabus-info.php>) web page. Make sure to visit this page to review and be aware of these university policies and resources.

Course Schedule

Week	Topic
01/26-01/30	Introduction, overview of unsupervised and supervised ML in bioinformatics
02/02-02/06	Essentials of machine learning in bioinformatics, NLP and text mining
02/09-02/13	Sequence classification with Linear and Logistic Regression
02/16-02/20	Language models using k-mers and word embeddings
02/23-02/27	Vector space representations: clustering & visualization with PCA, t-SNE, UMAP
03/02-03/06	Hidden Markov Models and Markov chains

03/09-03/13	Review for midterm with problem-solving exercises / <i>Midterm 1</i>
03/16-03/20	Sequence classification with Naive Bayes
03/23-03/27	Deep Learning introduction, fundamentals and architectures
03/30-04/03	<i>Spring recess</i>
04/06-04/10	Deep Learning in bioinformatics: CNNs, LSTMs, Recurrent Neural Networks (RNNs), Long Short Term Memory (LSTM) neural networks for sequence modelling
04/13-04/17	Word embeddings and language models with neural networks, transformers, BERT, transfer learning
04/20-04/24	Review for midterm with problem-solving exercises / <i>Midterm 2</i>
04/27-05/01	Efficient sequence searching, min-hashing, locality-sensitive hashing, vector quantization
05/04-05/08	Case studies using deep learning in bioinformatics / Project discussion / Project presentations
05/11	Project presentations. Review, wrap-up
	Final exam on Wednesday, May 13, 7:45-9:45 PM

The schedule is subject to change with fair notice.