EXMD 521/BMDE 521: Computational methods for single-cell genomics data analytics

General information	
Course #:	EXMD521/BMDE521
Section #:	001
Term and year:	Fall 2023
Course pre-requisite(s):	MATH 203, MATH 223, MATH 323, and one course in programming (or equivalent background with permission of instructor). Additional background in Calculus and Probability is required (e.g., MATH 222 and MATH 204).
Course co-requisite(s):	None
Course schedule (class day(s) and time):	Wed. 8:30-11:30
Number of credits:	3
Instructor information	
Name and title:	Dr. Jun Ding, Assistant Professor, Department of Medicine, Associate Member, Department of Biomedical Engineering, McGill University
Email:	jun.ding@mcgill.ca
Office hours:	1-3PM Friday
Communication plan:	The students can find me at my office (Glen site EM3.2212) during office hours for questions. For those students who cannot be in the office in person for various reasons, I will also be available by Zoom at https://mcgill.zoom.us/j/3418431568 during office hours.
TA information	
Name/Email:	Yumin Zheng (yumin.zheng@mail.mcgill.ca
Office hours:	

Course overview

With the advances in sequencing biotechnologies, the complexity of biomedical datasets has been everincreasing in the past decade. Compared to conventional bulk sequencing, the emerging single-cell measurements (particularly the multi-omics ones) can comprehensively profile biological processes in high resolution and large scale. It provides unrivaled opportunities to derive a deeper understanding of the biological mechanisms underlying various bioprocesses, which can revolutionize the development of novel diagnostic and therapeutic regimens. However, single-cell sequencing data is much more highdimensional, large-scale, sparse, and noisy compared to the bulk sequencing counterparts, which presents new computational challenges in methods for the representation, inference, and learning of single-cell data in support of biomedical research. In the past years, numerous computational methods (especially machine learning approaches) have been developed to address the above challenges, which substantially promotes single-cell applications in a large variety of studies (e.g., finding disease markers and drug candidates). This course will discuss popular computational methods (particularly machine learning approaches) for various single-cell data analysis, modelling, and visualization tasks that could drive novel biomedical discoveries with hands-on real-world examples and exercises.

Instructor message regarding course delivery

- Besides the in-person classes, the course will also be delivered via the Zoom platform to promote course availability. The detailed Zoom link is to be determined.
- All the course contents, learning materials, and resources will be available to the students via the myCourses platform.
- Students could seek support from McGill student service (<u>https://www.mcgill.ca/studentservices/</u>) if they are feeling overwhelmed by their academic work and/or would like to further develop their time and workload management skills.

Learning outcomes

By the end of the course, the students are expected to understand/master the following:

- What is single-cell genomics data, and why do they help solve complex biomedical problems?
- How to preprocess and normalize single-cell genomics data (single-cell RNA-seq, single-cell ATAC-seq, single-cell spatial transcriptomics, etc.)
- How to analyze single-cell genomics data (reducing dimensions, denoising/imputation, clustering cells, annotating cell populations, inferring pseudo time/cellular trajectories, cell-cell interaction inference) with existing methods
- How to develop novel computational models to address emerging computational challenges (e.g., single-cell spatial data analysis) in single-cell genomics data analytics?
- How to integrate single-cell multi-omics data to understand cellular dynamics?

Instructional methods

- Lectures and programming exercises with analysis cases in Python (JupyterLab)
- Students will need to install Anaconda and JupyterLab (all freely available) to participate in the course
- All the students should participate in the course (either in person or on Zoom) for the lectures
- If students anticipate that they cannot participate in certain course components, they should inform the instructor at least 10 hours before the lecture.

Note that instructional methods are subject to change based on public health protocols (<u>https://www.mcgill.ca/maxbellschool/programs/resources-current-students/covid-19-protocols-and-resources</u>).

Expectations for student participation

The students should attend all the classes. While in-person participation is the standard, online participation via Zoom is also accepted in case the students notify the instructor at least 10 hrs before the lecture by email.

Class recordings

All the lectures will be recorded, and the videos will be shared via the myCourses platform. Students should check their emails and myCourses for course updates at least once a week. In addition, students can download the myCourses Pulse mobile app to stay connected and on track.

Intellectual property considerations

I ask for everyone's cooperation in ensuring that the videos and associated materials are not reproduced or placed in the public domain. This means that each of you can use it for your own purposes, but you cannot allow others to use it by posting it online or giving it or selling it to others who may copy it and make it available. Thank you very much for your help with this.

Required course materials

Course material, prepared by the Lecturer, will be available to the registered students via myCourses.

Course content

In this course, we will discuss the computational methods (mostly machine learning models) for common single-cell data analysis tasks: (1) Data preprocessing; (2) Dimensionality reduction; (3) Cell embedding learning and Cell Clustering; (4) Trajectory Inference; (5) Gene Regulatory Network Inference; (6) Cell-Cell Interaction inference (7) Single-cell Spatial data modelling (8) Multi-omics Data Integration. These machine learning methods could be categorized into three significant learning challenges: (1) Representation: how to infer good embeddings for representing the cells and subsequent analyses (2) Inference: how to infer the probability for a particular cellular event given the model. (3) Learning: Learn the model (structure and parameters) that could represent the probability density of cells (e.g., in terms of gene expression). Each single-cell data analysis task could span one or multiple learning challenges. Therefore, we will organize the lectures into different single-cell data analysis tasks, and each task will be discussed based on the corresponding learning challenges.

All the single-cell machine learning methods covered in this course will be discussed in the context of "real-world" biomedical applications. For each method, we will provide real-world examples and handson datasets for students to learn and practice the methods learned in the class.

Week 1:

Brief Introduction of Single-cell Genomics Data

Expectation: After this lesson, the students should have a better understanding of single-cell technologies (e.g., single-cell RNA-seq, single-cell ATAC-seq). For example, why single-cell RNA-seq is more powerful than its bulk counterparts. The students should also know how single-cell measurements/experiments were performed after the lesson.

- What is single-cell genomics data (e.g., single-cell transcriptomics, single-cell epigenetics)
- Why is it powerful?
- ✤ A few application examples of single-cell genomics

Required reading: <u>https://www.nature.com/articles/nrg.2015.16</u>

Week2:

Comparison of Different Single-cell Sequencing Platforms

Expectation: the students should understand the pros and cons of different single-cell genomics platforms and know how to choose the best single-cell platform for their specific application scenarios. The students are also expected to design single-cell experiments (e.g., # of cells, # of reads) and estimate the approximate cost after this lesson.

- 10x genomics/BD/C1 Fluidigm single-cell platforms
- UMI-based vs. full-length single-cell sequencing
- Cost per cell vs. how many cells needed

Required reading:

1)https://www.nature.com/articles/s41587-020-00748-9

2)https://www.ingentaconnect.com/content/ben/cg/2020/00000021/00000008/art00007

3)https://satijalab.org/costpercell/

4) https://satijalab.org/howmanycells/

Week3:

Quantify Single-cell Gene Expression from Single-cell RNA-seq

Expectation: The students should be able to quantify the single-cell expression from the single-cell RNAseq with cellranger or other established pipelines after the lesson.

- 10x genomics (UMI-based) cellranger pipeline (e.g., quality control, reads mapping, counting reads)
- SMART-seq2 (full-length) gene expression quantification
- A real-world single-cell expression quantification example (hands-on tutorial)

Required reading:

1) https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cellranger

2) https://pubmed.ncbi.nlm.nih.gov/24385147/

Week4:

Standard Single-cell RNA-seq Analysis Pipeline Overview

Expectation: After this lesson, the students should be able to perform standard single-cell RNA-seq data analysis (e.g., identify different cell populations) using established pipelines such as Scanpy or Seurat.

- Single-cell RNA-seq data preprocessing
- Dimensionality reduction
- Clustering cells
- ✤ Annotating cell clusters with cell type or sub-type information.
- Trajectory and pseudotime inference
- Cell-cell interaction inference

Required reading:

1)Scanpy pipeline: https://scanpy.readthedocs.io/en/stable/

- 2) <u>https://www.kallistobus.tools/tutorials/scrna-seq_intro/python/scrna-seq_intro/</u>
- 3) https://www.embopress.org/doi/full/10.15252/msb.20188746

Week5:

Classical Single-cell Dimensionality Reduction and Visualization methods

Expectation: the students should be able to reduce the high-dimensional single-cell data with classical pipelines after this lesson.

- Principal component analysis (PCA)
- ✤ t-SNE
- Uniform Manifold Approximation and Projection

Required reading:

1) https://www.frontiersin.org/articles/10.3389/fgene.2021.646936/full

- 2) <u>http://www.cs.toronto.edu/~hinton/absps/tsne.pdf</u>
- 3) https://arxiv.org/pdf/1802.03426.pdf

Week6:

Single-cell data Denoising/Imputation/Augmentation

Expectation: the students should be able to perform single-cell data denoising, imputation, and augmentation with existing methods after the lesson. The students are also expected to learn how to develop computational models for similar tasks.

- Single-cell denoising methods (e.g., DCA)
- Single-cell data imputation methods (e.g., MAGIC)

Required reading:

1) https://www.nature.com/articles/s41467-018-07931-2

2) https://doi.org/10.1016/j.cell.2018.05.061

Week7: Classical Methods for Single-cell Data Clustering

Expectation: The students should understand the most commonly used classical methods for clustering single-cell data. The students are also expected to understand the limitations of conventional single-cell clustering methods and think of strategies to improve the clustering performance.

- K-means/Spectral clustering
- Louvian/Leiden clustering methods
- Monocle/Seurat/Scanpy

Required reading:

1) <u>https://www.nature.com/articles/s41598-019-41695-z/</u>

- 2) https://cole-trapnell-lab.github.io/monocle3/docs/clustering/
- 3) https://cole-trapnell-lab.github.io/monocle3/docs/clustering/
- 4) https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html

Week8:

Deep-learning based Methods for Single-cell Data Dimension Reduction and Clustering

Expectation: The student should understand and use the deep learning (e.g., Variational autoencoder - VAE) based methods for dimension reduction and clustering of the single-cell data

- Auto-encoder (AE)/Variational auto-encoder (VAE) basics
- Example deep learning pipeline -scVI
- Example deep learning pipeline- UNIFAN

Required reading:

- 1) <u>https://jaan.io/what-is-variational-autoencoder-vae-tutorial/</u>
- 2) https://www.nature.com/articles/s41592-018-0229-2
- 3) https://pubmed.ncbi.nlm.nih.gov/36036832/

Week9:

In class mid-term, no lecture will be given.

Week10:

Computational Methods for Cell-type Annotation of Single-cell clusters

Expectation: the students should be able to annotate single-cell clusters with cell type after this lesson.

- Cell-type annotation based on biomarkers
- Cell-type annotation based on cell type label transfer
- Comprehensive cell-type annotation platforms such as Cellar

Required reading:

- 1) http://bioconductor.org/books/3.13/OSCA.basic/cell-type-annotation.html
- 2) https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1795-z
- 3) <u>https://www.nature.com/articles/s41467-022-29744-0</u>

Week11:

Single-cell Trajectory Inference Methods

Expectation: The students should be able to perform cellular trajectory analysis on the single-cell data after the lesson.

- Probabilistic graphical models for trajectory inference (scdiff/cshmm)
- Monocle3/Seurat for trajectory inference
- Other deep neural networks based trajectory inference

Required reading:

1) <u>https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5848617/</u>

2) https://www.nature.com/articles/s41586-019-0969-x

Week12:

Computational Models for Cell-cell Interaction Inference

Expectation: The students should be able to perform cell-cell interaction analysis with existing pipelines after the lesson. The students should also understand the limitations of existing cell-cell interaction inference pipelines and potential strategies for developing improved methods.

- Cell-cell interaction inference by cellphonedb
- Cell-cell interaction inference by cellchat
- Cell-cell interaction inference by Trasig
- Cell-cell interaction inference with single-cell spatial data

Required reading:

- 1) https://www.nature.com/articles/s41596-020-0292-x
- 2) https://www.nature.com/articles/s41467-021-21246-9
- 3) <u>https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02629-7</u>

Week13:

Multimodal Methods for Single-cell Data Analytics

- Multimodal methods that integrate single-cell omics data for learning joint embeddings and subsequent analyses
- Computational methods that transfer the knowledge from one single-cell modality to another

Required reading:

- 1) https://www.nature.com/articles/s41592-021-01264-7
- 2) https://www.sciencedirect.com/science/article/pii/S0092867421005833
- 3) https://www.nature.com/articles/s41467-021-22197-x

Week	Date	Description	Course materials	Assignments
				due
1	Aug.30	Brief Introduction	https://www.nature.com/articles/nrg.2015.16	Sept.6
		of Single-cell		
		Genomics		
		Data		
2	Sept.6	Comparison of		Sept.13
	•	Different	1) <u>https://www.nature.com/articles/s41587-020-00748-9</u> 2) <u>https://satijalab.org/costpercell/</u>	•
		Single-cell	3) https://satijalab.org/howmanycells/	
		Sequencing		
	6 1 4 9	Platforms	1) https://www.10vgapamics.com/wwpact/circle.coll.gopa_ovprossion	C
3	Sept.13	Quantify Single-cell	1)https://www.10xgenomics.com/support/single-cell-gene-expression 2) https://pubmed.ncbi.nlm.nih.gov/24385147/	Sept.20
		Gene		
		Expression		
		from Single-		
		cell RNA-seq		
4	Sept.20	Standard	 <u>https://scanpy.readthedocs.io/en/stable/</u> <u>https://www.kallistobus.tools/tutorials/scrna-</u> 	Sept.27
		Single-cell RNA-seq	seq_intro/python/scrna-seq_intro/ 3)https://www.embopress.org/doi/full/10.15252/msb.20188746	
		Analysis	3) <u>III(bs.//www.embobress.org/doi/fuil/10.15252/IIIsb.20188746</u>	
		Pipeline		
		Overview		
5	Sept.27	Classical	 https://www.frontiersin.org/articles/10.3389/fgene.2021.646936/full http://www.cs.toronto.edu/~hinton/absps/tsne.pdf 	Oct.4
		Single-cell	3) <u>https://arxiv.org/pdf/1802.03426.pdf</u>	
		Dimensionality Reduction and		
		Visualization		
		methods		
6	Oct.4	Single-cell	1) <u>https://www.nature.com/articles/s41467-018-07931-2</u> 2) https://doi.org/10.1016/j.cell.2018.05.061	Oct.11
		data		
		denoising,		
		imputation,		
		and		
		argumentation		
7	Oct.11	No class	Fall reading back	
8	Oct.18	Mid-term	Mid-term Exam (2 hrs) 1) https://www.nature.com/articles/s41598-019-41695-z/	Nov 1st
9	Oct.25	Classical Methods for	2) https://cole-trapnell-lab.github.io/monocle3/docs/clustering/	Nov.1st
		Single-cell	 3) <u>https://cole-trapnell-lab.github.io/monocle3/docs/clustering/</u> 4) <u>https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html</u> 	
		Data		
		Clustering		
10	Nov.1st		 <u>https://jaan.io/what-is-variational-autoencoder-vae-tutorial/</u> <u>https://www.nature.com/articles/s41592-018-0229-2</u> 	Nov. 8
			3) https://pubmed.ncbi.nlm.nih.gov/36036832/	

11	Nov.8	Computational Methods for Cell-type Annotation of Single-cell clusters	 http://bioconductor.org/books/3.13/OSCA.basic/cell-type- annotation.html https://genomebiology.biomedcentral.com/articles/10.1186/s13059- 019-1795-z https://www.nature.com/articles/s41467-022-29744-0 	Nov.15
12	Nov.15	Single-cell Trajectory Inference Methods	 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5848617/ https://www.nature.com/articles/s41586-019-0969-x 	Nov.22
13	Nov.22	Computational Models for Cell-cell Interaction Inference	 <u>https://www.nature.com/articles/s41596-020-0292-x</u> <u>https://www.nature.com/articles/s41467-021-21246-9</u> <u>https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02629-7</u> 	Nov.29
13	Nov.29	Multimodal Methods for Single-cell Data Analytics	1) <u>https://www.nature.com/articles/s41592-021-01264-7</u> 2) <u>https://www.sciencedirect.com/science/article/pii/S0092867421005833</u> 3) <u>https://www.nature.com/articles/s41467-021-22197-x</u>	Dec.6
14			Final Course project code and report Due	Dec.14

Evaluation

As undergraduate students often have heavier course loads and lack sufficient skillsets or background knowledge required for specific single-cell data analytics, particularly those analyses by advanced machine learning methods, the evaluation for undergraduates will differ from their graduate counterparts. Namely, the undergraduate and graduate students will be evaluated and scored separately to ensure equality and inclusion in the course. The detailed evaluation strategy is described below.

1) Midterm:

The midterm will be a ~2 hours in-person test on the theoretical basics and application of computational models for single-cell RNA-seq data analytics. Example questions include single choice, multi-choice, designing genomics pipelines, and computational models to address example single-cell data analytics needs. (30%)

2) Homework assignments: Students will be given weekly homework assignments (on the covered topics). Given the relatively heavier course load, the homework assignments for undergraduate students will be less intensive (in terms of both quantity and difficulty) than their graduate counterparts. For instance, the number of questions/problems in the assignment for undergraduate students will be 40% of that for graduate students.

3) Final course project: The students will be asked to develop a machine learning model for a specific single-cell data analysis request (e.g., dimensionality reduction, clustering, cell type annotation, trajectory inference, cell-cell interaction, multi-omics data integration). The course project report will be a thesis-like summary with 3500-5000 words (7-10 pages with a font size of 12, single line spacing), which should include introduction, method, result, and discussion sections. Undergraduate students can finish the project in groups, with a maximum of 2 undergraduates in each group.

Exam descriptions

A mid-term example will be held in person on week 8. It will be on the theoretical basics and application of computational models for single-cell RNA-seq data analytics from the topics covered before week 9. Example questions include single choice, multi-choice, designing genomics pipelines, and computational models to address single-cell data analytics needs. The students must also finish a final course project (Due on week 13, Apr.11). The final project will develop and apply a computational pipeline for specific single-cell data analytics, including dimension reduction, clustering, trajectory inference, data imputation, gene regulatory network reconstruction, and cell-cell interaction inference.

Assignment descriptions

A homework assignment will be given after each lecture and is due on the date of the following lecture. The assignment will be on the single-cell data analysis topic covered in the lecture. The students are expected to submit the code in the Jupyter notebook electronically to the TA via email. The assignments for undergraduate students will be different from their graduate counterparts.

Means of evaluation organizer

Name of assignment or exam	Due date	% of final grade
Mid-term	Week 8	30%
Homework assignment	Weekly	30%
Final course project	Week14	40%

McGill policy statements

Language of submission

"In accord with McGill University's <u>Charter of Students' Rights</u>, students in this course have the right to submit in English or in French written work that is to be graded. This does not apply to courses in which acquiring proficiency in a language is one of the objectives." (Approved by Senate on 21 January 2009)

« Conformément à la Charte des droits de l'étudiant de l'Université McGill, chaque étudiant a le droit

de soumettre en français ou en anglais tout travail écrit devant être noté, sauf dans le cas des cours dont l'un des objets est la maîtrise d'une langue. » (Énoncé approuvé par le Sénat le 21 janvier 2009)

Academic integrity

"McGill University values academic integrity. Therefore, all students must understand the meaning and consequences of cheating, plagiarism and other academic offences under the <u>Code of Student</u> <u>Conduct and Disciplinary Procedures</u>" (Approved by Senate on 29 January 2003) (See <u>McGill's guide</u> <u>to academic honesty</u> for more information).

« L'université McGill attache une haute importance à l'honnêteté académique. Il incombe par conséquent à tous les étudiants de comprendre ce que l'on entend par tricherie, plagiat et autres infractions académiques, ainsi que les conséquences que peuvent avoir de telles actions, selon <u>le</u> <u>Code de conduite de l'étudiant et procédures disciplinaires</u> » (Énoncé approuvé par le Sénat le 29 janvier 2003) (pour de plus amples renseignements, veuillez consulter le <u>guide pour l'honnêteté</u> <u>académique de McGill.</u>)

Additional statements

- "The <u>University Student Assessment Policy</u> exists to ensure fair and equitable academic assessment for all students and to protect students from excessive workloads. Students and instructors are encouraged to review this Policy."
- "© Instructor-generated course materials (e.g., handouts, notes, summaries, exam questions) are protected by law and may not be copied or distributed in any form or in any medium without explicit permission of the instructor. Note that copyright infringements can be subject to follow-up by the University under the Code of Student Conduct and Disciplinary Procedures."
- "I will notify you if part of a class is being recorded. By remaining in classes that are recorded, you agree to the recording, and you understand that your image, voice, and name may be disclosed to classmates. You also understand that recordings will be made available in myCourses to students registered in the course. Please consult me if you have concerns about privacy, and we can discuss possible measures that can be taken."
- EDI statements. Underrepresented minority students from an educational background different from the Canadian systems and thus may have difficulty understanding the lectures will be provided extra office hours (2hrs) every two weeks.
- "The University is committed to maintaining teaching and learning spaces that are respectful and inclusive for all. To this end, offensive, violent, or harmful language arising in course contexts may be cause for disciplinary action."

- "As the instructor of this course, I endeavor to provide an inclusive learning environment. However, if you experience barriers to learning in this course, do not hesitate to discuss them with me and/or the <u>Office for Students with Disabilities</u>."
- "Many students may face mental health challenges that can impact not only their academic success but also their ability to thrive in our campus community. Please reach out for support when you need it; many <u>resources</u> are available on-campus, off-campus, and online."
- McGill University is on land which long served as a site of meeting and exchange amongst Indigenous peoples, including the Haudenosaunee and Anishinabeg nations. We acknowledge and thank the diverse Indigenous people whose footsteps have marked this territory on which peoples of the world now gather.
- "Content warning: Please be aware that some of the content in this course contains language or images that may be difficult for some students. It is included in this course because it directly relates to [list relevant learning outcome]."
- "If you have difficulty affording food or if you lack a safe and stable place to live, and believe that these circumstances may affect your performance in this course, I encourage you to contact the <u>Dean of Students</u>, who can connect you with support services. If you feel comfortable doing so, please let me know as well so we can discuss how I can best support your learning." [adapted from <u>Goldrick-Rab</u>, 2017]
- "<u>End-of-course evaluations</u> are one of the ways that McGill works towards maintaining and improving the quality of courses and the student's learning experience. You will be notified by email when the evaluations are available. Please note that a minimum number of responses must be received for results to be available to students."
- "In the event of extraordinary circumstances beyond the University's control, the content and/or evaluation scheme in this course is subject to change."
- McGill has policies on sustainability, paper use, and other initiatives to promote a culture of sustainability at McGill. See the <u>Office of Sustainability</u>.
- Guidelines for the use of mobile computing and communications (MC2) devices in classes at McGill have been approved by the APC. Consult the <u>Guidelines</u> for a range of sample wording that may be used or adapted by instructors on their course outlines.]