Name:

NetID:

Discussion Section:

Course Heading:

Keep your answers concise and to the point.

Long responses won't earn extra credit.

NetID: Score:

1. Multiple Selection

(a) Select all fast alignment methods covered in the lecture: (5pt)

(A) FASTQ
(B) BLAT
(C) BWA
(D) STAR
(E) SVM

BCD

(b) Select all the types of factors that could have an effect on the macrophenotype of a person: (5pt)

(A) Environment
(B) Endophenotype
(C) Wearable data
(D) Genotype

ABD

(c) In the Watts-Strogatz (WS) network generation model, what value should we set for the rewiring parameter to get a fully random network? (4pt)

(A) $p = 0$
(B) $0 < p < 1$
(C) $p = 1$

C

(d) For the following truth table, when doing image segmentation, what is the Dice Index of this model? (4pt)

		Predicted	
		Yes	No
Real	Yes	1000	500
	No	500	2000

(A) 2/3
(B) 1/3

(C) 1/4
(D) 1/2

A

(e) Select all FALSE statements about single-cell RNA-seq data analysis: (5pt)

(A) Popular data visualization tools t-SNE and UMAP both aim to preserve local structures (e.g., distances) between cells when projecting high-dimensional data into a low-dimensional space.
(B) Droplet-based single-cell RNA-sequencing technologies isolate cells into droplets and sequence each droplet individually.
(C) Homogeneous cell populations are generally better studied using single-cell RNA-sequencing than bulk RNA-sequencing.
(D) Pseudotime in trajectory analysis corresponds to the actual timepoints at which samples are collected during a biological process.

BCD

(f) Select all FALSE statements about biomedical data privacy: (5pt)

(A) Because biomedical data are inherently sensitive, the only way to protect the privacy of human subjects is to avoid sharing any data.
(B) An anonymous person's genotype profile may be re-identified through their relatives.
(C) The noisy nature of molecular omics data (e.g., gene expression levels) provides full protection against attempts to extract private information.
(D) Differential privacy mechanisms protect privacy by introducing noise into released statistics.

AC

2. Choose True/False for the following statements: (20pt)

(a) [False / True] Local reassembly is not a way to identify structural variants in genome sequencing.

(b) [False / True] Among the population sampled in the 1000 Genomes project, there are more common variants than rare variants.

(c) [False / True] The sum of the true positive rate and the false positive rate of a supervised model should always be 1.

(d) [False / True] In SVD, if the formula can be described as $A = USV^T$, then the sum of all elements in the matrix S is negative.

(e) [False / True] The characteristic of the small world network is that it has a large clustering coefficient and a small characteristic path length.

(f) [False / True] Backward propagation is used to calculate the gradient of the loss function with respect to each weight in a neural network model.

(g) [False / True] The reparameterization trick solves the problem of back propagation of VAE by moving randomness out of the network.

(h) [False / True] LSTM and attention mechanisms are brought forward to solve the long-term dependence problem of RNNs.

(i) [False / True] A transformer model does not need to have both encoders and decoders.

(j) [False / True] GNN is not only good for regression and classification on graphical data, but also could generate new biomolecules if represented properly.

FFFFT TTTTT

3. Fill in the blanks:

(a) ___Peak calling___ is the process to find the enriched segments of DNA and identify potential transcription factor binding sites when analyzing ChIP-seq data. (2pt)

(b) Depending on the nature of output labels (whether they are categorical or quantitative), supervised machine learning models can be divided into __Regression__ models and __Classification__ models. (4pt)

(c) The x-axis of an ROC plot represents _____D_____ of the model, while the y-axis represents _____A_____. (4pt, use the following option to answer)

(A) sensitivity (B) 1-sensitivity (C) specificity (D) 1-specificity

(d) Suppose you model a small molecule as a graph, with each atom as a node, and each chemical bond as an edge. If you want to predict the properties of this molecule using GNN, then you are performing __graph__-level tasks with this GNN. (2pt)

(e) Diffusion models utilize a ___A____ forward, ___C____noising process and learn to reverse it, whereas flow-matching models learn a __D___, ___B___ velocity vector field to transport noise to data. (8pt, use the following option to answer)

(A) stochastic (B) continuous (C) step-by-step (D) deterministic

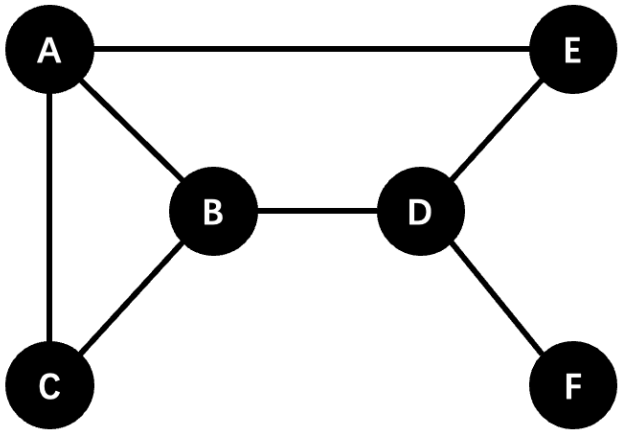
4. Link each unsupervised learning method with its category: (8pt)

NetID: Score:

Clustering Method	Category
K-means	Connectivity-based Methods
tSNE	Centroid-based Methods
Hierarchical clustering	Density-based Methods
DBSCAN	Distribution-based Methods

1 - 2
2 - 4
3 - 1
4 - 3 (each wrong answer -2 pts)

5. For the following graph, answer the questions: (10pt)



(a) The degree of node E?

2

(b) The clustering coefficient of node B?

1/3

(c) The clustering coefficient of node D?

0

(d) The clustering coefficient of node C?

1

(e) The shortest path length between node C and node E?

2

Each wrong answer -2 pts

6. Given the input matrix and kernel, set the stride to be 1 and no paddings for convolution. Then, for the output matrix after convolution, apply max pooling with a 2x2 filter and a stride of 1. (14pt)

Input matrix:

0	6	4	0
4	3	4	6
9	0	4	9
1	2	1	5

Kernel:

1	-1
1	-1

(a) Fill the following empty matrix with the output of the convolution:

Answer:

NetID: Score:

-5	1	2
10	-5	-7
8	-3	-9

(b) Fill the following empty matrix with the output of the max pooling:

Answer:

10	2
10	-3

One wrong answer -1 pt