Each question worth 1 point

1. Students offered many intelligent answers. We were looking for general/abstract answers (e.g. simplify a process to understand it; or predict new data to replace or guide experiments), but we also accepted concrete examples from class.
2. Students offered many intelligent answers, such as the tradeoffs between accuracy and comprehensibility, or the influence of unknown factors.
3. Students offered many intelligent answers, such as predicting which drug will bind to a molecular target.
4. Mention a processing purpose such as denoising or compressions; and an interpretative purpose, such as visualization or understand most important features.

For full credit: ≥1 processing purpose [orange] AND ≥1 interpretive purpose [green]

0.75 credit for mentioning an interpretive purpose only

0.5 credit for mentioning a processing purpose only

1. Supervised – predict labels using labels. Unsupervised – cluster points without using labels. Use supervised when have labels for some, trying to predict labels for others. Use unsupervised when no/few labels present, or when goal is to understand clusters/structure of the data.

0.33 points earned for mentioning supervised uses labels and unsupervised does not use labels. 0.33 points for saying that supervised is when you have labeled data. 0.33 points for saying you unsupervised when trying to understand natural clusters/structure.

1. Apply nonlinear transformation to data, often to a higher dimensional space (). [1/3] Advantage: find nonlinear clusters. [1/3]; Disadvantage: overfitting OR difficult to find right transformation [1/3]
2. Good rule splits a parent node into two children each more pure than the parent. [== Increase homogeneity within clusters == descrease inhomogeneity within clusters] [1 point]. Not too specific [0.25 recovery points, if main educational objective not mentioned]
3. Iterative process == repeat E-M steps [0.2] (until convergence [0.1 points]). Correct description of E-step: calculate likelihood of data [0.3] Correct description of M-step [0.3]. Start with initialization step [0.1 points].
4. First PCA is the vector from a set of orthogonal vectors which points in the same direction as the maximum variance of the data that can be explained by one of these orthogonal vectors.
5. Rugged – protein folding energy landscape with many, steep local minima. [2/3] Misfolded protein diseases more likely with rugged landscapes [1/3]
6. B [random walk] < C [self-avoiding random walk] < A [fully extended chain == A > C > B [No partial credit]
7. Red: no steric clash / higher probability / all residues; Yellow: allowable steric clash / lower probability / smaller residues. [no partial credit]
8. C 1 points; B 0.5 points; B+C 0.5 points
9. Repulsion arises from electrostatic repulsions between electron clouds [0.2] when clouds overlaps (== r < sigma) [0.2]. Attraction arises from van der Waals forces [0.2]. r = distance between particles [0.2]; sigma = fixed distance at which inter-particle potential is zero [0.2]. Some students may have interpreted the question differently, and are encouraged to explain why they interpreted the question the way they did.
10. A [1 point]. Recovery credit: Eligible reads must originate 1-89 nucleotides 5’ of the breakpoint to include ≥ 1 base both upstream and downstream of the breakpoint. [0.1] It is important to identify which fraction of bases have a read whose 5’ end overlaps that base. [0.1] With 30 reads per typical site and 90 bp-long reads, one-third of bases have a read whose 5’ end overlaps that base. [0.1] This means that we expect that one third of 89 ~ 30 reads will overlap the site of the breakpoint [0.1]. But because the deletion is heterozygous, only half of these reads will arise from the relevant chromosome. [0.1]
11. B-only [1 point]; B+C+explanation [0.5 points]; B+C without explantion [0 points] “None of the above: Long reads!” [0.25 points]. FYI: Because of the long and repetitive nature of this sequence and insertion, split-reads cannot detect the insertion. [0.1 points] Read-depth will not be able to pinpoint the breakpoint [0.1 points], but with enough coverage it could detect increased read counts pertaining to the inserted sequence [0.1]. Paired-end reads can sometimes be more successful than split reads in resolving SVs in repetitive regions [0.1 points] but in this case the insertion is too large and the sequence is too repetitive [0.1 points].