# BE 275 Midterm, Fall 2021

## Question 1 (10 points)

Assuming a constant and random fixed rate, the number of neuron firings for a given neuron within a set period of time follows a Poisson distribution:

$$p(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where $k$ is the number of firings, and $\lambda$ is the average number of firings in that given period. You collect recordings of firing numbers across a population of neurons using a fluorescent reporter and a microscope. Each neuron has the same firing rate. Each microscopy field has roughly 50 neurons with the reporter, and you image 20 different regions.

a) Provide an expression for the variance in the number of events across neurons (using the distribution above). You do not need to simplify or solve.

b) What are at least three things you can say about the distribution of average firing numbers across the fields of view in your experiment (relative to the distribution for single neurons)?

c) You expect that there is actually a population of neurons not firing at all. Meanwhile, the active neurons fire at a rate of once per second. What is the probability of a neuron having no firing events, given it is active? How about if it is inactive/non-firing?

d) You expect that 5% of your neuron population is inactive/non-firing. For how many seconds do you need to collect data to be 99% sure that a neuron with no events is inactive?

## Question 2

Day *et al*, *Canc Res*, 2021 build a model for the signaling determinants of glioblastoma response to SHP2 inhibition. To do so, they measure a variety of signaling proteins over time. Their results are shown in Figure 2 on the next page.

a) You decide to analyze the molecular measurements (e.g., ATF2, JNK) by PCA as well. Would you expect the scores and loadings to be the same, similar, or completely different? Explain.

b) Describe the effect of PC2 on cell phenotype (e.g., Cell death, S, $G_2$). How is PC2 related to SHP2 siRNA treatment?

c) Let's say the authors decide to use a 4-component model instead of a 3-component one. How would the plots in (C) and (F) change?

d) How do R2X, R2Y, and Q2Y vary with the number of components? How would R2X look relative to R2X using PCA?

e) You reproduce Figure 2C, but the points look mirrored along the y-axis compared to the figure (all the negative PC1 points are at positive PC1 positions). What has happened? Are your results the same as theirs? Would any other plots also change?

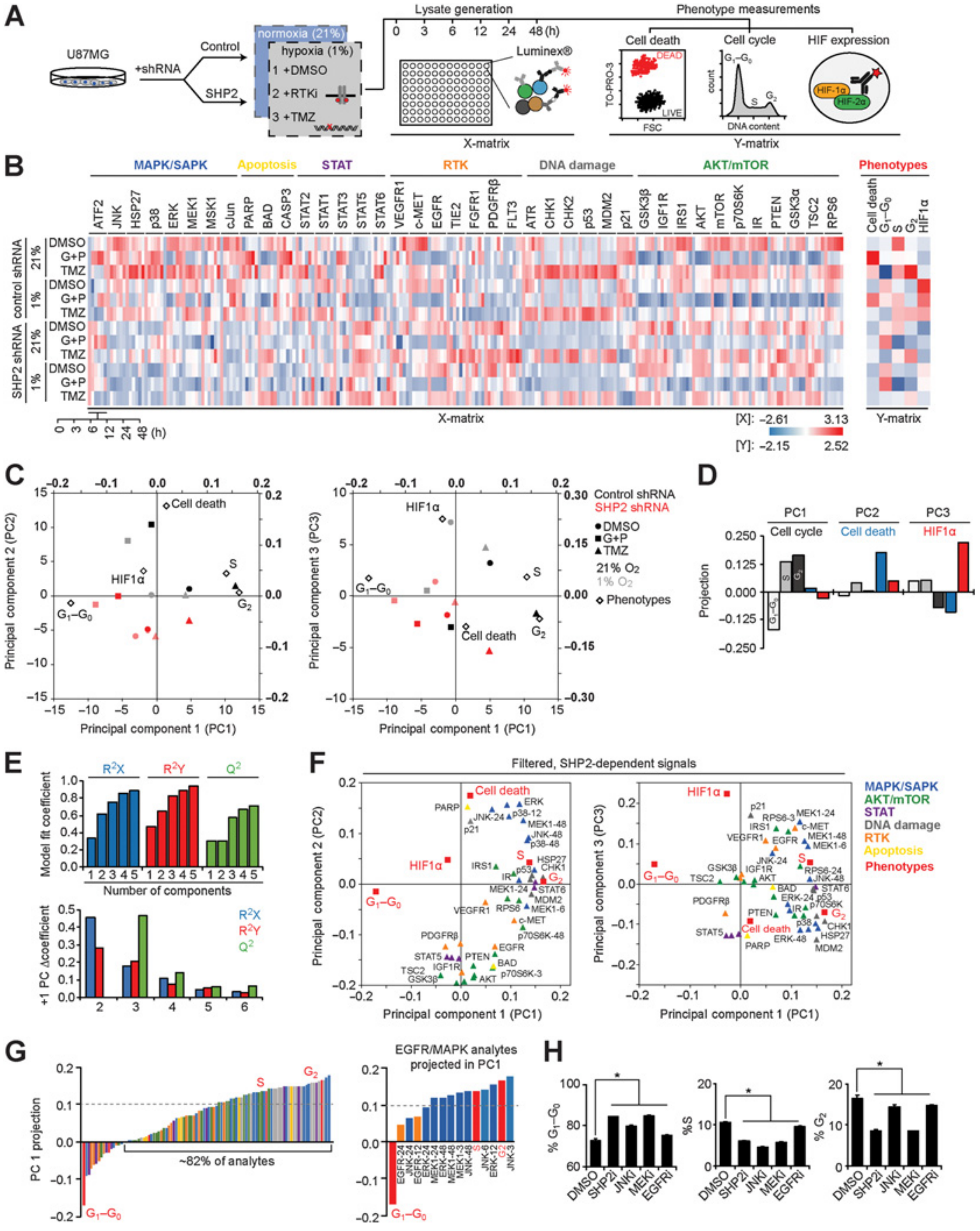
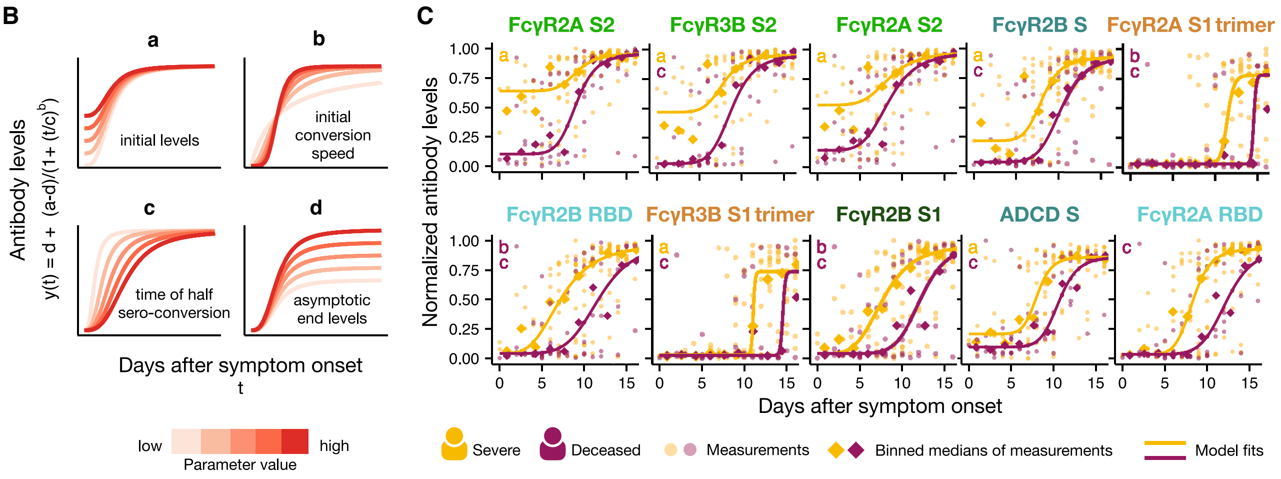


**Figure 2.**

A data-driven model predicts SHP2-regulated signaling governing GBM cell response to therapy. **A,** U87MG cells expressing control or SHP2 shRNA were cultured in 21% or 1% oxygen and treated with gefitinib (G) and PHA665752 (P), temozolomide (TMZ), or DMSO. Cells were lysed 0, 3, 6, 12, 24, and 48 hours after treatment, and lysates were analyzed via Luminex. **B,** Mean-centered, variance-scaled signaling and phenotypic measurements are represented by a heat map, with Luminex kits indicated. Supplementary Table S2 provides analyte posttranslational modifications. **C,** Bi-plots of scores (bold numbered axes; top right) for conditions and loadings (bottom left) of Y-matrix phenotypes for PC2 or PC3 versus PC1. **D,** Phenotype projections into PCs 1–3 are plotted. **E,** Model fit ($R^2X$, $R^2Y$) and predictive ($Q^2$) coefficients are plotted versus PC. Coefficient changes (Δ) with additional components are shown. **F,** Loadings are plotted of signals (triangles, color coded by Luminex kit) and phenotypes (red squares) for the three-component model, displaying only SHP2-regulated analytes. Times are indicated where two or more time points projected in opposite directions for an analyte. Loadings are listed in Supplementary Table S5. **G,** PC1 signal and phenotype projections are plotted, smallest to largest. Strongly projecting EGFR (orange) and MAPK (blue) analytes, and cell-cycle phases (red) are highlighted. **H,** U87MG cells were treated for 48 hours with 10 μmol/L SHP099 (SHP2i), 20 μmol/L SP600125 (JNKi), 5 μmol/L CI-1040 (MEKi), 10 μmol/L gefitinib (EGFRi), or DMSO, and cell-cycle distribution was analyzed. Error bars indicate mean ± SEM of three replicates; *, $P < 0.05$ for the indicated comparisons from Tukey *post hoc* testing following one-way ANOVA.

## Question 3

Zohar *et al*, *Cell*, 2020 analyzes the dynamics of antibody response during SARS-CoV-2 infection. Each of the plots below depicts a certain measure of antibody quality/quantity, separated by whether subjects survived after being admitted to the ICU. They summarize these dynamics by fitting the longitudinal data to a logistic curve model using non-linear least squares.



a) What are two algorithmic challenges (difficulties in arriving at a solution) that arise when using non-linear least squares but do not in the linear case (OLS)?

b) What do you need to provide to a non-linear least squares method to fit the model?

c) You reimplement this analysis, and want to independently check that the fitting process successfully fit. What is something you could check to verify the solution converged?

d) You want to test whether the b parameter is significantly different between severe and deceased outcomes. How can you do this?

e) You wish to cross-validate your model to verify that it is not over-fit to the data. In the study, each subject was measured at 2–3 timepoints. How should you decide which samples to leave out in each fold? Explain your reasoning.

## Question 4

A mammogram is a diagnostic test for breast cancer with a sensitivity and specificity of roughly 97% and 64.5%, respectively. A completely healthy, asymptomatic woman shows a positive test and is recommended for a biopsy. The incidence of breast cancer in the general population is 1 per 8 women.

a) Write out Bayes' law, and then rewrite it as the probability of the woman having a breast tumor given her positive test.

b) Sensitivity is true positives over all positives, while specificity is true negatives over all negatives. Therefore, the false positive rate is 1 – specificity and the false negative rate is 1 – sensitivity. How many false and true positives are expected in a cohort of 1000 tests?

c) Calculate the probability of the woman having breast cancer given her positive test result.

d) What could we do to further ensure positive tests are giving us true results? (You can't improve the test itself.)

e) A common form of breast biopsy has a sensitivity and specificity of 91% and 98%, respectively. What the chance **the biopsy is positive**, given the information from above?

## Question 5

a) What is cross-validation and what does it estimate?

b) Outline the steps for performing cross-validation in order, including normalizing your data by z-scoring.

c) How do predictions from cross-validation systematically differ from the actual model?

d) Why are multiple folds necessary? What are two trade-offs when selecting a number of folds?

e) What does bootstrapping provide?

f) Outline the steps for bootstrapping a PLSR model. In the end, what are the resulting outputs of doing this?

## Question 6

You are developing a glucose monitor that automatically dispenses insulin. You develop a *very* simple model of insulin and glucose regulation in the body:

$$\frac{\delta I}{\delta t} = -\gamma I + \theta G$$

$$\frac{\delta G}{\delta t} = \alpha - \beta I G$$

$I$ and $G$ indicate the amount of insulin and glucose, respectively. $\alpha$ indicates some constitutive production of glucose, $\theta$ some production of insulin by the body when glucose is present, $\beta$ the removal of glucose by insulin, and $\gamma$ clearance of insulin.

a) Is there a steady-state amount of insulin and glucose? If so, how much?

b) Can this model ever oscillate? If so, under what parameter values?

c) Sketch out what the function would look like that you would pass to an ODE solver. Describe the inputs and outputs.

d) Say you obtain the solution on the right, plotted in phase space. What does this say about solutions that begin at the red dot? Explain your reasoning.