Bi 1: The Great Ideas of Biology
Homework 7
Due Date: Thursday, May 25, 2023

> "Problems worthy of attack prove their worth by hitting back."
>
> Piet Hein

In the last homework, you worked out a statistical mechanical theory of gene expression regulation. Now, we'll put that theory to the test by analyzing the results of real-world experiments!

1. The Theory-Experiment Dialogue

In class, we described the role of transcription factors as the molecular agents that tune the level of gene expression, either by inhibiting the ability of RNA polymerase to bind the promoter (repressors), or by recruiting the polymerase to the promoter (activators). In your homework, you then turned the cartoon-level model that represents one of these processes into a concrete mathematical form using statistical mechanics. But it is never enough to simply write down a formula; we also need to put it to the test. Our expression for the fold-change in gene expression featured several tunable parameters, including the repressor copy number R, and the binding energy $\Delta\varepsilon_r$ describing the protein-DNA interaction between the repressor and its binding site on the genome. Ideally, our experiment will use these parameters as "knobs" to be tuned in order to assess the agreement of the results with our model. To access these parameters, we'll exploit the tools of modern molecular biology.

We can tune R by mutating a region of the DNA known as a ribosomal binding site (RBS). As the name suggests, this is a sequence downstream of the transcription start site that is recognized and bound by the ribosome. Therefore, once a gene has been transcribed into mRNA, a ribosome will arrive at the RBS in order to initiate its translation into protein. Mutating the RBS allows us to adjust the strength of the interaction between it and the ribosome, altering the probability of binding. In turn, this controls the frequency with which translation is initated, and hence the number of proteins that end up being made.

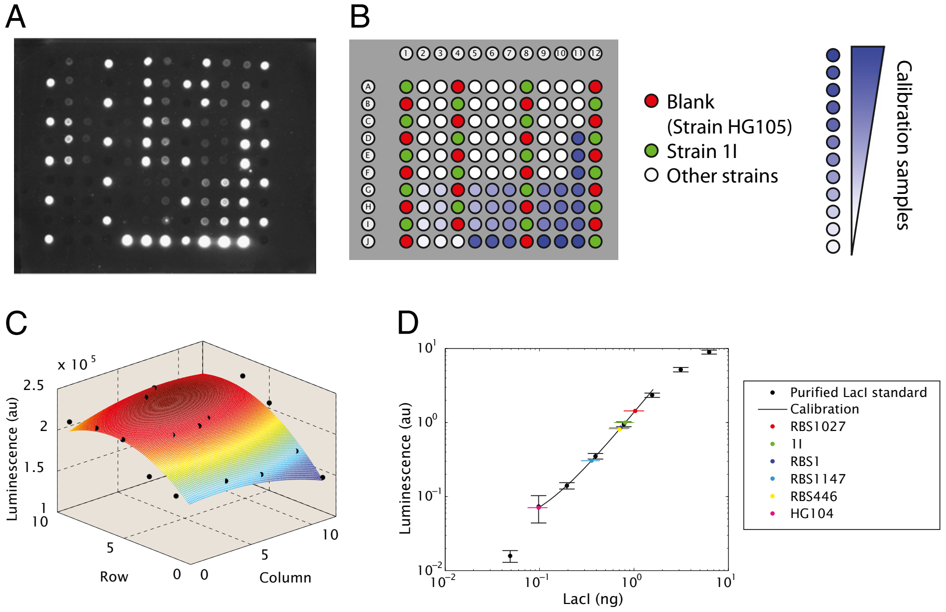


Figure 1: **Results of immunoblot experiments to measure LacI repressor protein copy number.** (A) A typical example of a real immunoblot luminescence image. (B) Legend indicating layout of strains in (A). The interleaved pattern of 1I and blank strains is chosen to help correct for nonuniform illumination, which is a common technical artefact in these experiments. White spots are actual experimental strains of interest. Blue spots contain purified LacI protein at known concentration, mixed with cell cultures of similar concentration to the experiments that lack the LacI gene. This makes it possible to convert brightness to mass of protein, and from mass to copy number using the known mass of LacI. Knowing the number of cells in the experimental strains, we can further convert their brightness to an average number of LacI proteins per cell. (C) The aforementioned uneven illumination profile, which we correct for during data processing. (D) The luminescence vs protein mass for the calibration samples is fit to a low-order polynomial to produce a calibration curve. Then the luminescence of the unknown strains can be converted to protein mass. From Garcia and Phillips (2011).

Tuning the number of repressors is a good start, but we also have to verify how many there are in the cell. To do so, we can use techniques such as fluorescent protein fusions or quantitative Western blots.

Your dataset

In this homework, you will analyze images from experiments used to measure fold-changes in the expression of a fluorescent protein under control of the LacI repressor. When there are many copies of LacI around, the fluorescent protein will be lowly expressed; when there are only a few copies of LacI, the fluorescent protein will be expressed at higher numbers. Your objective is to analyze fluorescence microscopy images of live cells in order to extract fold-change values, then compare your results with those predicted by the theory you developed last week.

The dataset also contains LacI copy number values that were determined by purifying repressor protein up to known quantities as a standard. This standard was then compared to measurements obtained by breaking open millions of bacteria before “fishing” out repressors by using antibodies that bind to them. The repressors bound to antibodies were subsequently modified so that they emit light, and absolute counts were obtained by comparing the amount of light emitted by these repressors to that from the purified standard. Figure 1 gives a broad sense of the data produced by these experiments.

We note that the fluorescence-based approach described here is by no means the only way to measure fold-changes experimentally. Another approach is known as a LacZ assay (Figure 2), which involves tracking the expression of the titular LacZ enzyme. LacZ digests sugars; when fed a particular colorless sugar called ONPG, the resulting products appear yellow in solution. Hence, the color of the mixture is directly proportional to LacZ expression. This system provides another way to observe and tinker with the kind of phenomena we’re discussing. Remember that curating multiple orthogonal, independent, and consistent measurements of a single phenomenon is essential if we wish to prove to ourselves (and the world) that we *truly understand something.*

First, a prediction

The main task in this homework is to test our thermodynamic theory of simple repression from last week by analyzing fluorescence microscopy images. But — like we just mentioned — there are other experimental methods of measuring the same thing. Do they agree?

Recall our theoretical expression for fold-change from last week,

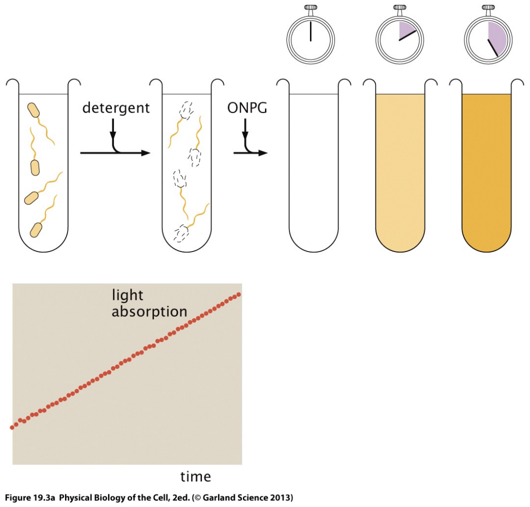

detergent
ONPG
light
absorption
time
Figure 19.3a Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Figure 2: **Schematic of a LacZ assay.** Cells are lysed using detergent, then a sugar (ONPG) is added. LacZ cleaves ONPG (which is colorless) into simpler sugars (which appear yellow in solution). Thus, the initially colorless mixture acquires color as more product accumulates. The slope of this line is proportional to the amount of LacZ enzyme present.

$$\text{fold-change} = \frac{1}{1 + \frac{R}{N_{NS}} e^{-\Delta\varepsilon_r / k_B T}} \tag{1}$$

In Equation (1), N_{NS} is the genome size, and R is the copy number of repressor proteins. Both of these are known; recall from your last homework that $N_{NS} = 4.6 \times 10^6$ for *E. coli*, and we just described the process for determining R. We're measuring fold-change using the proxy measurement of fluorescence intensity, which is essentially proportional (with some caveats to be cleared up shortly).

For now, the point is that the only unknown parameter is $\Delta\varepsilon_r$: the binding energy between LacI and DNA. More precisely, $\Delta\varepsilon_r$ is the *difference* in energy between having a repressor bound to its target site and having it bound to any random sequence in the genome. If we can independently determine $\Delta\varepsilon_r$, we can make a *parameter-free* prediction to which we can compare our microscopy data. Luckily, we have such a method!

Not only can we engineer strains with known R, but we can also change the DNA sequence where the repressor itself binds! These regions are sometimes called *operators*, and different operators have different $\Delta\varepsilon_r$. A list of some different strains and their respective operators can be seen in Table 1. Using

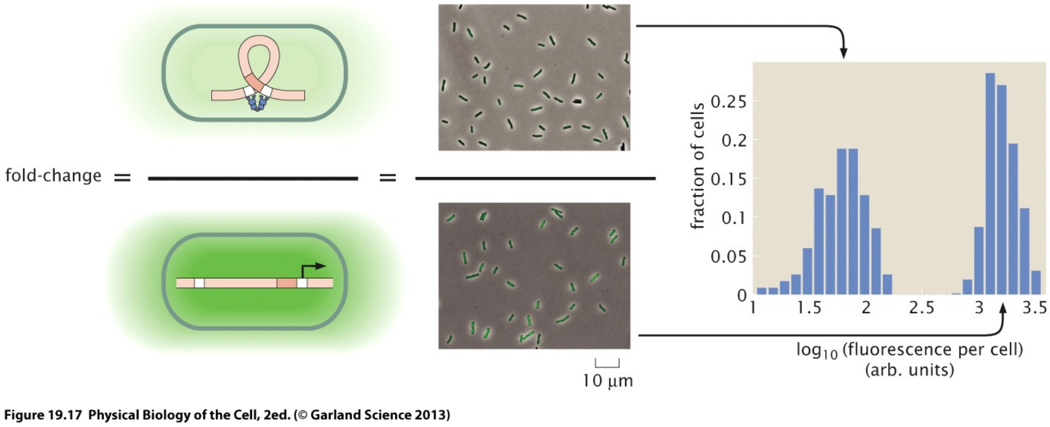

fold-change =
=
fraction of cells
0.25
0.2
0.15
0.1
0.05
0
1
1.5
2
2.5
3
3.5
log$_{10}$ (fluorescence per cell)
(arb. units)
10 µm
Figure 19.17 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Figure 3: **Fold-change in gene expression, in cartoon form.** Recall from the last set that fold-change is a ratio of gene expression levels in cells with and without repressor proteins present. Since the expressed gene is a fluorescent protein, expression levels are proportional to fluorescence intensity. In other words, when working with fold-changes, we need not convert the arbitrary fluorescent units into absolute units of mRNA or protein.

the LacZ assay described above, we can measure fold-changes based on each strain's operator, and use the results to infer $\Delta\varepsilon_r$ for each particular operator sequence.

Question 1a

In preparation for the rest of the set, invert Equation (1) (i.e., solve for $\Delta\varepsilon_r$ as a function of the other variables).

Question 1b

For each strain in Table 1, use your result from the previous problem to calculate $\Delta\varepsilon_r$. Then plot your fold-change predictions for the microscopy experiments (i.e., plot Equation (1) as a function of R for each operator). Show all three curves on one plot, and let R range from 10^0 to 10^3. As a sanity check, be sure to also plot the provided data point for each operator.

Operator	Repressor #	Measured Fold-change
O1	260	2.77×10^{-3}
O2	260	1.24×10^{-2}
O3	260	4.77×10^{-1}

Table 1: Fold-change measurements from lacZ assays for Question 1b.

Enter real data

It's finally time to put our predictions to the test! We'll walk through how to analyze some fluorescence microscopy data from multiple strains of *E. coli*. The strains are identical except for their operator sequence (e.g. their $\Delta\varepsilon_r$) and repressor copy number (R) regulating our gene of interest, which is a fluorescent protein. There are 12 "experimental" strains, for each combination of operator (O1, O2, or O3) and repressor count per cell (22, 60, 124, or 260 molecules). There are also 2 "control" strains per operator: "delta" (which have 0 repressors) and "auto" (in which the fluorescent gene of interest is removed entirely). The auto strains are necessary because cells can display a low level of basal fluorescence even without the protein, a phenomenon called *autofluorescence*.

We assume that the amount of fluorescent protein is proportional to the total fluorescence of the cell. However, autofluorescence throws a wrench into this proportionality. So, to get an accurate measurement, we subtract off the average autofluorescence from the total fluorescence intensity of the cells. Hence our empirical formula for fold-change is

$$\text{fold-change} = \frac{\langle I_{R\neq 0}\rangle - \langle I_{\text{auto}}\rangle}{\langle I_{\text{delta}}\rangle - \langle I_{\text{auto}}\rangle} \tag{2}$$

where each average intensity $\langle I\rangle$ is the average over all cells in all images for that particular strain, and all intensities are for strains with the same operator.

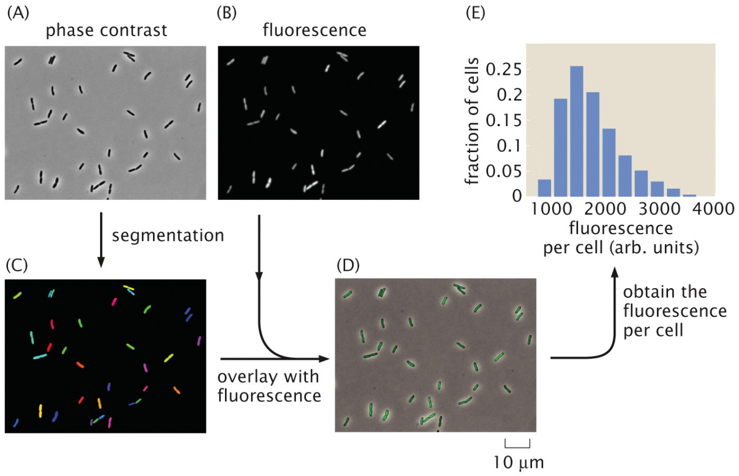


Figure 19.16 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Figure 4: **An image analysis workflow for fluorescence microscopy.** All strains of *E. coli* appear clearly under phase contrast (A), so that image is used to segment cells from the background. The fluorescence image (B) provides a readout of the gene expression in each cell. Using it as a mask (D) allows us to compute the distribution of fluorescence for the entire population of cells (E). Fold-change is calculated from the mean intensities, as shown in Figure 3.

Question 1c

Write a Python function called `segment_im` that takes as input a phase-contrast image, and returns an appropriate segmentation mask (as a binary image). All the concepts you need to do this were covered in the tutorial, but you will have to think carefully about how to put everything together. To get you started, we provide a header for you to follow:

```
def segment_im(im_phase, ip_dist=0.160,
                area_bounds=(1, 4), phase_thresh=0.3,
                show_mask=False):
```

All the arguments should be clear from the tutorial except `show_mask`. If `True`, it should create a figure that displays the segmentation mask overlaid on top of the original image, as demonstrated in the tutorial. If `False` (the default behavior), it should skip this step, increasing speed when processing the over 1000 images in this dataset. You should demonstrate to your own satisfaction that the function works as expected by testing it on at least a few images from a few different strains and visually evaluating the results. As in real science, there are a few images in the dataset with poor illumination that will be impossible to segment. Remember that the identification of single cells (and rejection of cell clusters and other "garbage") need not be flawless, but the error rate does need to be sufficiently low as not to significantly impact the statistical robustness of your results. We are not asking for rigorous testing here, merely common sense.

For your submission, demonstrate the function on the following list of images from the dataset:

```
[‘01_delta_phase_pos_01.tif’,
 ‘02_delta_phase_pos_01.tif’,
 ‘03_auto_phase_pos_01.tif’,
 ‘01_R60_phase_pos_01.tif’]
```

by showing us the segmentation mask overlaid on the original image.

Question 1d

Write a function called `extract_intensities` that takes as input a phase contrast image and its associated fluorescent image. It should use your code from Question 1c to segment the phase image, then use that mask to compute fluorescent intensities for all segmented cells. The function should return a list containing the total fluorescent intensities of all segmented objects.

Demonstrate your function works as expected by plotting histograms of cell intensities for the corresponding fluorescent images from Question 1c. Again, we enclose a header to get you started:

```
def extract_intensities(im_phase, im_fluor):
```

Question 1e

Write a function called `strain_totals` that takes as input strings corresponding to a single repressor and single operator label in the image filenames, and returns a list of cell fluorescent intensities for *all* cells corresponding to the input labels.
This time we offer a header and a pseudocode sketch of the logic:

```
def strain_totals(op, rep):
    glob to get all filenames with op & rep
    for each position:
        read images from a single position
        get list of single-cell intensities
        append single-cell ints into 1 list
    return list of intensities for all cells
```

For instance, `op` could be `‘O1’`, `‘O2’`, or `‘O3’` and `rep` could be any of `‘delta’`, `‘auto’`, `‘R22’`, `‘R60’`, `‘R124’`, or `‘R260’`. The function should construct a filename pattern from the input strings, then use `glob.glob` to grab all images corresponding to the specified strain.[a]

Demonstrate your function works as expected by plotting histograms of cell fluorescent intensities for the O1 auto, O2 auto, O2 delta, and O3 delta strains. All should be fairly Gaussian distributed, with perhaps a few outliers on the high intensity side, especially for the delta strains.

[a] `glob` returns a list of filenames matching an input string with wildcards. For instance, if `op = ‘O1’` and `rep = ‘delta’` then

```
glob.glob(op + ‘_’ + rep + ‘_phase*.tif’)
```

will return a list with all filenames for all phase contrast images for the strain with O1 operator and no repressors. Note the use of string concatenation to avoid writing strings by hand, which will also be useful later.

Even with beautiful code, the next problem will take some time to run (something on the order of a minute, depending on your computer and your exact code implementation). This can make debugging tedious. So, we recommend you test all your functions above thoroughly and ensure they work as expected before moving on!

Question 1f

Building off the code you’ve written so far, compute the mean fluorescent intensity for all provided strains. Then, calculate the fold-change in expression for all repressor copy numbers and operator sequences.
A good idea would be to define some lists like

```
ops = [’O1’, ’O2’, ’O3’]
rep_names = [’R22’, ’R60’, ’R124’, ’R260’]
rep_nums = [22, 60, 124, 260]
```

to help with any loops. A pseudocode sketch might go something like

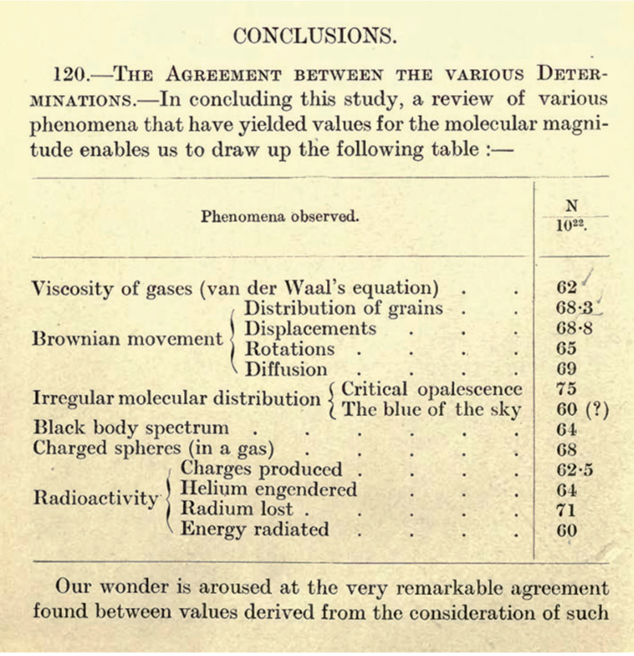
```
predefine storage arrays
loop over operators:
    compute auto & delta means
    loop over repressor #:
        compute mean intensity
        compute fold-change and store it
```

Your code should return the fold-change values in a 2D array. Remember that we need to subtract off the autofluorescence as specified in Equation (2).

Question 1g

Replot your prediction curves from Question 1b, but now overlay all the fold-change results from the microscopy data you just analyzed. Your results should fit the predictions reasonably well, with the exception of a few data points. *Briefly* discuss the possible experimental and/or computational reasons for these outliers.

(Hint: it may be informative to examine the mean fluorescent intensity of each strain, from before you computed fold-changes.)



CONCLUSIONS.

120.—The Agreement between the various Determinations.—In concluding this study, a review of various phenomena that have yielded values for the molecular magnitude enables us to draw up the following table :—

Phenomena observed.		$\frac{N}{10^{22}}$.
Viscosity of gases (van der Waal's equation)		62
Brownian movement	Distribution of grains	68·3
	Displacements	68·8
	Rotations	65
	Diffusion	69
Irregular molecular distribution	Critical opalescence	75
	The blue of the sky	60 (?)
Black body spectrum		64
Charged spheres (in a gas)		68
Radioactivity	Charges produced	62·5
	Helium engendered	64
	Radium lost	71
	Energy radiated	60

Our wonder is aroused at the very remarkable agreement found between values derived from the consideration of such

Figure 5: **Some of the many determinations of Avogadro's number from independent experimental techniques!** From *Atoms*, Jean Perrin, 1913.

Revisiting $\Delta\varepsilon_r$: least-squares regression

You have now seen firsthand that we have a great deal of gene expression data, on a variety of strains, produced via *orthogonal and complementary* methods (single-cell fluorescence and a bulk enzymatic assay). Once again, we note that comparing independent, empirical determinations of a single quantity is an important part of building scientific understanding. An excellent historical example of this is shown in Figure 5, presending Jean Perrin's curation of many efforts to calculate Avogadro's constant using wildly different physical approaches.

You made a prediction in Question 1b for where the data points analyzed in Question 1f should land. The agreement — or lack thereof — between that prediction and the results is some measure of our confidence in the theory's correctness. But we would like something stronger. Since we have measurements of fold-change from microscopy and from lacZ assays, can we compare these two independent determinations of the same quantity against each other to test our theory more decisively? What do we need to know to do so?

Since we are able to measure repressor copy number R independently with other experiments, the only unknown parameter in our theory seems to be $\Delta\varepsilon_r$. In Question 1b, we inferred the binding energies $\Delta\varepsilon_r$ for each operator by considering bulk lacZ assays for a *single* strain for each operator. But we have data from many similar experiments on (otherwise identical) strains with different repressor copy numbers. There was nothing special about the strains we used above. Each of these independent experiments contains information about its respective $\Delta\varepsilon_r$.

There are a variety of valid ways of thinking about this data. One common approach is to use least-squares, like we did in class, to fit all the data at once. In other words, we can pool all the data — from microscopy experiments and lacZ assays alike — and perform least-squares regression on the entire combined dataset to determine $\Delta\varepsilon_r$. The theory is then judged by the goodness of fit of all the data to this theory fit.

Here we propose an (arguably) more satisfying alternative: use the data from *one* independent experiment to determine best fit model parameters, *then* take this as a prediction that the other experiment must agree with if we are to believe the theory. In other words, take all the lacZ data for each operator and do least-squares to infer $\Delta\varepsilon_r$ for each operator. This provides a prediction that the microscopy data should agree with. (This is similar in spirit to many of the cross-validation techniques common in statistics and machine learning. Those methods typically require careful consideration of how to split data into subsets for training, testing, and validation. Here, we can be simplistic and choose a higher bar: that completely different experimental methods for measuring the same quantity should yield comparable results.)

Question 1h

Rather than inferring $\Delta\varepsilon_r$ from a single datum, use the least-squares techniques you learned in the tutorial to fit $\Delta\varepsilon_r$ using all the lacZ assay data for each operator. Download the data in `lacZ_data.csv` from the course website and load it using `pandas`. As in the tutorial, treat all the strains with O1 and any R as a dataset and perform least-squares to infer a $\Delta\varepsilon_r$ for O1. Then repeat this process for O2 and O3. Note that since fold-change is a multiplicative quantity, you should do least-squares fitting on the *log* of the fold-change, rather than fold-change itself. This is because fractional — not absolute — changes are what matter when discussing fold-changes.

How do the $\Delta\varepsilon_r$ values compare to the values you obtained by merely fitting a single point? Comment briefly.

Question 1i

The relationship between bulk lacZ measurements and single-cell microscopy experiments is not one-way. In this problem, reverse your analysis from the preceding question. In other words, use least-squares fitting to infer a binding energy for each operator from the *microscopy* fold-change data, and compare it to the lacZ numbers. How different are the inferred binding energies this time around? Can you explain any discrepancies?

Question 1j

This is it — the grand finale! It's time to plot everything you've calculated so far in one masterful graph.[a]

Plot the fold-change theory curves for each operator using your $\Delta\varepsilon_r$'s from Questions 1h and 1i. Also plot all the lacZ data points from the datafile, and all of the microscopy results from 1f. Now that you have everything in one place, comment on the agreement — or lack thereof — between the two experimental approaches (fluorescence versus LacZ). If there is disagreement, can you hypothesize whether it indicates a fault in the theory or in the measurements?

[a] To keep things visually clear, choose a different color for each operator, a different line style for the two theory curves, and a different dot style for the two experiment types. If you're using `matplotlib`, all of these options can be set with keyword arguments, e.g. `color='r'` will make that command's output red, `marker='v'` will make the datapoints into triangles, and `linestyle='-'` will connect points with a dashed line.