PMLS2/e: Datasets

This document corresponds to the second edition of *Physical models of living systems*. If you want materials related to the first edition, please instead visit www.physics.upenn.edu/biophys/PMLS/index.html .

Here is a list of the datasets mentioned in *PMLS*. Some are in plain text (human viewable, `.txt`, `.csv`), but others are not (MATLAB, `.mat`, Python `.npy` and `.npz`). You can get them all in a single archive here:
http://www.physics.upenn.edu/biophys/PMLS2e/Datasets/PMLS2eData.zip .
In each folder, a plain text description file named `_README.txt` mirrors information in the present document.

To get a file individually, try just clicking its link below (some browsers will then offer to save it to your hard drive). If your browser won't do that (for example, if a `.mat` file appears as a garbled page), try right-click (Windows) or ctrl-click (Mac), which should give you a context menu, including an item that allows you to download the file instead of attempting to display it.

Once you have the file, and if necessary have moved it to a convenient folder, you must next load it into your software. For MATLAB files:

- It may suffice to double-click the file in your computer's Finder.
- Or enter MATLAB and navigate to the folder containing the file, for example, by clicking the folder icon located on the upper left of the main MATLAB window, or using the `cd` command in the command window. Then you can double-click the file in MATLAB's `Current Directory` panel (upper left), or use the `load` command at the MATLAB command line or in a script. After this operation, your workspace will contain some variables with the data.
- Files with names ending in `.csv` or `.txt` are generic comma-separated-variable files. MATLAB can read such files by using the Import Wizard (`File>Import data`). But in most cases, the `.csv` file is just a duplicate of data also given in a `.mat` file. If you use MATLAB, and a `.mat` version exists, it's easier to use this file instead of the `.csv` or `.txt` version.
- Files with names ending in `.tif` are images, which can be read into MATLAB by using the `imread` command or Import Wizard.

For other software:

- Files with names ending in `.xls` or `.xlsx` are in Microsoft Excel format.
- Files with names ending in `.npy` or `.npz` are for Python users; the NumPy module can read them. Python can also read `.txt` and `.csv` files.

#1 = `HIVseries`: HIV infection time course. Files `HIVseries.mat` , `HIVseries.csv`, `HIVseries.npy` contain variable `"a"` with two columns of data. The first is the time in days since administration of a treatment to an HIV positive patient; the second contains the concentration of virus in that patient's blood in arbitrary units. [Data from A Perelson. Modelling viral and immune system dynamics. Nature Revs. Immunol. (2002) vol. **2** (1) pp. 28–36 (Box 1).]

#2 = `Population`: Files `population.mat` , `population.csv`, `population.npy`: First column: date in years CE. Second column: Estimated world population. [Data from
http://www.vaughns-1-pagers.com/history/world-population-growth.htm = perma.cc/2BMU-92RX.]

#3 = `ShotNoise`: Files `shotNoise.mat` and `shotNoise.npy`: Variable `A` gives the arrival times of 290 photon absorption events in an avalanche photodiode detector. Time is measured in units of 50 `ns`. Total duration is 5 `s`. [Data courtesy John F Beausang.]
`shotNoise.txt`: Same data.

#4 = `Brownian`: Jean Perrin's data on Brownian motion. Files g26perrindata.mat , g26perrindata.txt , g26perrindata.csv, g26perrindata.npy [data from J Perrin, *Les Atomes*]. The columns give x, y coordinates of the points in Figure 3.3a, in μm. Each such point in turn is the net displacement of a colloidal particle of radius 0.37 nm over 30 s of Brownian motion.

#5 = `Clusters`: Geographic locations of a set of incidents. Map data of incidents from
`http://bombsight.org/#14/51.4465/-0.0756` = perma.cc/78NH-CJUN .
Some points are duplicated because the map indicated multiple unresolved hits at that point.
File londonIncidents.npz contains:
- variable `all` = array of (x,y) pairs describing map coordinates of each incident, in arbitrary units
- variable `allrefs` = array of four (x,y) pairs describing reference points. In real space, in 2019, those points are:
upper left: intersection of Grosvenor Pl and Duke of Wellington Pl
lower left: intersection of Chelsea Bridge Rd and Ebury Br Rd
upper right: intersection of Tooley St and Tower Bridge Rd
lower right: intersection of Old Kent Rd and Trafalgar Ave.
(Files londonIncidentsAll.npy and londonIncidentsAllrefs.npy contain the same data in npy format.)

#6 = `LDexpt`: Luria-Delbrück experiment.
Files LDexpt.mat , LDexpt.npz: The variable `expcounts23` gives data for the experiment shown in Figures 4.6 and 4.8 of *PMLS*. Each culture studied had a certain number m of resistant bacteria, out of about $2.4 \cdot 10^8$ bacteria total. The data specify a histogram of m with variable-width bins: Bin #`i` includes outcomes for m in the range `bins(i)` through `(bins(i+1)-1)`. The number of cultures with m in this range is `expcounts23(i)`.
Another experiment, called #22 (not shown in the text), was similar. However, in this experiment, each culture was sampled: only 1/4 of the total culture was withdrawn, spread on a plate, and checked for mutants. The result are in `expcounts22`, with the same bin structure as in experiment #23. This experiment had about $2.8 \cdot 10^8$ bacteria per culture.
Files LDexpt23.csv, LDexpt22.csv, LDexpt23.npy, LDexpt22.npy: Same data with `bins` in the first column and `expcounts` in the second column.
[Data from Luria and Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. Genetics (1943) vol. **28** pp. 491–511, table 5 p505.]

#7 = `Stocks`: Files djiweekly2020.npz and djiweekly2020.txt contain variable `djiweekly` with weekly closing values of the Dow Jones Industrial Average during the years from 1921 to 2020.

#8 = `FlowCytometry`: File flowCytometry.npy: Contains measured fluorescence values for each of a large collection of individual immune-system cells.
[Data from Erez, A, Vogel, R, Mugler, A, Belmonte, A, and Altan-Bonnet, G (2018). Modeling of cytometry data in logarithmic space: When is a bimodal distribution not bimodal? Cytometry Part A: the Journal of the International Society for Analytical Cytology, **93**(6), 611–619. doi.org/10.1002/cyto.a.23333, which cites github.com/AmirErez/BimodalLogspaceCytA . We have extracted the third column of the data file `20140920-OT1-dynamics_Specimen_005_E5_E0.txt` to let you reproduce Fig. 3 of the article.]

#9 = `COVID`: File 20-08-03time-series-covid19-confirmed-US.csv : Data from github.com/CSSEGISandData/COVID-19 , accessed 3 Aug 2020.
Data in folder:
`https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series`
File: `time_series_covid19_confirmed_US.csv`
Description:

https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/README.md says "Time series table for the US confirmed cases, reported at the county level."

#10 = `BrownianTrap`: vanMameren-raw.txt: Observed positions of a particle at a series of times separated by (1/195000)s. Only positions along one axis are recorded, in nanometers. Data kindly supplied by J. van Mameren and C. Schmidt.

#11 = `CryoEM`: File 1dImages.npz contains variables:
`samples` = 3D array containing simulated noisy images, each 85 pixels wide, for each of 3 simulated datasets
`samples[i,j,k]` where `i` = which image, `j` = position in image, `k` = which noise level
`noiselevels` = array containing $1/\sqrt{\text{SNR}}$
`shiftSD` = SD of the jitter (random shift) applied to each simulated image, in pixels.

File 2dImages.npz contains variables:
`samples` = 4D array containing simulated noisy images, each 85×85 pixels for each of 3 simulated datasets
`samples[i, jx, jy, k]` where `i` = which image, `j` = position in image, `k` = which noise level
`noiselevels` = array containing $1/\sqrt{\text{SNR}}$
`shiftSD` = SD of the jitter (random shift) applied to each simulated image, in pixels.

File test2Dblur.tif contains a 2D image to be used as an initial template for refinement.

#12 = `Photodiodeblips`: Files g112APDtraces.csv and g112APDtraces.npy contain columns of data. Columns 1–2 correspond to higher illumination; 4–5 are medium illumination; 7–8 are for the lowest illumination. In each pair of columns, the first entry is time in seconds; the second is detector output in volts at that time. [Data courtesy John F Beausang.]

#13 = `Vesicle`: g293vesicle.mat: The variable `amplitudes` is a list of bin centers; `frequencies` gives the corresponding number of times at which amplitudes falling in each bin were observed. The first bin contains the trials in which a stimulus evoked no response (failures). g293vesicle.xlsx, g293vesicle.csv and g293vesicle.npy: Same data. The first column contains amplitudes. The second column contains frequencies.
To show more clearly the discreteness of the response, the mean number of neurotransmitter vesicles released in response to an action potential was reduced relative to normal by raising the concentration of magnesium in the solution surrounding the cells. The muscle cell was itself prevented from firing any action potential; thus, its electrical response was a proxy for the amount of neurotransmitter actually released. [Data from Boyd and Martin. The end-plate potential in mammalian muscle. J Physiol. (Lond.) (1956) vol. **132** (1) pp. 74–91, Fig. 8.]

#14 = `MyosinV`: Files g42myosinwalk.mat and g42myosinwalk.npz: The files contain these variables:
`yildizHistoRed` = stepping histogram for myosins taking only 70 nm steps;
`yildizHistoGreen` = stepping histogram for myosins taking $(70 - x)$ nm steps alternating with x nm steps.
In each case, each row consists of the pair ((center of histogram bar in seconds), (observed frequency)).
`redDelta=0.5, greenDelta=1.0` are the respective bin widths in seconds.
Files yildizHistoRed.csv, yildizHistoGreen.csv, yildizHistoRed.npy, and yildizHistoGreen.npy contain the same information as the corresponding variables in the `.mat` and `.npz` files.
[Data from Yildiz et al. Myosin V walks hand-over-hand: Single fluorophore imaging with 1.5-nm localization. Science (2003) vol. **300** (5628) pp. 2061–2065, Fig. 6.]

#15 = `LinearFitPoisson`: linearFitPoisson.mat, linearFitPoisson.csv, and linearFitPoisson.npy contain variables: `xvals`=(distance from radioactive source to detector, m)$^{-2}$; `counts`=(counts in detector in a fixed time

interval).

#16 = `Bursting`: GoldingData.mat and GoldingData.npz: Arrays `panelA1,2,3` contain data from three different trials. First column: time after induction, minutes. Second column: sample mean of the number of mRNA molecules per bacterium.
Array `panelC`: First column: $\log_{10}$ of the sample mean of mRNA count in steady state, for various degrees of induction. Second column: $\log_{10}$ of the corresponding variances of mRNA count.
Arrays `panelD1,2,3` contain data from three different trials. First column: time after induction, minutes Second column: natural logarithm of the fraction of cells with zero copies of mRNA.
Files panelA1.txt , panelA1.npy, panelA2.txt , panelA2.npy, panelA3.txt , panelA3.npy, panelC.txt , panelC.npy, panelD1.txt , panelD1.npy, panelD2.txt , panelD2.npy, panelD3.txt , panelD3.npy: Same data.
[Data from Golding et al. Real-time kinetics of gene activity in individual bacteria. Cell (2005) vol. **123** (6) pp. 1025–1036 (see Figure 10.6).]

#17 = `Colquhoun`: File colquhounData.npy: Column 2 contains the integer numbers of channels that closed during a 0.5 ms time window starting at the value in column 1 (in ms). (The first bin was unreliable and is not included.)
[Data from D Colquhoun and AG Hawkes, "Principles of the stochastic interpretation of ion-channel mechanisms" in *Single-channel recording* 2nd ed. (New York, Plenum, 1995) Eds. B Sakmann and E Neher. Caption says "(*R. temporaria,* synaptic channels, 50 nM acetylcholine, −80mV, 8 °C. DC Ogden, DJ Adams, and D Colquhoun unpublished data.)"]

#18 = `Kinesin`: [Data from Visscher et al. DOI: 10.1038/22146 .]
File Visscher.npz contains three arrays: `velocityVsATP`: Selected points from Fig 2 of the paper. Column 1: ATP concentration, μM. Column 2: Kinesin velocities, μm/s.
`forces`: Retarding force for each data point, pN.
`velocityVsForce`: Selected points from Fig 3A of the paper, all with [ATP]=5 μM. Column 1: Retarding force, pN. Column 2: Kinesin velocities, μm/s.

#19 = `HbMb`: Binding curves of two macromolecules for oxygen. hemoglobinmyoglobin.mat , hemoglobinmyoglobin.npz , hemoglobin.csv , hemoglobin.npy , myoglobin.csv , myoglobin.npy : Variable `hemoglobin`: first column, concentration in M; second column, probability to be unbound.
Variable `myoglobin`: same format.
[Data from FC Mills, ML Johnson, and GK Ackers. 1976. Oxygenation-linked subunit interactions in human hemoglobin. Biochemistry **15**:5350–5362 and A Rossi-Fanelli and E Antonini. 1958. Studies on the oxygen and carbon monoxide equilibria of human myoglobin. Arch Biochem Biophys **77**:478–492.]

#20 = `Autoregulation`: rosenfeld.mat and rosenfeld.npz: Synthetic governor. Variables named `regulatedX` are experimental data for regulated gene; `unregulated` is data for the unregulated system. In each variable, column 1 is time in units of cell cycles; column 2 is protein per cell normalized to its final value.
rosenfeld.csv , rosenfeld.xlsx: same data.
[Data from Rosenfeld et al. Negative autoregulation speeds the response times of transcription networks. J Mol Biol (2002) vol. **323** (5) pp. 785–793, Fig. 3.]

#21 = `Epidemic`: Measles outbreak data (Netherlands, 2000). Data from van Steenbergen, J. E., van den Hof, S., Langendam, M. W., van de Kerkhof, J. H. T. C., and Ruijs, W. L. M. (2000). Measles Outbreak—Netherlands, April 1999–January 2000. Morbidity and Mortality Weekly Report, 49(14), 299–303.
File SIAMRev.Hethcote.txt: First column, time [days]. Second column, number of cases.

#22 = `Novick`: Novick/Weiner data. g149novickA.mat , g149novickA.txt , g149novickA.npy: From their Fig. 1 (high inducer). First column: Time in hours. The e-folding time was about 3 hours. Second column: Fraction of maximum beta-galactosidase activity.
g149novickB.mat , g149novickB.txt , g149novickB.npy: From their Fig. 4 (moderate inducer). First column: Time in hours. The e-folding time was about 3 hours. Second column: Fraction of maximum beta-galactosidase activity.
[Data from Novick and Weiner. Enzyme induction as an all-or-none phenomenon. Proc Natl Acad Sci USA (1957) vol. **43** (7) pp. 553–566.]

#23 = `Superspreaders`: Laxminarayan.txt , Laxminarayan.npy : Distribution of the number of SARS COV2 infections attributable to a single individual (first column). Second column is observed frequency [Fig 2A2 of Laxminarayan, R., et al. (2020). Epidemiology and transmission dynamics of COVID-19 in two Indian states. Science (New York, N.Y.), 370(6517), 691–697. doi.org/10.1126/science.abd7672].
Laxminarayan.npz : Same data in variables named `ell` and `frequency`.