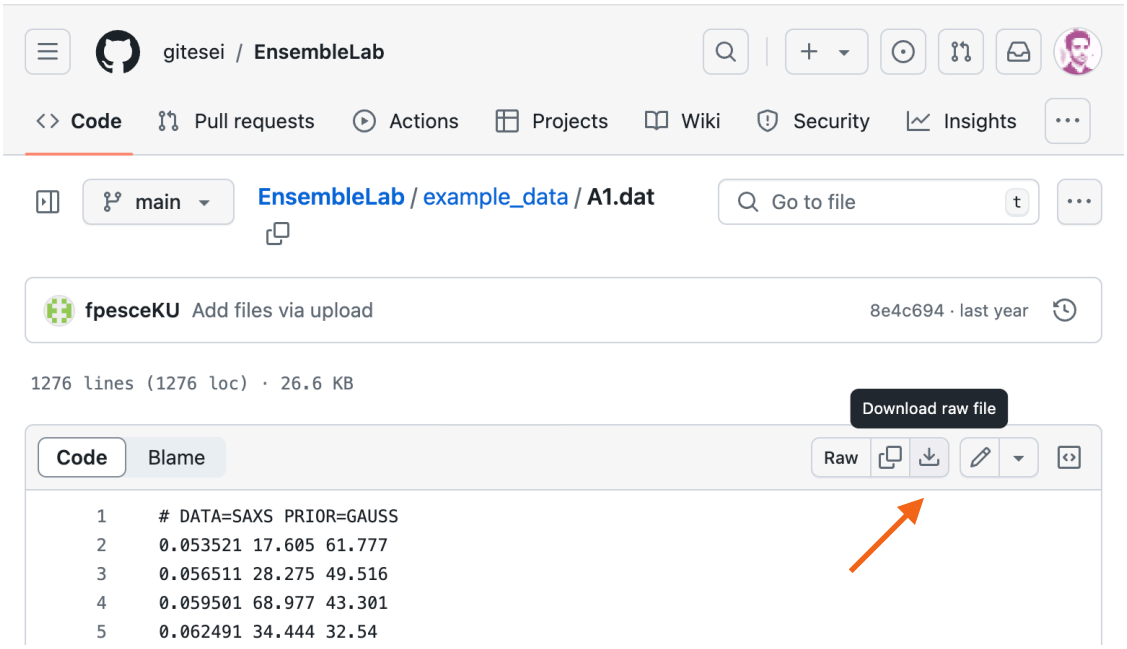
EnsembleLab class tutorial

Authors: Francesco Pesce and Giulio Tesei

1. Open the following link (**Google Chrome** is required for everything to work properly) https://colab.research.google.com/github/gitesei/EnsembleLab/blob/main/EnsembleLab.ipynb

2. Make sure to have **GPUs** enabled: go to “Runtime”, select “Change runtime type”, and select “GPU”.

3. The aim of the tutorial is to run molecular dynamics (**MD**) simulations of intrinsically disordered regions (**IDRs**) and integrate simulations with small-angle X-ray scattering (**SAXS**) data. At https://github.com/gitesei/EnsembleLab/tree/main/example_data you will find sequences and SAXS data for several IDRs. Choose the IDP you want to work with and download its sequence (“.fasta” file) and SAXS data (“.dat” file). Otherwise, you can use your own data. To download files from GitHub, open the file and click on the download button (see image below). If this button is not present (it is not if you are not logged in with a GitHub account), click on “Raw” and then “File>Save page as…” from Chrome’s menu.



4. In **0 - IDP sequence and data**, enter the NAME of the IDP that you have chosen, and paste its amino acid **SEQUENCE** in one-letter code. Select “**EXPERIMENT: SAXS**” and set the environmental conditions for the simulations (temperature, ionic strength, and pH). These should be set so as to reproduce the experimental conditions used for the SAXS

experiment (reported here). For further details on the SAXS experiments see DOI: 10.1016/j.bpj.2022.12.013. When all settings are in place, run the cell. A prompt will appear that allows to upload the file with SAXS data to session storage.

5. Run the four **Preliminary operations** cells **one by one**, waiting for the execution of each cell to complete (a green check mark will appear) before running the next. When the execution of "Preliminary operations: setting the environment (i)" is complete, the session will restart and Colab will report a message informing you that the session has crashed. That is required for all packages to work properly.

6. After all "Preliminary operations" are executed and complete, you can execute all the other cells either one by one or all at once by selecting "Run after" from the "Runtime" menu.

7. Cell **1 - Run MD simulation** will run a molecular dynamics simulation. The default option "AUTO" will set the simulation time depending on sequence length. The longer the IDR, the larger the ensemble of conformations it can adopt. Moreover, the reconfiguration time of IDRs increases with increasing sequence length. Therefore, longer sequences will require more sampling. Typical simulation times range from ca. 5 min (a 71 ns-long simulation of an IDR of 70 residues), 7 min (71 ns-long simulation of an IDR of 140 residues), to 34 min for a 373 ns-long simulation of an IDR of 351 residues. If GPUs are momentarily unavailable on Google Colab, proceed to the next cell. You can also set the simulation time (in ns) yourself instead of relying on the "AUTO" option. Remember that the simulation time (i.e. the extent of sampling) will affect the estimates for calculated averages and associated errors (see next point).

8. Cell **2 - Simulation analysis** will plot the distributions and averages of some structural parameters: radius of gyration, R_g; asphericity, Δ; prolateness, S; end-to-end distance, R_{ee}; and apparent Flory scaling exponent, ν. ν is calculated from a nonlinear fit to the ensemble-averaged inter-residue distances, $\sqrt{\langle R_{ij}^2 \rangle}$, as a function of sequence separations, $|i-j|$.

9. Cell **2.1 - Sequence analysis**, will calculate the following sequence descriptors: fractions of K, R, D, E, and aromatic residues; average stickiness, $\langle \lambda \rangle$; sequence hydropathy decorator, SHD; sequence charge decorator, SCD; charge segregation parameter κ; fraction of

charged residues, FCR; and net charge per residue, NCPR. $\langle\lambda\rangle$, SHD, SCD, κ, and FCR will be used to estimate the apparent Flory scaling exponent, ν_{SVR}, using a support vector regression model trained on simulations of tens of thousands of sequences of human IDRs (see https://doi.org/10.1038/s41586-023-07004-5).

10. In cell **3 - Ensemble reweighting against experimental data** the Bayesian/Maximum-entropy (BME) approach is used to reweight the MD simulations so that it better matches the SAXS data. This is done by first calculating SAXS curves for each of the n frames of the simulation trajectory using Pepsi-SAXS. Then BME reweighting is performed by minimizing the functional $\mathscr{L}(w_1 \dots w_n) = \frac{m}{2}\chi^2(w_1 \dots w_n) - \theta\, S_{rel}(w_1 \dots w_n)$, where $(w_1 \dots w_n)$ are the statistical weights associated with each frame of the simulation, χ^2 quantifies the agreement between simulation and SAXS, S_{rel} quantifies how much the new weights are different from the initial ones.
The agreement between reweighted simulation data and experiment is quantified as

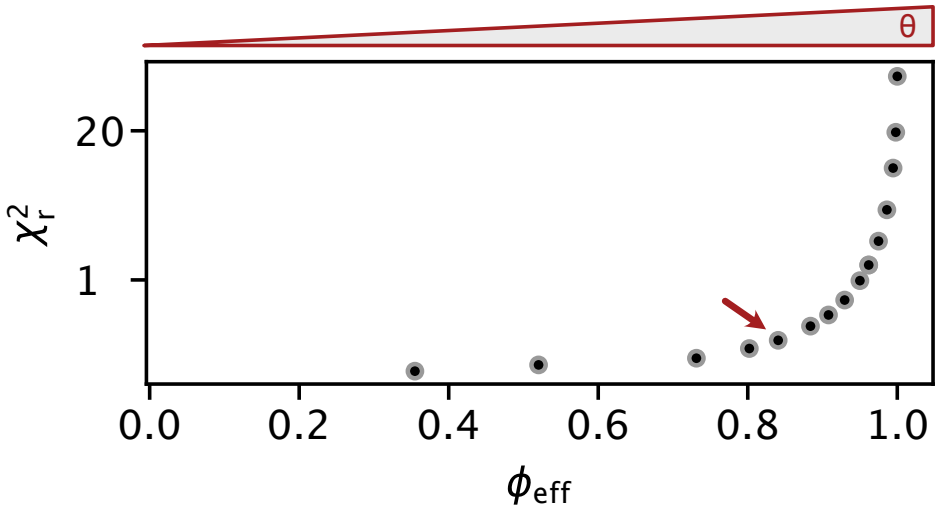
$$\chi^2(w_1 \dots w_n) = \frac{1}{m}\sum_i^m \frac{\left(\sum_j^n w_j I_j^{CALC}(q_i) - I^{EXP}(q_i)\right)^2}{\sigma(q_i)^2},$$

where $\sigma(q_i)$ is the error on the experimental SAXS intensities, $I^{EXP}(q)$.
The extent of deviation of the new weights from the equal weights of the simulation trajectory, $w_j^0 = 1/n$, is quantified by the Kullback-Liebler divergence:

$$-S_{rel}(w_1 \dots w_n) = \sum_j^n w_j \ln\left(\frac{w_j}{w_j^0}\right) = \sum_j^n w_j \ln\left(w_j \times n\right).$$

θ is a free parameter that must be tuned to strike a balance between obtaining a good agreement with the experimental data (low χ^2) and retaining as much information as possible from the starting simulation (high S_{rel}).

11. To fine-tune this hyperparameter, you will scan θ and for each value plot χ^2 vs $\phi_{eff} = \exp(S_{rel})$. A good value of θ is located at the elbow of the curve (see figure below).



12. Cell **3.2 - Analyze reweighted ensemble** This cell shows comparisons between SAXS curves and conformational properties from experiments (grey) and from the simulation trajectory before (blue) and after BME reweighting (red). What is the relative error of the predicted R_g, $(R_g^{sim} - R_g^{exp})/R_g^{exp}$, before and after BME reweighting?

13. As a first exercise, go back to cell **3.1** and switch the "**THETA_LOCATOR**" option from "AUTO" to "INTERACTIVE". After rerunning this cell, you can use a drop-down menu to select the θ value to use. Try selecting different θ values, both high and low, and run cell **3.2** again. How do the structural observables and the fit to SAXS change in response to changes in θ?

14. As a second exercise, go back to cell 1.1 and check the "Break_CALVADOS" box. That will add random noise to the λ values of the CALVADOS model (the amino-acid stickiness parameters). The resulting force field will likely not reproduce the experimental data accurately and the effect of reweighting will be more evident.

15. Finally, cell **4 - Download results** will trigger the download of a zip archive containing the data from the simulation and reweighting. The **README** file explains the content of the zip archive. The archive contains data that can be used to reproduce the plots from the notebook and the simulation files. The archive also contain trajectory and topology files that can be used to visualise the conformational ensemble using molecular visualisation software such as VMD, PyMol or Chimera.