# 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://colab.research.google.com/github/gitesei/EnsembleLab/tree/main/example_data you will find sequences and SAXS data for 3 IDPs.

4. In **0 - IDP sequence and data**, enter the **NAME** of the IDP that you want to work with (Sic1, NLS, or IBB), and paste its amino acid **SEQUENCE** in one-letter code in the text box in the first cell of the notebook. To compare with experiments, simulations will be run at the same solution conditions as the experiments. SAXS and Förster Resonance Energy Transfer (FRET) experiments were performed at 293 K (Sic1; https://doi.org/10.1021/jacs.0c02088) and 296 K (NLS and IBB; https://doi.org/10.1073/pnas.1704692114). For NLS and IBB, we also use the experimental pH=7.4 to set the charge on the histidine residues (assuming pKa=6). The ionic strength of the samples were 0.19 M for Sic1, and 0.15 M for NLS and IBB.

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 normal and required for all packages to work properly. Rerun the first two cells (it won’t take long).

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 the actual 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 IDP of 70 residues), 7 min (71 ns-long simulation of an IDP of 140 residues), to 34 min for a 373 ns-long simulation of an IDP of 351 residues. If GPUs are momentarily unavailable on Google Colab, proceed to the next cell.

8. Cell **2 - Simulation analysis**, will plot the distributions and averages of some structural parameters: radius of gyration, $R_g$; end-to-end distance, $R_{ee}$; apparent Flory scaling exponent, $\nu$; and conformational entropy per residue, $S_{conf}/N$ (see https://doi.org/10.1021/ct500684w). $\nu$ 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 $\kappa$; fraction of charged residues, FCR; and net charge per residue, NCPR. $\langle\lambda\rangle$, SHD, SCD, $\kappa$, and FCR will be used to estimate the apparent Flory scaling exponent, $\nu_{\mathrm{SVR}}$, using a support vector regression model trained on simulations of tens of thousands of sequences of human IDRs (https://doi.org/10.1101/2023.05.08.539815). What is the relative error of $\nu_{\mathrm{SVR}}$ with respect to the $\nu$ value obtained from the simulation trajectory?

10. In cell **3.1 - Execute reweighting** the Bayesian/Maximum-entropy approach is used to reweight the MD simulations so that it better matches the SAXS data. This is done by minimizing the functional $\mathscr{L}(w_1 \ldots w_n) = \frac{m}{2}\chi^2(w_1 \ldots w_n) - \theta\, S_{rel}(w_1 \ldots w_n)$, where $(w_1 \ldots w_n)$ are the statistical weights associated with each frame of the simulation, the $\chi^2$ 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

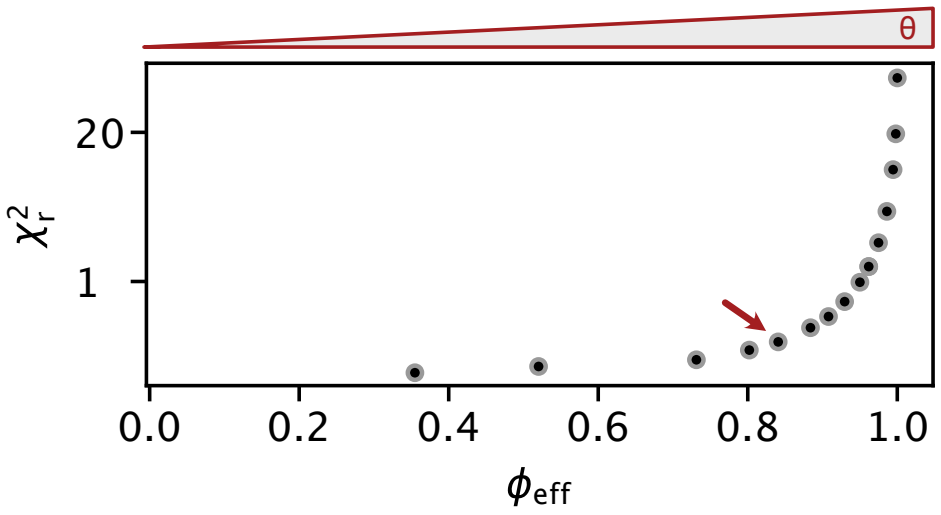
$$\chi^2(w_1 \ldots w_n) = \frac{1}{m} \sum_i^m \frac{\left( \sum_j^n w_j I_j^{CALC}(q_i) - I^{EXP}(q_i) \right)^2}{\sigma(q_i)^2},$$

where $\sigma_i(q_i)$ is the error on the experimental SAXS intensity. 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

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

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

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




12. Cell **3.2 - Analyze reweighted ensemble** shows a comparison between the SAXS curve and the $R_g$ distribution from the simulation trajectory before and after BME reweighting. What is the relative error of the predicted $R_g$ before and after BME reweighting? Estimate the relative change in $R_g$, $R_{ee}$, and $\nu$ upon BME reweighting. Do these quantities increase or decrease to a similar extent?

13. As an 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 slider to select the $\theta$ value to use. Try selecting different $\theta$ values, both high and low, and then run cell **3.2** again. How do the structural observables and the fit to SAXS change in response to changes in $\theta$?

14. 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.

15. Bonus: cell **5 - Calculate FRET efficiency** uses the program FRETpredict (https://github.com/KULL-Centre/FRETpredict) to calculate the average FRET efficiency from the all-atom trajectory using a rotamer library approach. Briefly, for each trajectory frame, conformational ensembles of the fluorescent probes are inserted at the labeled residues to sample the distributions of dye-dye separations and relative orientations. In this calculation, we assume that the complete conformational sampling of both protein and dyes is achieved within the fluorescence lifetime of the donor. What is the relative error of the average FRET efficiency from simulations with respect to the reference experimental value?

16. Does the agreement between predicted and experimental average FRET efficiency improve after reweighting the trajectory using SAXS data and the BME procedure (see cell **5.1 - Reweight FRET efficiency**)?