EnsembleLab class tutorial

1. Open the following link (**Google Chrome** is required for everything to work properly) https://colab.research.google.com/github/fpesceKU/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 proteins (**IDPs**) and analyze simulations in connection with small-angle X-ray scattering (**SAXS**) data. At https://github.com/fpesceKU/EnsembleLab/tree/main/example_data you can find sequences and SAXS data for 11 IDPs. Choose the IDP you want to work on and download its sequence (“.fasta” file) and SAXS data (“.dat” file). Otherwise, you can use your own data.

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, pH, ionic strength). These should be set so as to reproduce the experimental conditions used for the SAXS experiment (reported here). 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 three “**Preliminary operations**” cells **one by one**, waiting for the execution of each cell to be 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 related to the session crashing. That is normal and required for all packages to work properly.

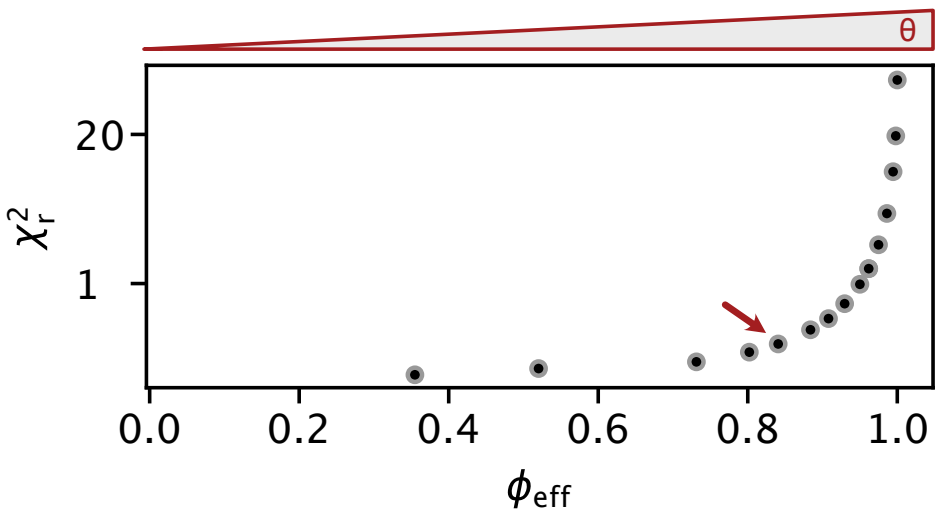
6. After all “Preliminary operations” are executed and complete, you can execute the following cells altogether (executing cell-by-cell or selecting “Run after” from the “Runtime” menu).

7. Cell “**2.1 Launch MD simulation**” will run the actual simulation. The default option “AUTO” will set the simulation time depending on the sequence length. The longer is an IDP, the larger the ensemble of conformations it can adopt. 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 with 351 residues. 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 amount of sampling) will affect the estimates for calculated averages and associated errors (see next point).

8. Cell "**2.2 Calculate structural observables from simulation**", will plot the distributions and averages of some structural parameters: radius of gyration (R_g), hydrodynamic radius (R_h), end-to-end distance (D_{ee}), and scaling exponent (ν) from fitting the scaling profile. These calculated averages (as any other measurement) will be affected by an error. As an optional exercise, you can vary the simulation time and see how this affects calculated averages and errors.

9. In cell "**3.2 Execute reweighting**" the Bayesian/Maximum-entropy approach is used to reweight the MD simulations so that it better matches SAXS data. This is done by minimizing the functional $\mathscr{L}(w_0 \dots w_n) = \frac{m}{2}\chi^2(w_0 \dots w_n) - \theta \cdot S_{rel}(w_0 \dots w_n)$, where $(w_0 \dots w_n)$ are the statistical weights associated with each frame of the simulation, the χ^2 quantifies the agreement between simulation and SAXS, S_{rel} quantifies how much the new weights are different from the initial ones. θ is a free parameter that must be tuned so as to ensure a reweighted distribution with a good agreement with the experimental data (low χ^2) while retaining as much information as possible from the starting simulation (high S_{rel}).
From a practical perspective what one does is scan multiple values for θ and plot the resulting χ^2 vs. $\phi_{eff}(= \exp(S_{rel}))$. A good value for θ will be located in the elbow of the curve, i.e. the largest extent of χ^2 minimization together with the least extent of deviation from the starting simulation (see figure below).



10. Cell "**3.3 Analyze reweighted ensemble**" will show the fit of the simulation to the experimental SAXS curve and the same quantities from cell **2.2**, but reweighted against SAXS.

11. As an exercise, go back to cell **3.2** and switch the "**THETA_LOCATOR**" option from "AUTO" to "INTERACTIVE". After rerunning this cell, you can use a slider to select the θ value to use. Try selecting different θ values, both upstream and downstream of the curve, and then run cell **3.3** again. How do the structural observables and the fit to SAXS change in response to changes in θ?

12. Finally, cell "**4. Download results**" will trigger the download of a zip archive containing the data from the simulation and reweighting. A **README** file is included that 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. In the simulation folder, you will find 10 pdb files that are representative of the different sizes that the IDP that you have simulated can adopt. You can visualize these structures with e.g. ChimeraX (recommended) and PyMol.