This directory contains all the inputs and outputs from an EOM run on PAR15 in 100mM NaCl.

* Accounting of all PDB subdirectories for each PAR sample (these are derived from the molecular dynamics trajectories from each respective simulation):
  + PAR15na: 2066 PDB files – this is the one in the example dataset
  + PAR15naMg: 2204 PDB files
  + PAR22na: 2298 PDB files
  + PAR22naMg: 2276 PDB files

Additionally, these measures were performed:

* Have experimental data in subdirectory saxs\_data
* Rename all PDBs to {0,1,2,3,…}.pdb convention
* Compute CRYSOL scattering of all PDBs in directory. This will output a .int file. This is done by running “STARTHERE\_computePoolIntensities.m”, which has several dependencies that are also in the folder. You no longer need all individual PDBs and can just work from this .int output when calling EOM. The settings utilized in CRYSOL are present in that script.
* Use GAJOE v1.3 manually through the command window, after obtaining the .int file with CRYSOL intensities.
  + You do NOT need to recompute all CRYSOL intensities to re-do GAJOE – just the .int file
  + We tested GAJOE v2.1 but found there is no difference except additional output of Rflex and Rsigma values, which are difficult to interpret since our pool is derived from a long MD simulation
  + We want to use GAJOE v1.3 because the outputs it gives are compatible with the established downstream analysis framework (GAJOE v2.1 would require recoding it due to different output styles)

This is what it should look like in the command window (this is the ATSAS EOM interface). I am running on Windows but it should be similar on a Unix based OS:

* Text

  Description automatically generated with low confidence
* to run GAJOE1.3 from Powershell, type “.\gajoe13” with the gajoe13.exe executable in the working directory
* Parameters used:
  + 20 curves per ensemble (previously 5)   
    1000 generations (default is 1000, 10000 is the max – start with 1000, look at convergence file to see if X2 is converging – if not then go to 10000 generations)   
    50 ensembles   
    10 mutations per ensemble  
    20 crossings per generation   
    100 times to repeat the genetic algorithm (previously 50)

**Documentation of production runs:**

PAR15na (GA001)

Graphical user interface, text

Description automatically generated Final X2=0.140  
[…] . It failed to create the analysis files (size histograms) because it couldn’t find size\_list file, but no problem we will create those histograms ourselves downstream anyways.

Ran in <3 min. Looking at the convergence file, there is essentially no change in X2 at later generations. There is no need to up to 10000 generations. Final fit to SAXS data looks pretty good as well.

Table

Description automatically generated

PAR15naMg (GA001)

Graphical user interface, text

Description automatically generated with medium confidence Final X2=0.154

PAR22na (GA001)

Graphical user interface, text

Description automatically generatedFinal X2=0.400

For this one, the final fit to low q is not perfect, and the X2 seems like it could still be trying to reach convergence. Maybe here we run for 10000 generations instead of 1000. But let’s do the next one first.

PAR22naMg (GA001)

Graphical user interface, text

Description automatically generatedFinal X2=0.312

Fit looks better and more converged than PAR22na. Note that we are reporting the Final X2 values of the last generation. The X2 of the final curve compared to experimental data might be slightly different.

**Production runs part 2 – 10000 generations rather than 1000.**

Try first on PAR22na, the worst result previously. If this works better then we’ll probably need to do 10000 generations on all four cases for consistency.

PAR22na, 10000 generations (GA004; GA002&3 were command re-entrant mulligans):

Graphical user interface, text

Description automatically generated with medium confidence Final X2=0.379

This took around 20min. Still not too bad. X2 convergence a bit better, not much tho and it seems like it has reached convergence.

PAR22naMg, 10000 generations (GA002):

Graphical user interface, text

Description automatically generated Final X2=0.300 (extremely stable)

PAR15na, 10000 generations (GA002):

Text

Description automatically generated Final X2=0.139

Not really that much better than last run at 1000 generations.

PAR15naMg, 10000 generations (GA002):

Text

Description automatically generated Final X2=0.154

No practical improvement compared to 1000 generations.

Retrying PAR22na, 10000 generations (GA006; GA005 was a command re-entrant mulligan)  
Truncating out the first three points of the scattering curve (new qmin=0.0157A-1. This is because there is an aberrant point that might be dragging down the entire fit in the Guinier regime.

Chart, line chart, scatter chart

Description automatically generated

Text

Description automatically generated with medium confidence Final X2=0.383

Results:

PAR15na:

'Mean Rg of full pool: 23.5776'

'Standard error in Rg of full pool: 0.12441'

'Mean Rg of EOM pool: 23.9416'

'Standard error in Rg of EOM pool: 0.13202'

'Mean Dmax of full pool: 75.8338'

'Standard error in Dmax of full pool: 0.35864'

'Mean Dmax of EOM pool: 76.7702'

'Standard error in Dmax of EOM pool: 0.40188'

'Mean Ree of full pool: 52.4023'

'Standard error in Ree of full pool: 0.46314'

'Mean Ree of EOM pool: 55.9462'

'Standard error in Ree of EOM pool: 0.54357'

PAR15naMg:

'Mean Rg of full pool: 19.5742'

'Standard error in Rg of full pool: 0.08868'

'Mean Rg of EOM pool: 25.7914'

'Standard error in Rg of EOM pool: 0.11512'

'Mean Dmax of full pool: 66.5837'

'Standard error in Dmax of full pool: 0.26988'

'Mean Dmax of EOM pool: 82.7046'

'Standard error in Dmax of EOM pool: 0.41487'

'Mean Ree of full pool: 40.131'

'Standard error in Ree of full pool: 0.34123'

'Mean Ree of EOM pool: 57.9845'

'Standard error in Ree of EOM pool: 0.57747'

PAR22na:

'Mean Rg of full pool: 28.1178'

'Standard error in Rg of full pool: 0.11954'

'Mean Rg of EOM pool: 31.516'

'Standard error in Rg of EOM pool: 0.16561'

'Mean Dmax of full pool: 89.8216'

'Standard error in Dmax of full pool: 0.3476'

'Mean Dmax of EOM pool: 98.2681'

'Standard error in Dmax of EOM pool: 0.56797'

'Mean Ree of full pool: 62.1408'

'Standard error in Ree of full pool: 0.46821'

'Mean Ree of EOM pool: 69.303'

'Standard error in Ree of EOM pool: 0.67601'

PAR22naMg:

'Mean Rg of full pool: 22.6326'

'Standard error in Rg of full pool: 0.13679'

'Mean Rg of EOM pool: 27.0042'

'Standard error in Rg of EOM pool: 0.12702'

'Mean Dmax of full pool: 75.0717'

'Standard error in Dmax of full pool: 0.37276'

'Mean Dmax of EOM pool: 86.3239'

'Standard error in Dmax of EOM pool: 0.32713'

'Mean Ree of full pool: 47.8218'

'Standard error in Ree of full pool: 0.44208'

'Mean Ree of EOM pool: 50.3611'

'Standard error in Ree of EOM pool: 0.55096'

For PAR15 run for 10000 generations, the Rg seems to be greater for PARna than PARnaMg. Upon checking the PAR15 runs for 1000 generations, they make more sense:

PARna (1000 generations):

'Mean Rg of full pool: 19.5742'

'Standard error in Rg of full pool: 0.08868'

'Mean Rg of EOM pool: 25.7074'

'Standard error in Rg of EOM pool: 0.10914'

'Mean Dmax of full pool: 66.5837'

'Standard error in Dmax of full pool: 0.26988'

'Mean Dmax of EOM pool: 82.8669'

'Standard error in Dmax of EOM pool: 0.40821'

'Mean Ree of full pool: 40.131'

'Standard error in Ree of full pool: 0.34123'

'Mean Ree of EOM pool: 57.8246'

'Standard error in Ree of EOM pool: 0.57744'

PARnaMg (1000 generations):

'Mean Rg of full pool: 23.5776'

'Standard error in Rg of full pool: 0.12441'

'Mean Rg of EOM pool: 23.9411'

'Standard error in Rg of EOM pool: 0.13263'

'Mean Dmax of full pool: 75.8338'

'Standard error in Dmax of full pool: 0.35864'

'Mean Dmax of EOM pool: 76.7966'

'Standard error in Dmax of EOM pool: 0.40322'

'Mean Ree of full pool: 52.4023'

'Standard error in Ree of full pool: 0.46314'

'Mean Ree of EOM pool: 55.5448'

'Standard error in Ree of EOM pool: 0.54614'

Let’s use 1000 generation runs for PAR15 and 10000 generation runs for PAR22, under the rationale that for PAR22, the PARna run did not converge after 1000 generations but for PAR15 it did.

Post EOM MD vs SAXS comparisons (here I show CRYSOL scattering from raw MD):

PAR15na

Chart, line chart

Description automatically generated

PAR15naMg

Chart

Description automatically generated

PAR22na

Chart, histogram

Description automatically generated

PAR22naMg

Chart, radar chart

Description automatically generated

SAXS .dat files are saved and plotted in publication figures.

Hierarchical clustering (plots will be shown in SI Figures):

In {Dmax,Rg} space:

PAR15na:

Text, letter

Description automatically generated

Representative structures: 497 761 702 516

PAR15naMg:

Text, letter

Description automatically generated

Representative structures: 944 2106 966

PAR22na:

Text, letter

Description automatically generated

Representative structures: 580 497 343 62

PAR22naMg:

Text, letter

Description automatically generated

Representative structures: 32 42 247

In {Ree,Rg} space:

PAR15na:

Text, letter

Description automatically generated

Representative structures: 542 168 810 505 516

PAR15naMg:

A picture containing text, person, document, crowd

Description automatically generated

Representative structures: 2148 2121 1460 1592 1620

PAR22na:  
Text, letter

Description automatically generated

Representative structures: 574 398 88 67

PAR22naMg:

Text, letter

Description automatically generated

Representative structures: 24 253 137