Supplementary File I

Options of `cryosieve`

`--reconstruct_software`

CryoSieve utilizes RELION's reconstruction module and postprocessing module for the 3D reconstruction and postprocessing step. In the `--reconstruct_software` option, provide the correct command prefix to invoke `relion_reconstruct` or `relion_reconstruct_mpi`. For instance, replaced it with the absolute directory path, if RELION binaries are not located in a directory listed in the `PATH` variable. Optionally, the multi-processing version of RELION's reconstruction is revoked by using the following formats: `--reconstruct_software "mpirun -n 5 relion_reconstruct_mpi"`. Ensure to enclose the string within double quotation marks (`""`), so that it is treated as a single argument.

`--postprocess_software`

The `--postprocess_software` option follows a similar pattern as `--reconstruct_software`. However, it is not mandatory for the sieving procedure. CryoSieve automatically skip the postprocessing step, if this parameter is left empty. Omit this parameter in the command to skip the postprocess step.

`--i`

The `--i` option specifies the path of the input star file, which can be either a relative or an absolute path.

`--o`

The `--o` option serves as the prefix for all output files. The folder path specified in the "–o" parameter will be automatically created. Assign it as `XXX` to generate resulting files with names like `XXX_iter0.star` and `XXX_iter0_half1.mrc`.

`--mask`

The `--mask` options refers to the path of the input mask file.

`--angpix`

The `--angpix` option specifies the pixel size in units of Angstroms.

`--num_iters`, `--frequency_start` and `--frequency_end`

The parameters `--num_iters`, `--frequency_start`, `--frequency_end`, and `--retention_ratio` play a crucial role in controlling the performance of

CryoSieve. CryoSieve operates as an iterative sieving process. The number of iterations is determined by the value specified in the `--num_iters` parameter. Within each iteration, the number of particles retained is calculated by multiplying the `--retention_ratio` with the number of particles from the previous iteration. CryoSieve employs a specific scoring function to sort particles. This scoring function compares the compatibility of information above a certain frequency. The cut-off frequency starts at the value specified in the `--frequency_start` parameter and linearly increases to reach the value specified in the `--frequency_end` parameter. The units for `--frequency_start` and `--frequency_end` are Angstroms. For high-quality datasets, it is recommended to use higher frequency settings, such as setting `--frequency_start` to 40 and `--frequency_end` to 2. For datasets containing a large number of particles or with low resolution, it is advisable to use a lower `--frequency_start`, such as setting `--frequency_start` to 60 and `--frequency_end` to 4. Experience shows that setting `--frequency_end` to the resolution of the reconstruction density map of the dataset yields better results. The `--retention_ratio` is generally set to 0.8, indicating that 20% of particles are discarded in each iteration. `--num_iters` is usually set to 10, but if the resolution of reconstruction keeps increasing with the number of iterations, it can be set to a larger number, like 15 or 25, to discard more useless particles.

`--num_gpus`

The `--num_gpus` option specifies the number of GPUs to be utilized during the execution of CryoSieve.

`--sym`

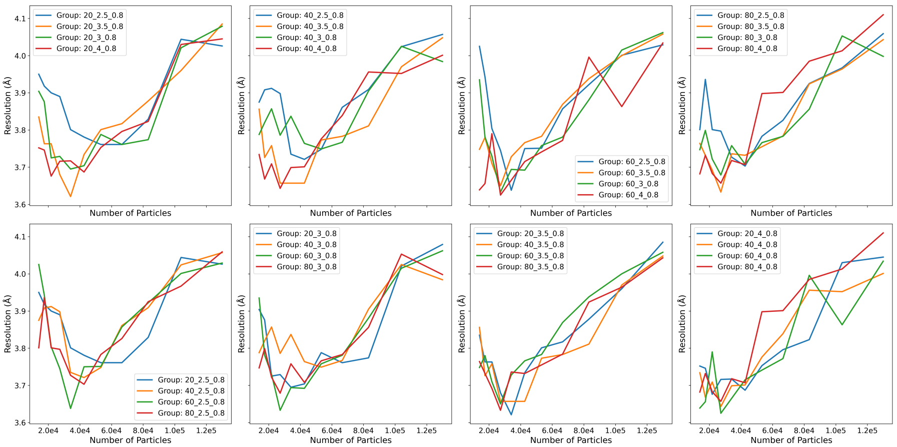
The `--sym` parameter is the molecular symmetry.

`--balance`

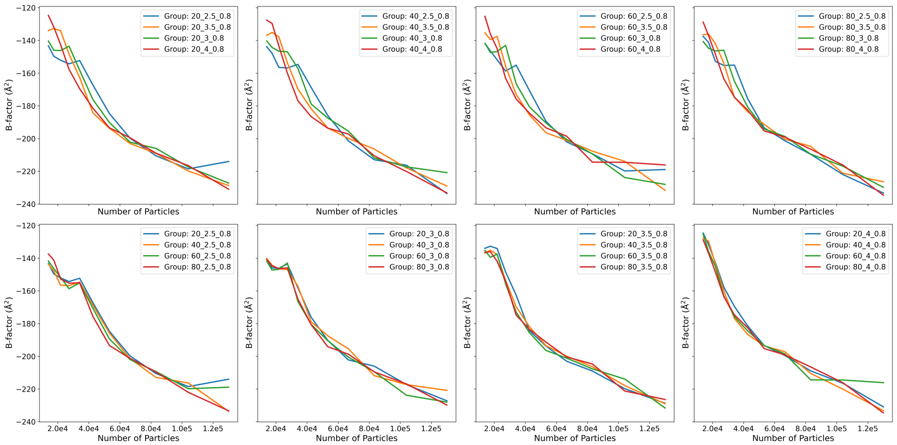
If the `--balance` flag is set, CryoSieve ensures that the number of sieved particles in the gold-standard split of two half sets is equal.

Choosing `--frequency_start` and `--frequency_end` options of `cryosieve`

Based on experimental observations (Supplementary Figure 1 and 2), the performance of CryoSieve remains notably consistent across a range of parameters, except when employing extreme values for start frequency or end frequency. Thus, it is strongly advised to utilize the default settings for start and end frequencies specified in CryoSieve.



Supplementary File Figure 1: **Comparision of the resolutions derived from cryoSPARC homogeneous refinement jobs, using star files generated across various iterations of CryoSieve under differing parameters as input.** The groups are labeled following a `start-frequency_end-frequency_retaining-ratio` format. For instance, `60_3_0.8` represents a configuration where the starting frequency is set to 60 Å, the end frequency to 3 Å, and the sieving retention ratio for each iteration is 80%. X-axis represents the number of particles retained after different iterations of CryoSieve, while the y-axis shows the resolution reported by re-estimating poses via cryoSPARC.



Supplementary File Figure 2: **Comparison of the B-factor derived from cryoSPARC homogeneous refinement jobs, utilizing star files generated by various iterations of CryoSieve with differing parameters as input.** This figure follows the same conventions as Supplementary Figure 1, with the only difference being that the Y-axis denotes the B-factor.