SCIPION

A Software Framework for Integration, Reproducibility and Validation in Cryo-EM

Biocomputing Unit,
Instruct Image Processing Center, CNB-CSIC
J.M. de la Rosa Trevín

CSIC

instruct Integrating Biology

UAM UNIVERSIDAD AUTONOMA DE MADRID

BIO COMPUTING Unit

instruct image Processing Center

We are witnessing a Revolution in Cryo-EM

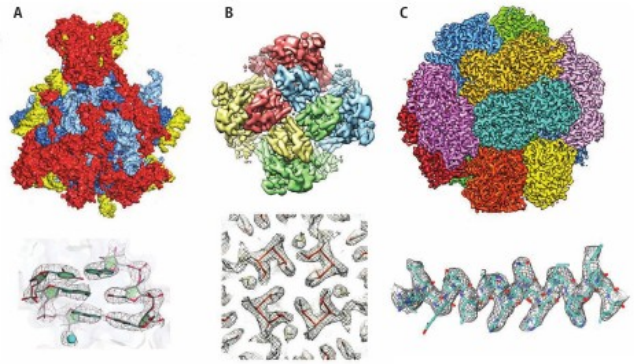
BIOCHEMISTRY

The Resolution Revolution

Werner Kühlbrandt

Advances in detector technology and image processing are yielding high-resolution electron cryo-microscopy structures of biomolecules.

Precise knowledge of the structure of macromolecules in the cell is essential for understanding how they function. Structures of large macromolecules can now be obtained at near-atomic resolution by averaging thousands of electron microscope images recorded before radiation damage accumulates. This is what Amunts *et al*. have done in their research article on page 1485 of this issue (*1*), reporting the structure of the large subunit of the mitochondrial ribosome at 3.2 Å resolution by electron cryo-microscopy (cryo-EM). Together with other recent high-resolution cryo-EM structures (*2*–*4*) (see the figure), this achievement heralds the beginning of a new era in molecular biology, where structures at near-atomic resolution are no longer the prerogative of x-ray crys-



eLIFE elifesciences.org

INSIGHT

MICROSCOPY

Cryo-EM enters a new era

Advances in detector hardware and image-processing software have led to a revolution in the use of electron cryo-microscopy to determine complex molecular structures at high resolution.

WERNER KÜHLBRANDT

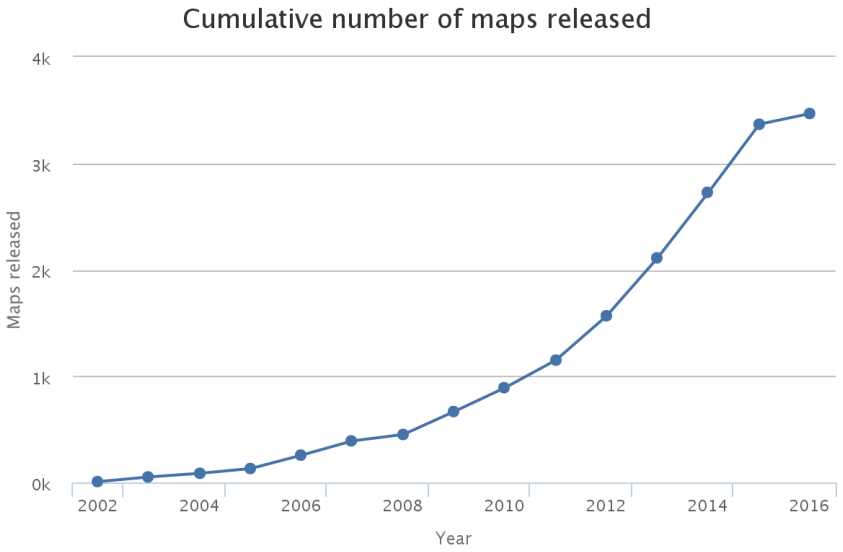
CSIC

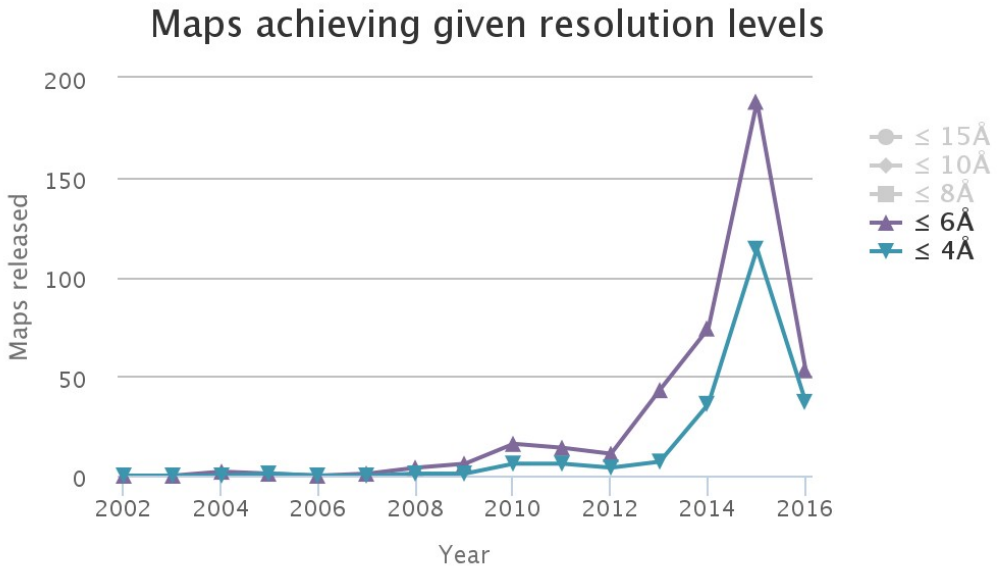
instruct Integrating Biology

BIO COMPUTING Unit

instruct image Processing Center

Maps released at EMDB increase..
and the achieved resolution





CSIC

instruct Integrating Biology

BIO COMPUTING Unit

instruct image Processing Center

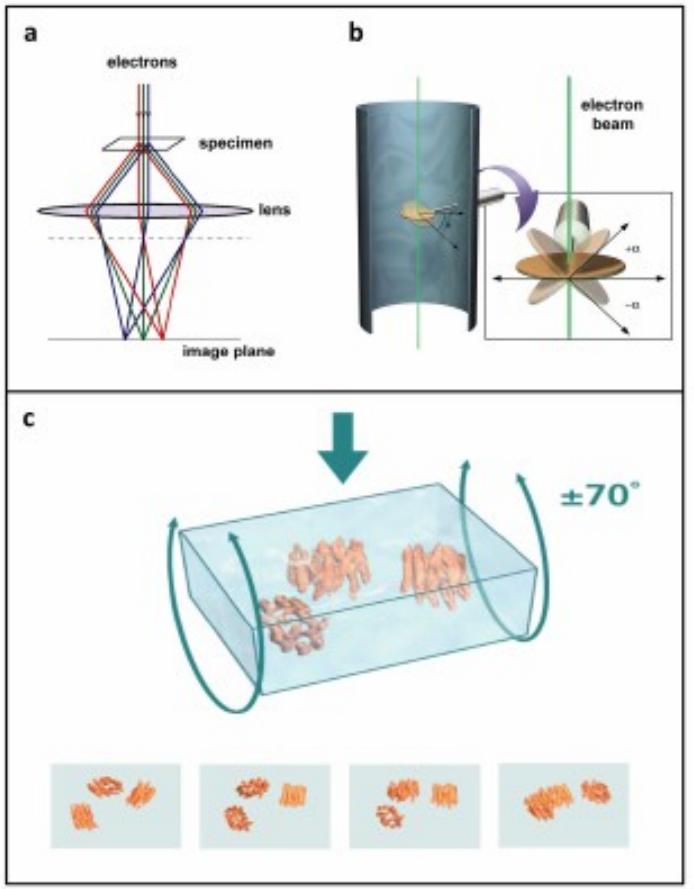
Key contributing factors

1. Improvement of microscopes, specially the Direct Detector Devices (DDDs).

2. Development of image processing software and increasing computing power.

Cryo-EM encompass different experimental techniques

1. Single Particles
2. Tomography
3. Helical
4. Crystallography



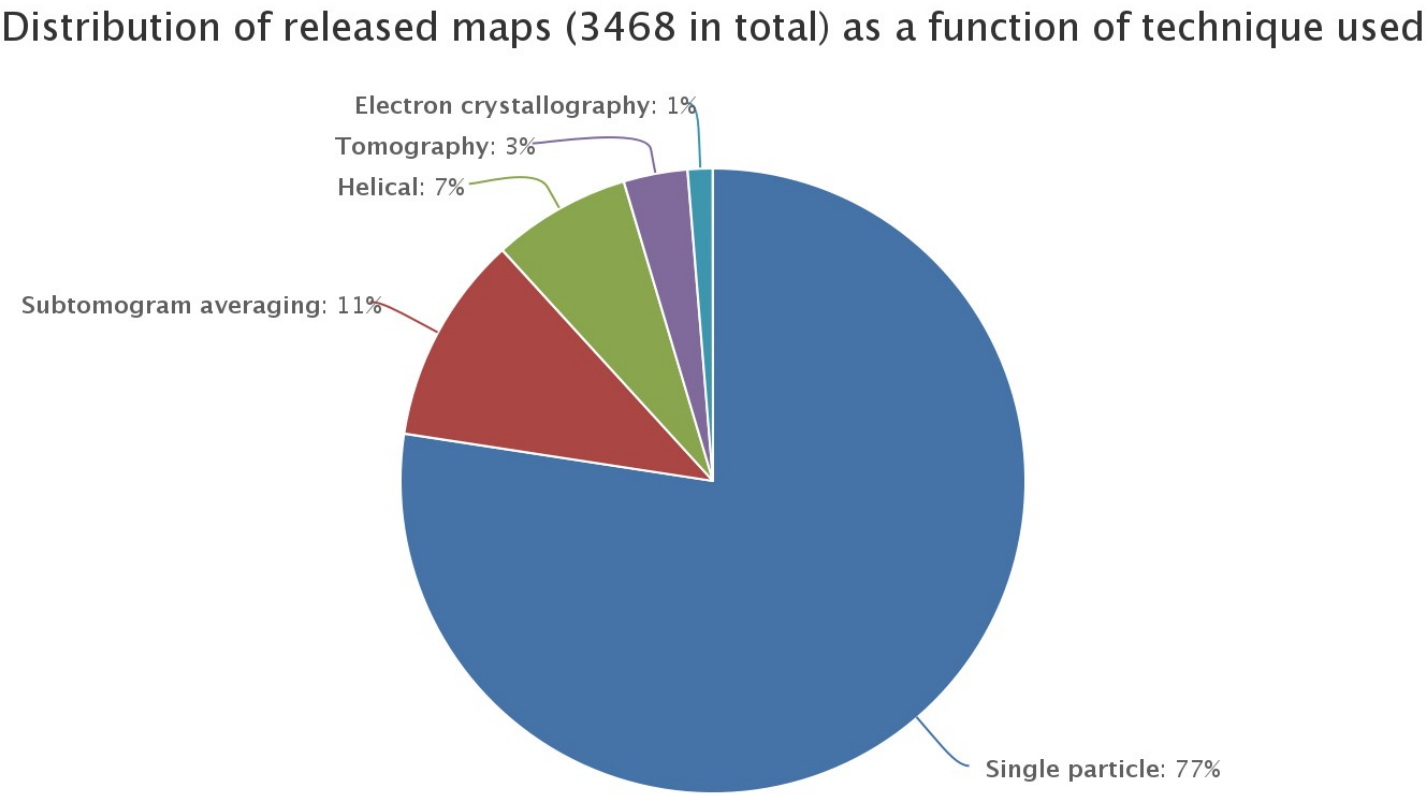
CSIC

instruct Integrating Biology

BIO COMPUTING Unit

instruct image Processing Center

Single-Particles is the most used technique



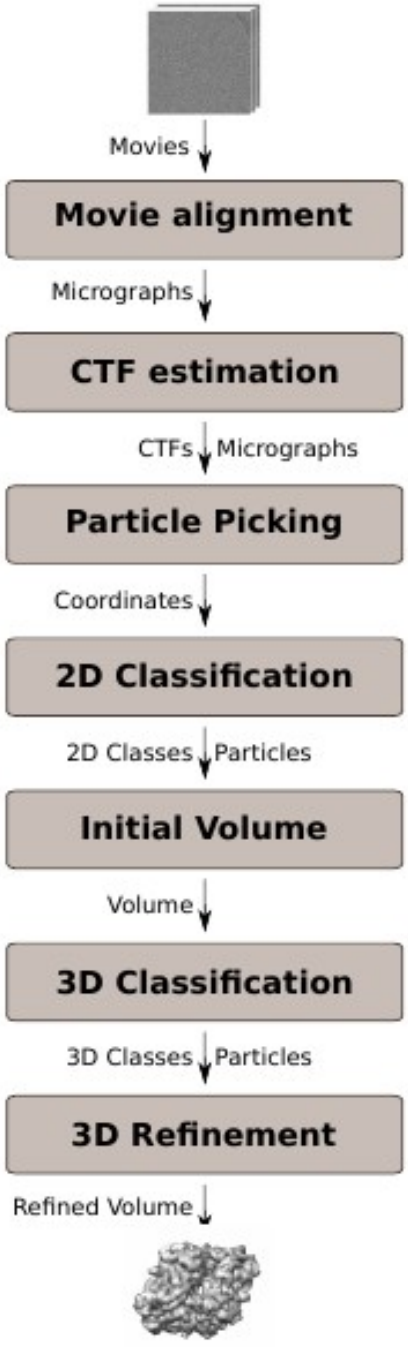
CSIC

instruct
Integrating
Biology

BO Computing Unit

instruct
image
Processing
Center

Single-Particles Workflow: from Movies to 3D Model

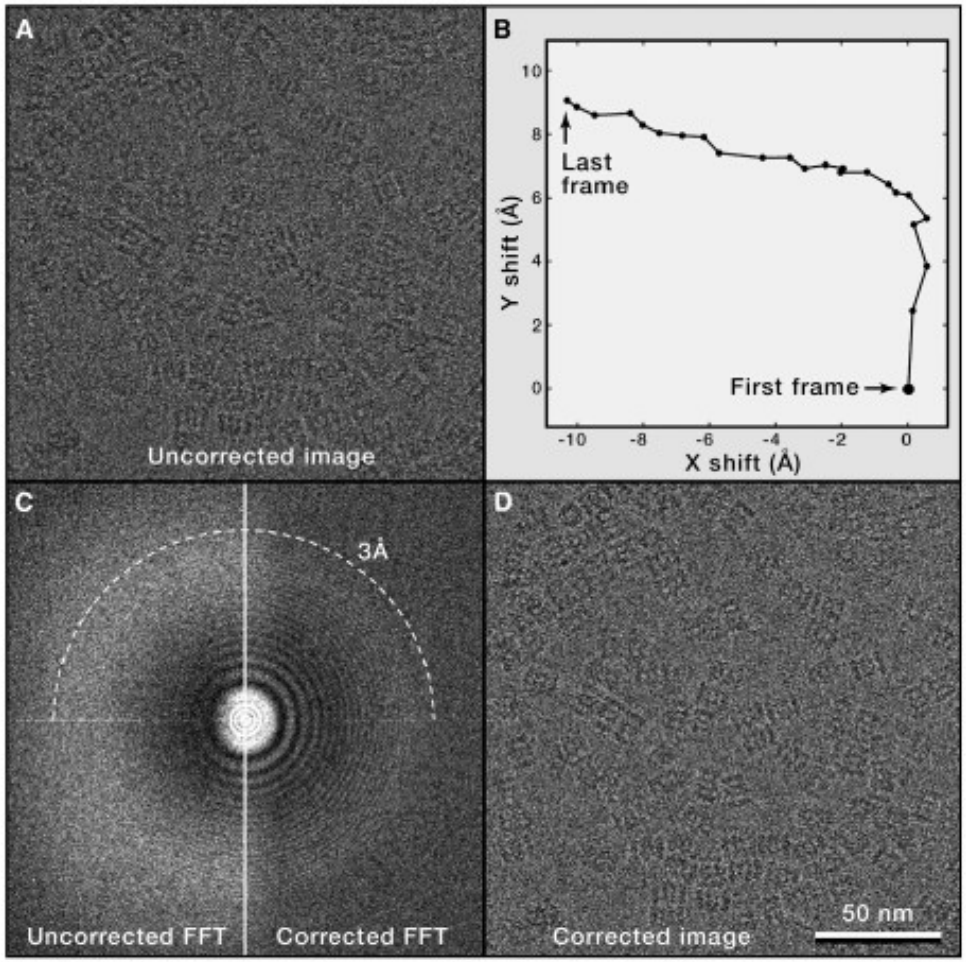


CSIC

instruct Integrating Biology

instruct image Processing Center

Motion correction between movie frames

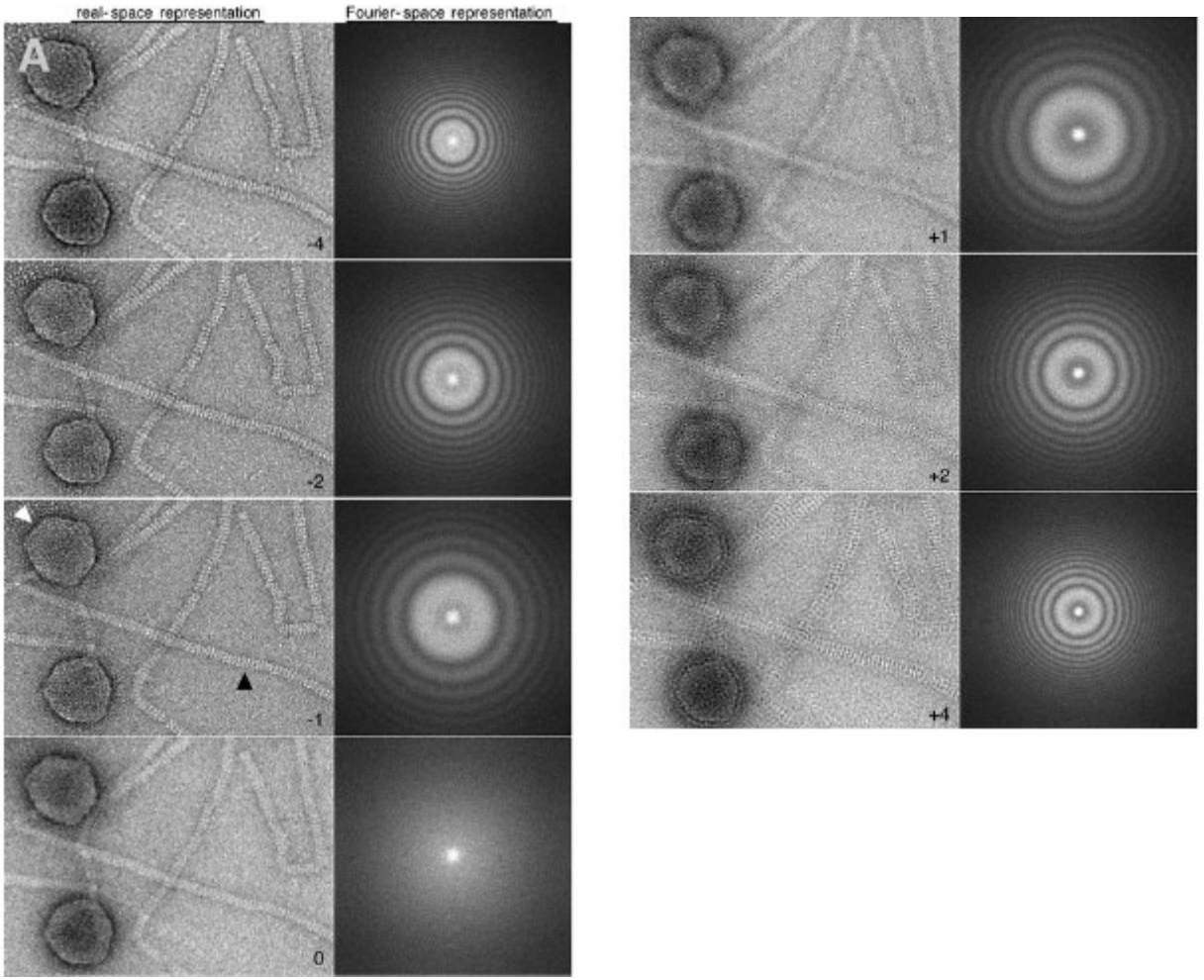


CSIC

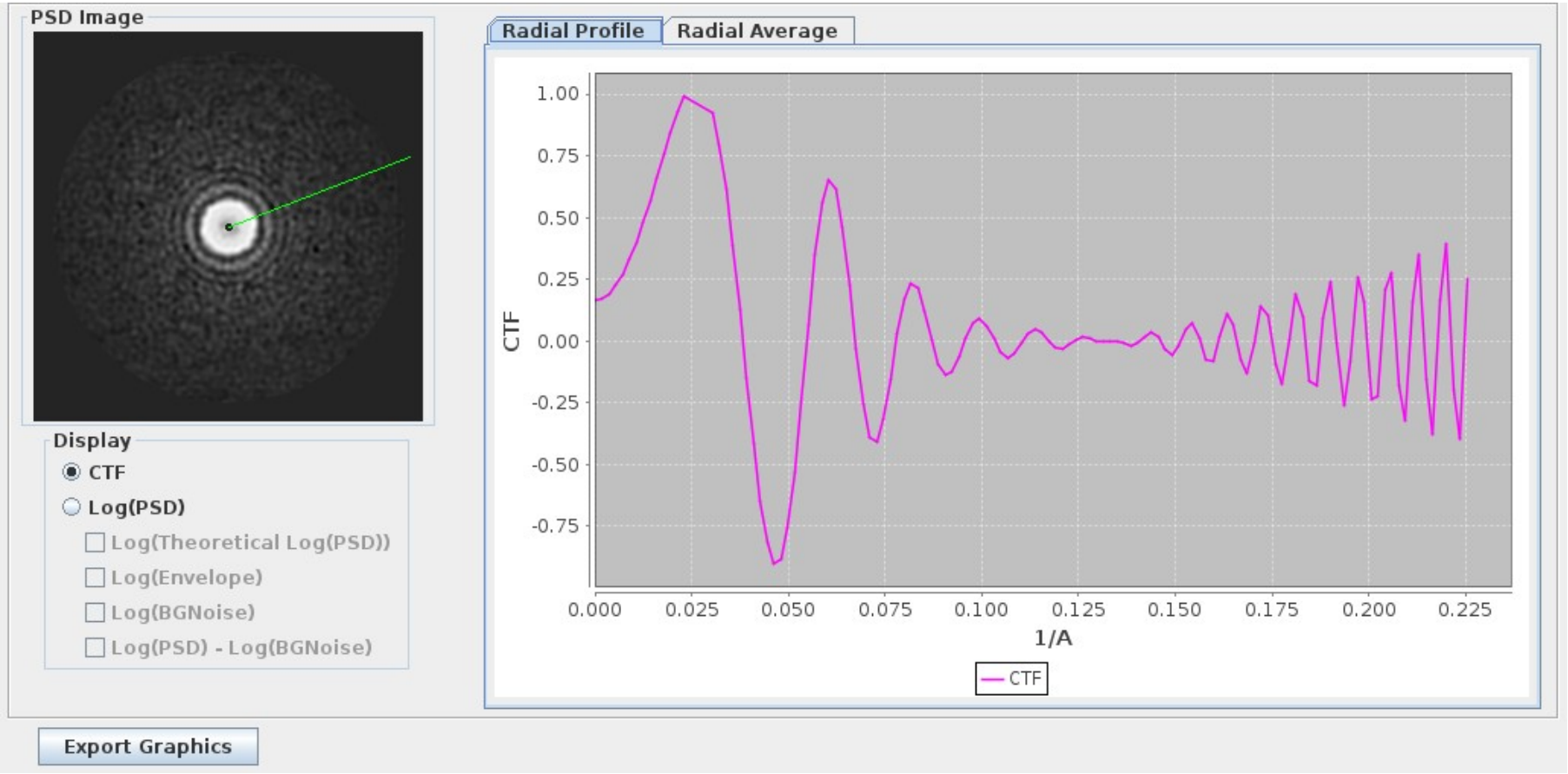
instruct Integrating Biology

instruct image Processing Center

To enhance contrast, images are taken out-of-focus



EM images are affected by the CTF



CSIC

instruct Integrating Biology

BO Computing Unit

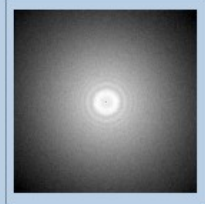
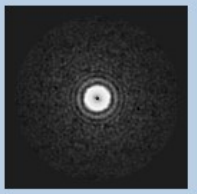
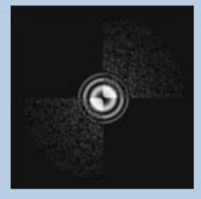
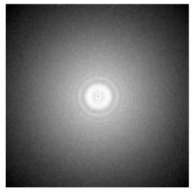
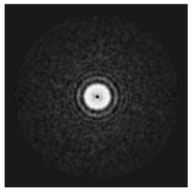
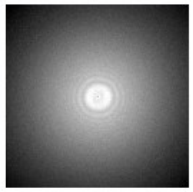
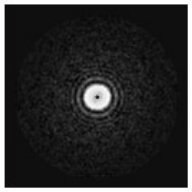
instruct image Processing Center

CTF serves to discard bad micrographs

File Display Tools Help

50 1 Cols 10 Rows 3

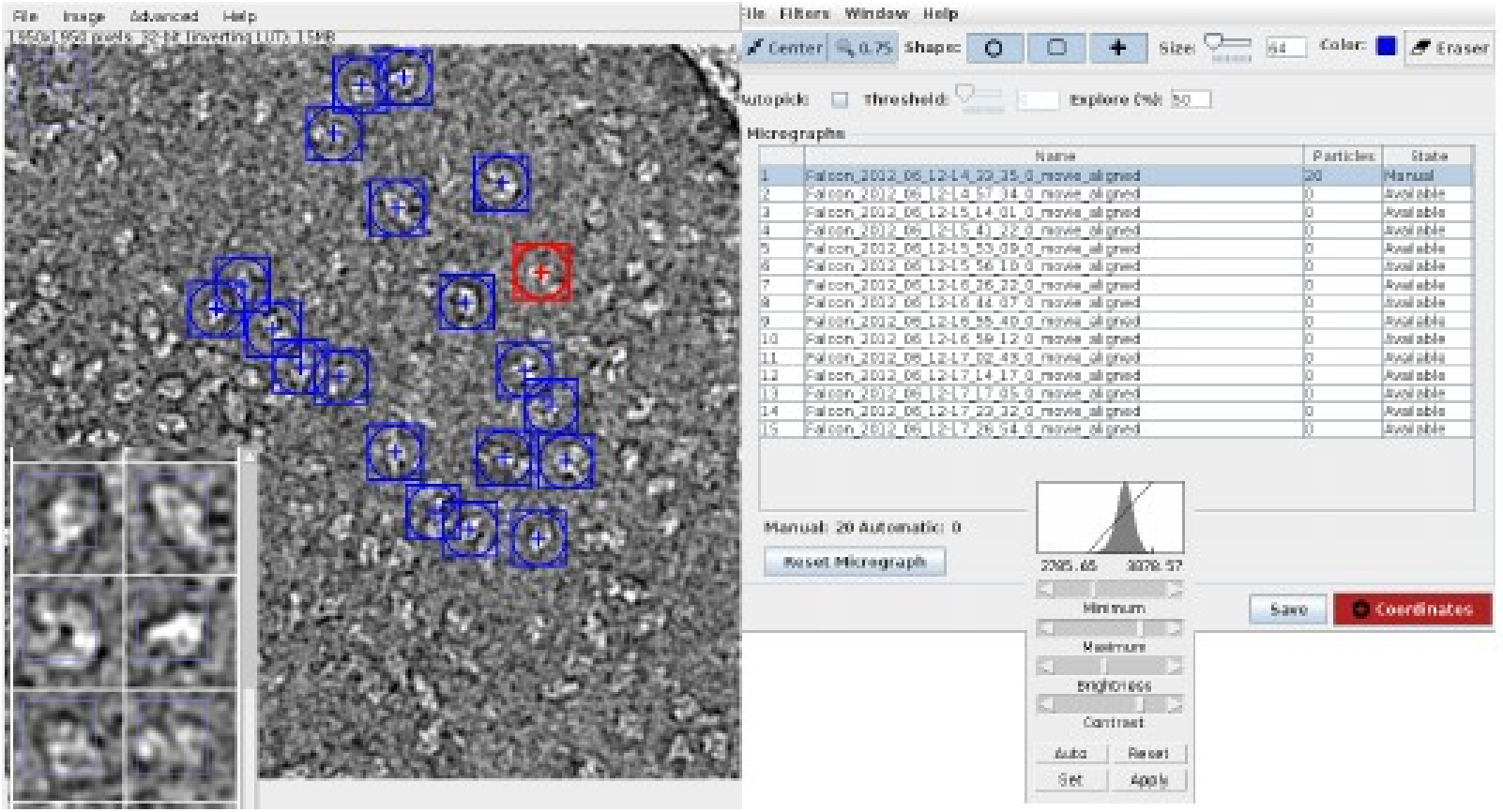
Block CTF

	id	enabled	_psdFile	_xmipp_enhanced_psd	_xmipp_ctfmodel_quadrant	_defocusU	_defocusV	_defocusAngle	_defocusRatio	_m
1	1	☑				33977.3984	33087.3008	63.7310	1.0269	Runs/000002_ProtImportMicrographs/extra/FoilHo
2	2	☑				33782.5000	32638.1992	91.5366	1.0351	Runs/000002_ProtImportMicrographs/extra/FoilHo
3	3	☑				33236.5000	32524.1992	80.1429	1.0219	Runs/000002_ProtImportMicrographs/extra/FoilHo

Close Micrographs Recalculate CTFs

CSIC instruct Integrating Biology BIO COMPUTING Unit instruct image Processing Center

We need to select 'particles' from good micrographs



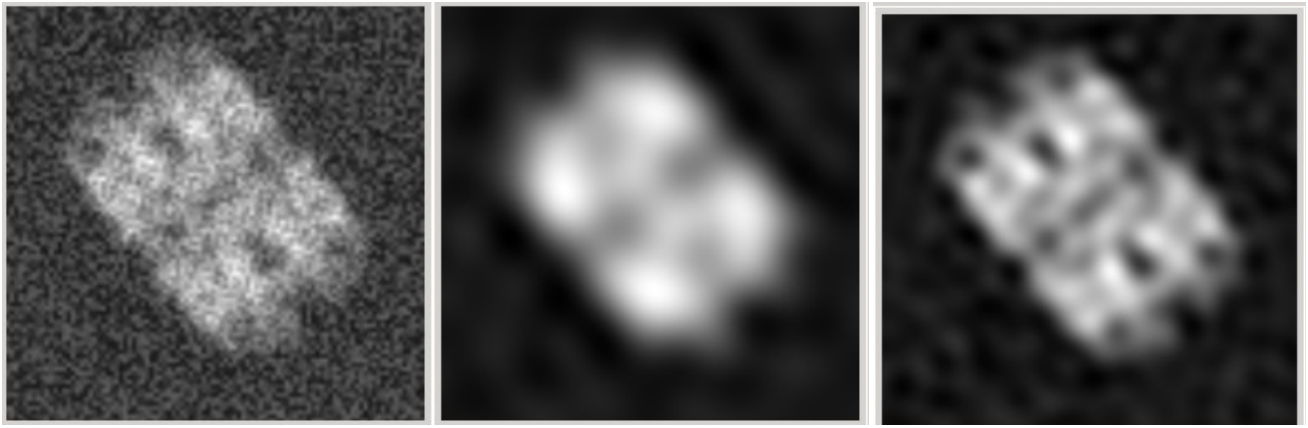
There different types of picking algorithms

1. Manual/Supervised
2. Automatic:
   - Reference free
   - Template-based

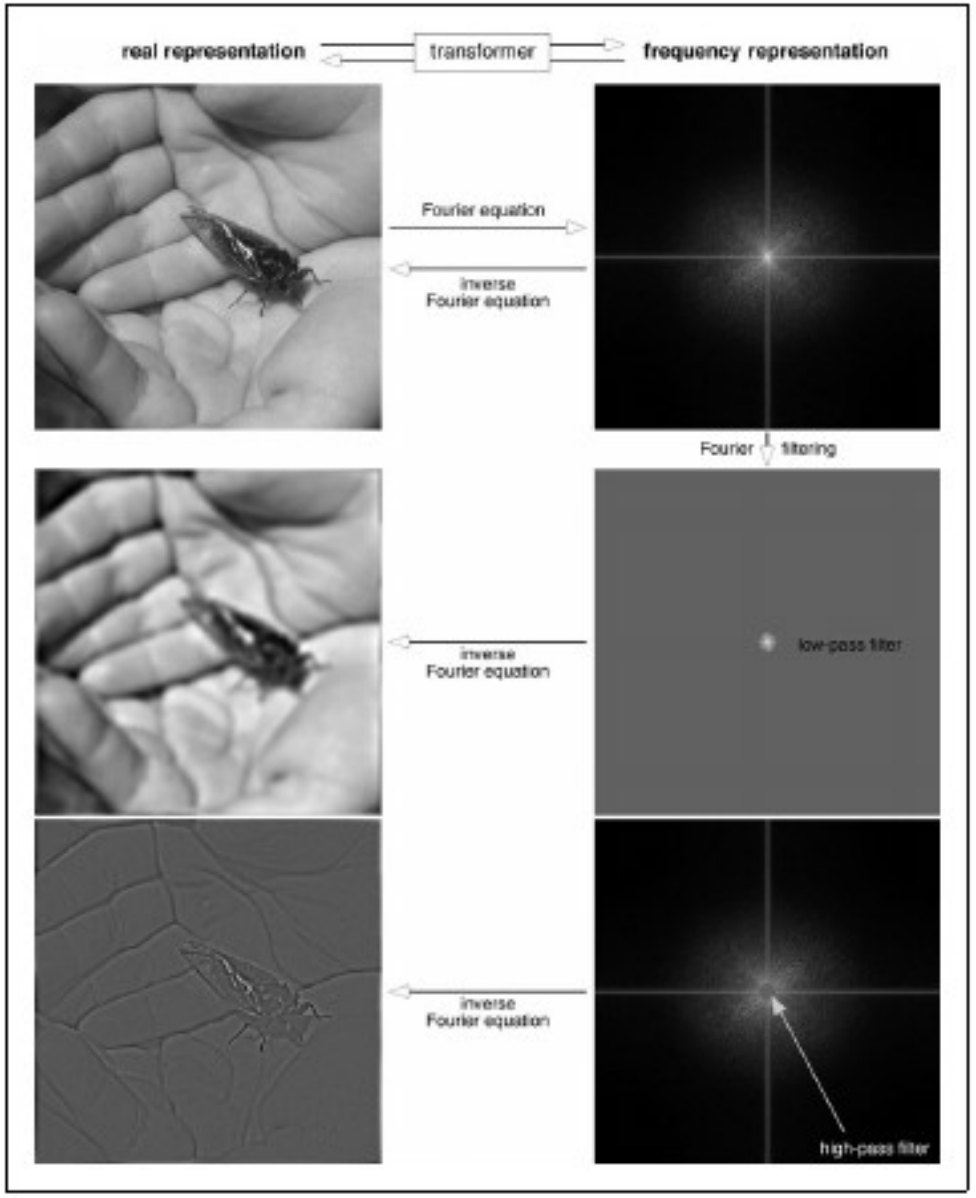
Some pre-processing operations are commonly used

Fourier “Top-hat” low-pass filter

Original image | radius=0.07 | radius=0.13



Fourier-Transforms are at the core of EM image processing

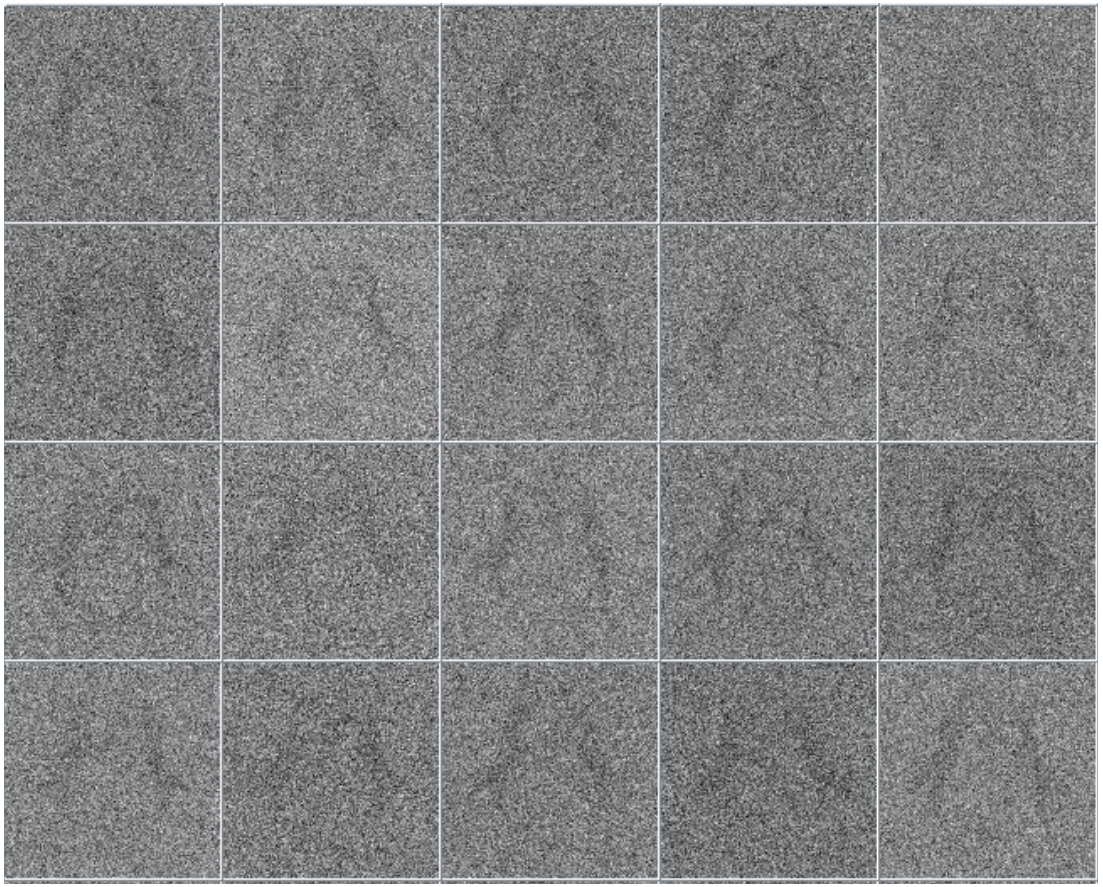


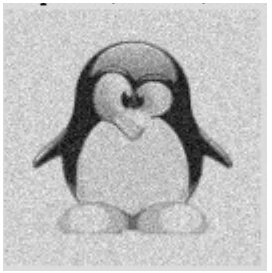
CSIC

instruct Integrating Biology

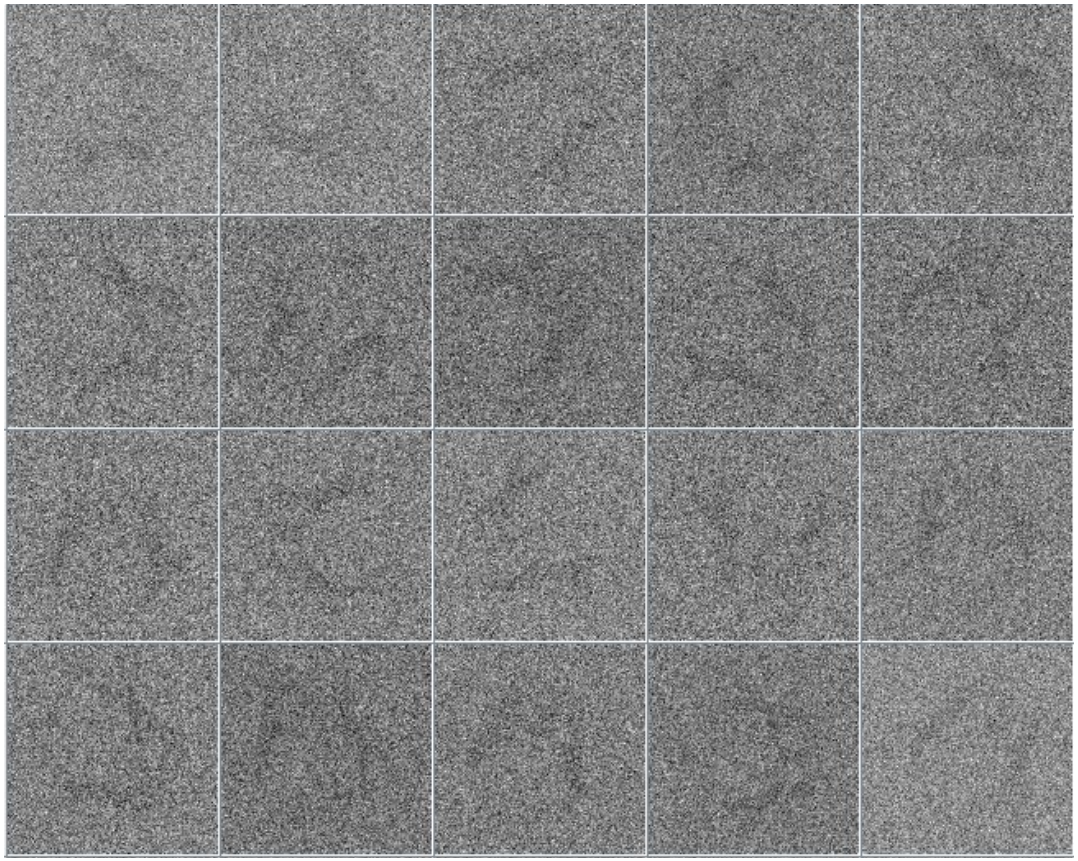
instruct image Processing Center

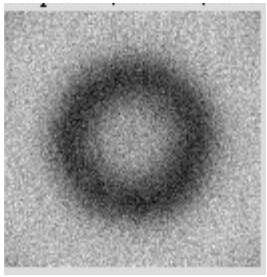
Averaging is required for extracting the signal



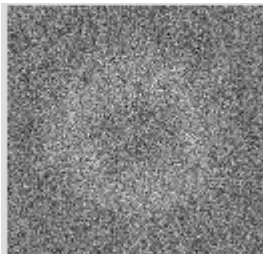


Images need to be aligned before averaging





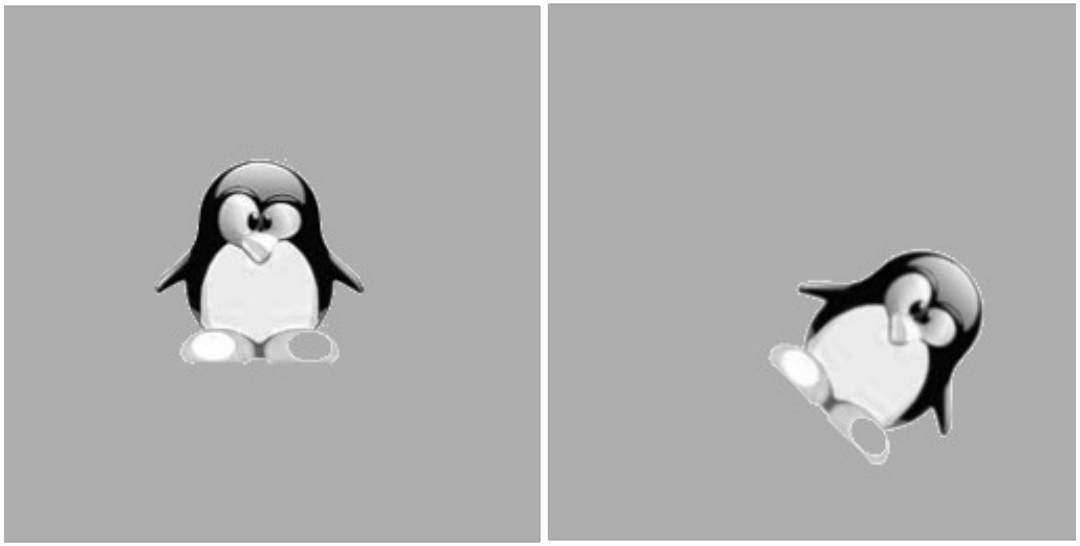
Average image



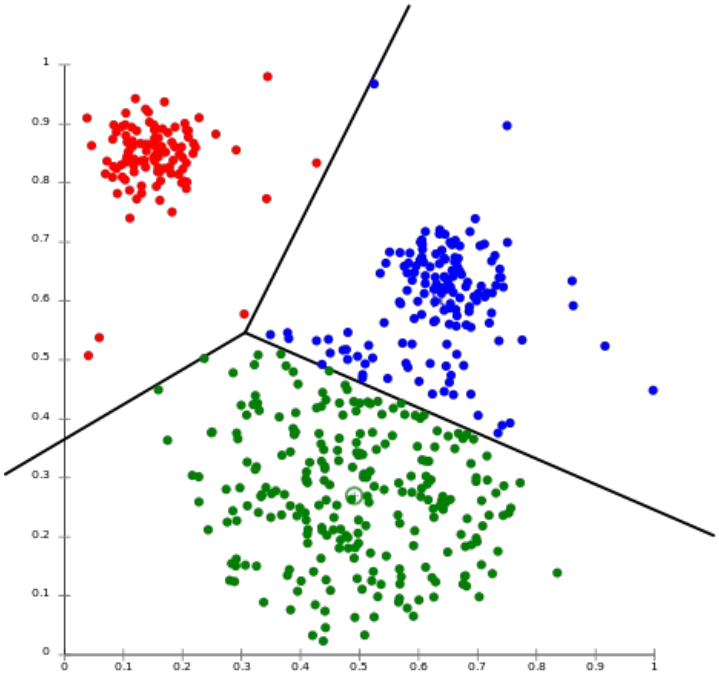
Standard deviation

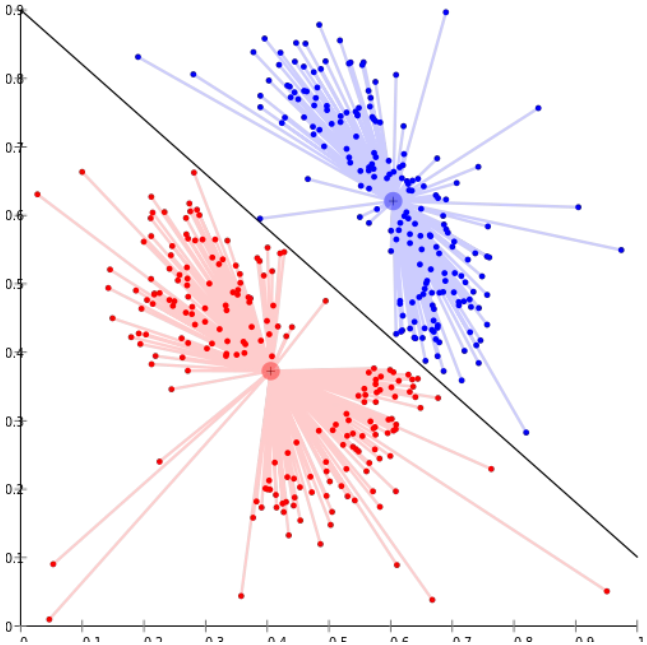
2D alignment search for three free parameters

Three free parameters:
in-plane angle, x and y shifts



Most of 2D classification are based on K-means

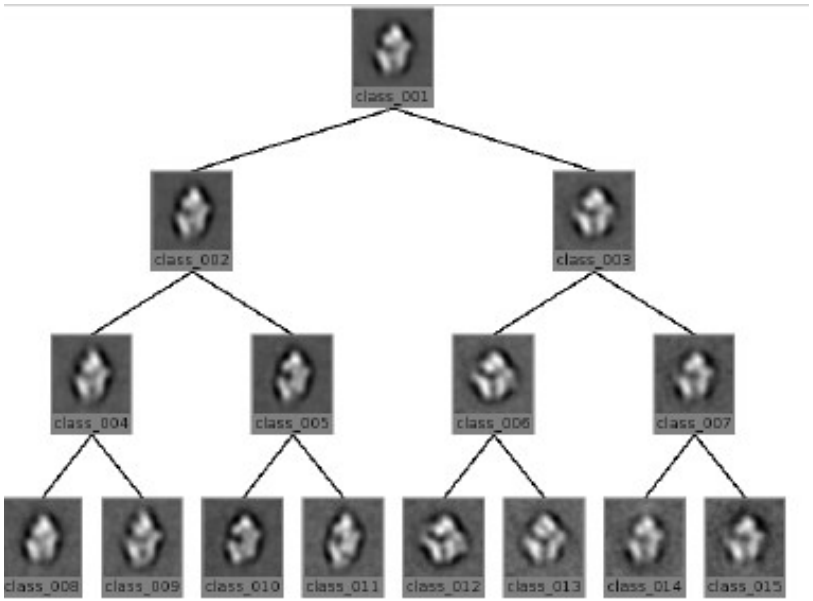


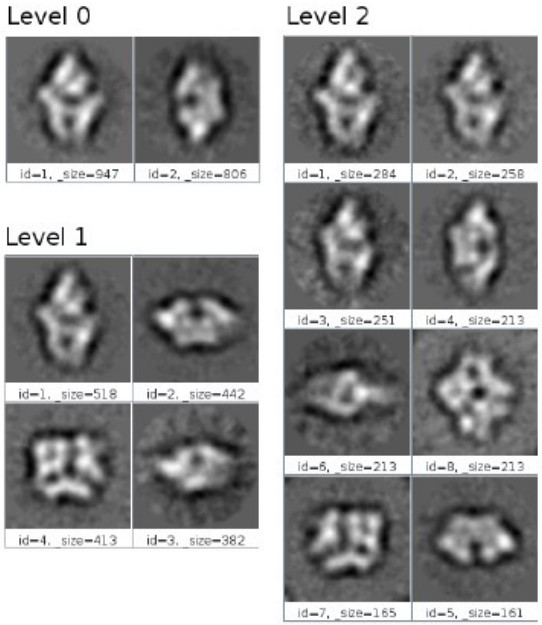


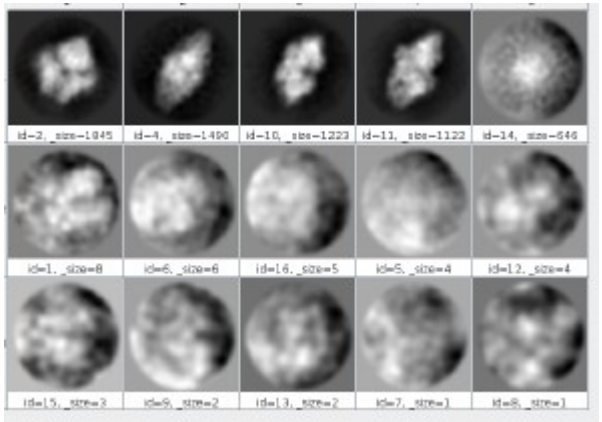
Several 2D classification strategies are implemented in different packages

1. Multivariate Data Analysis (SPIDER/IMAGIC)
2. KerdenSOM (Xmipp)
3. Maximum Likelihood (Xmipp, Relion)
4. Clustering (Xmipp-CL2D and Sparx-ISAC)

Classification results from different algorithms





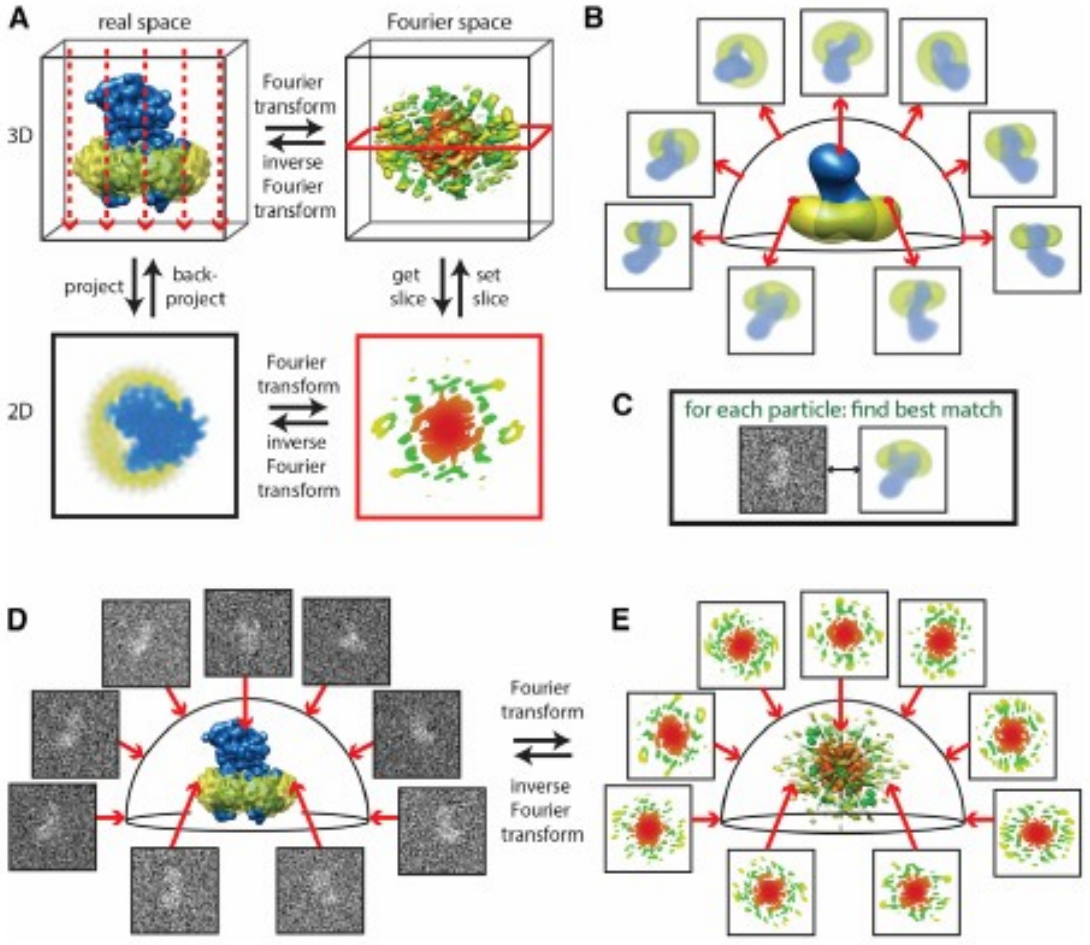


CSIC

instruct Integrating Biology

instruct image Processing Center

Basic concepts of 3D reconstruction

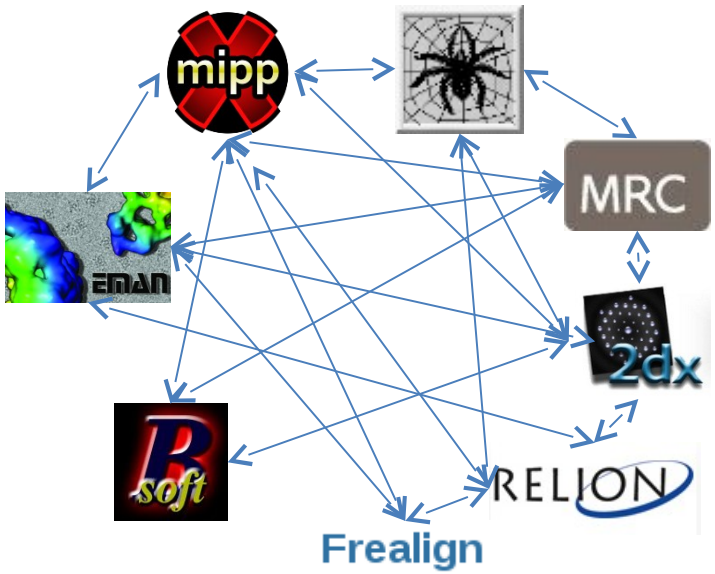


Initial volume problem

1. Random Conical Tilt
2. Angular reconstitution (common lines)
3. Bootstrapping methods
4. Simulating annealing

The EM field needs software integration

Using different EM software packages is now like the tower of Babel



CSIC

instruct Integrating Biology

BIO COMPUTING Unit

instruct image Processing Center

Appion is certainly a pioneer work in terms of software integration (and our main inspiration)

It is increasing the number of external tools added to current EM packages (such as Eman2, Xmipp3 or Relion)

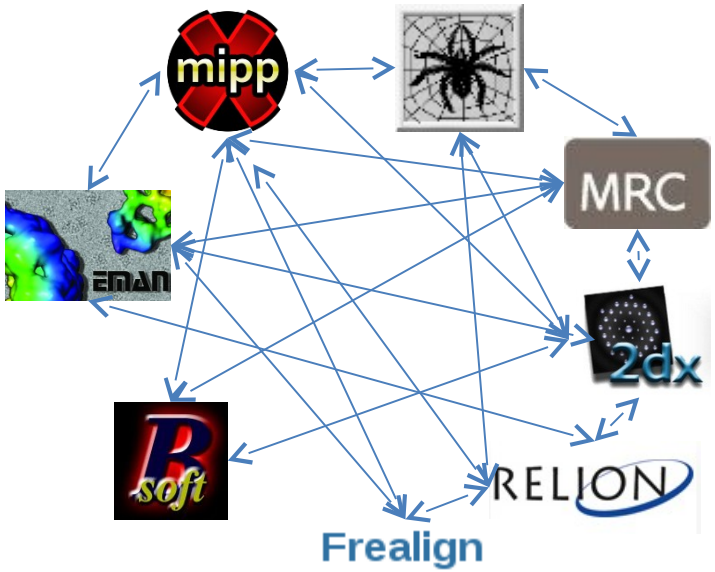
But still is complicated to smoothly combine different software tools in a consistent manner.

Scipion Principles

1. Integrate EM software packages to be used in the same project.
2. Full project traceability, improving reproducibility.
3. Execute complete workflows in an automated manner.
4. Easy to install and use.
5. Easy to extend with new protocols.

Goal 1: Integrate EM software packages to be used in the same project.

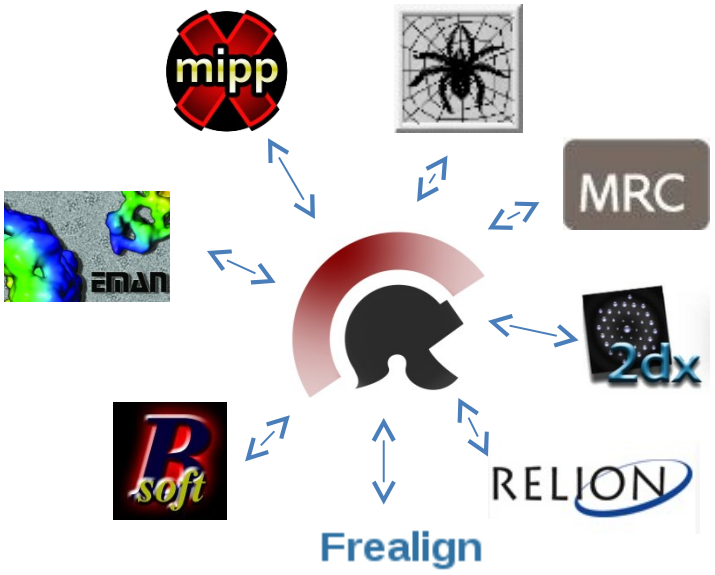
All vs All is hard to maintain and extend



All conversions: N*N

New package: 2*N

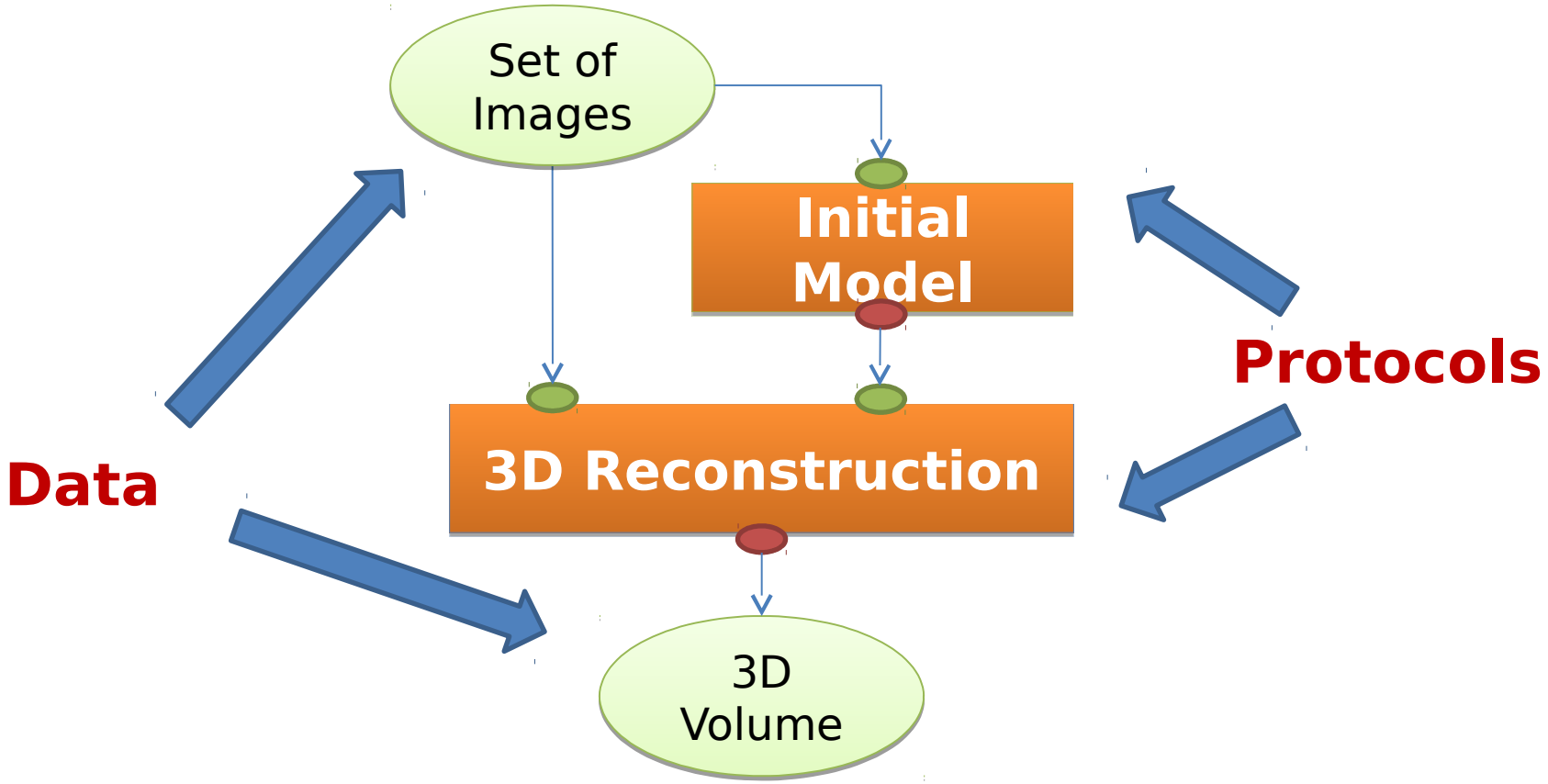
It is better to have a common format



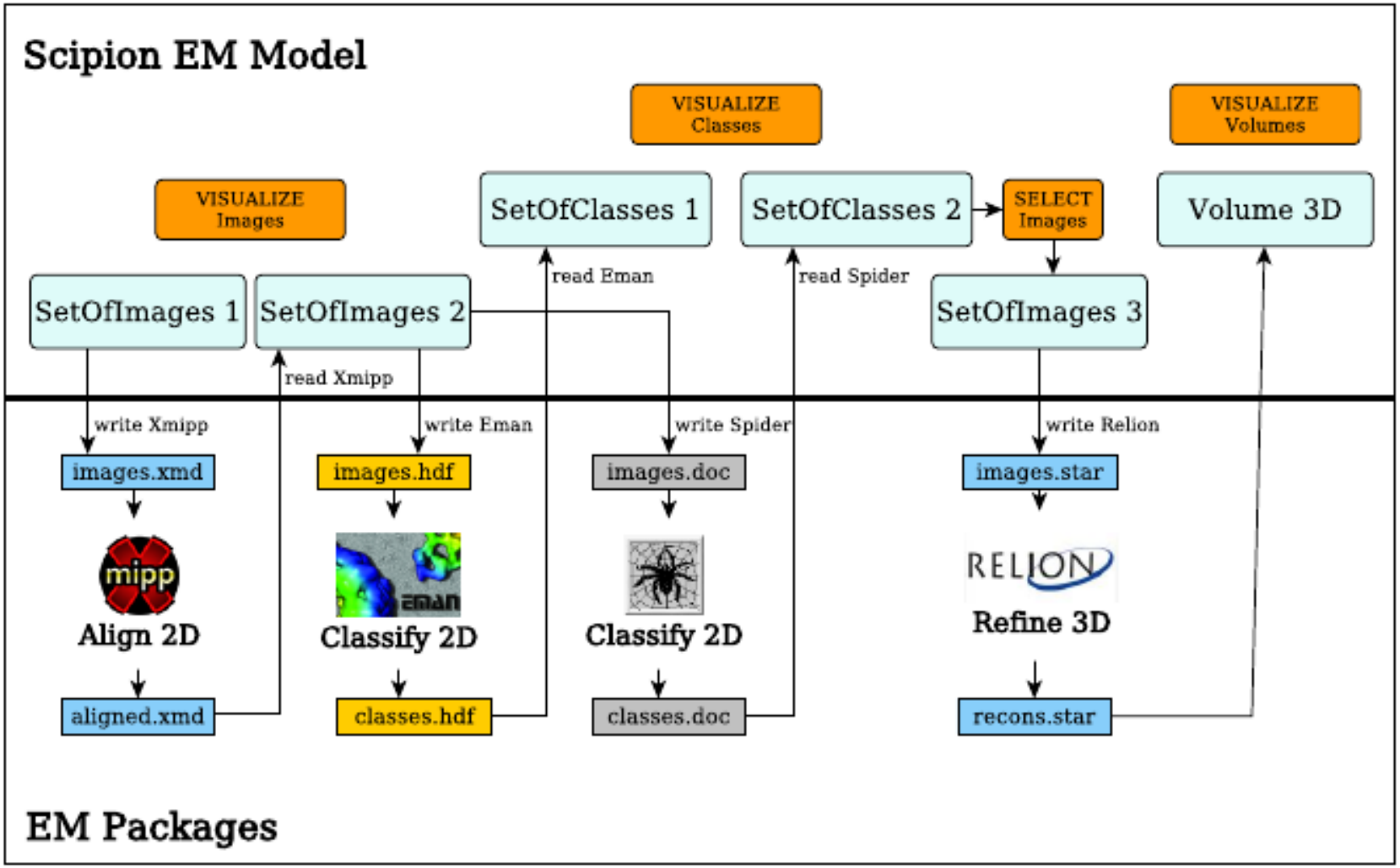
All conversions: N+N

New package: 2

We created a model of the EM domain

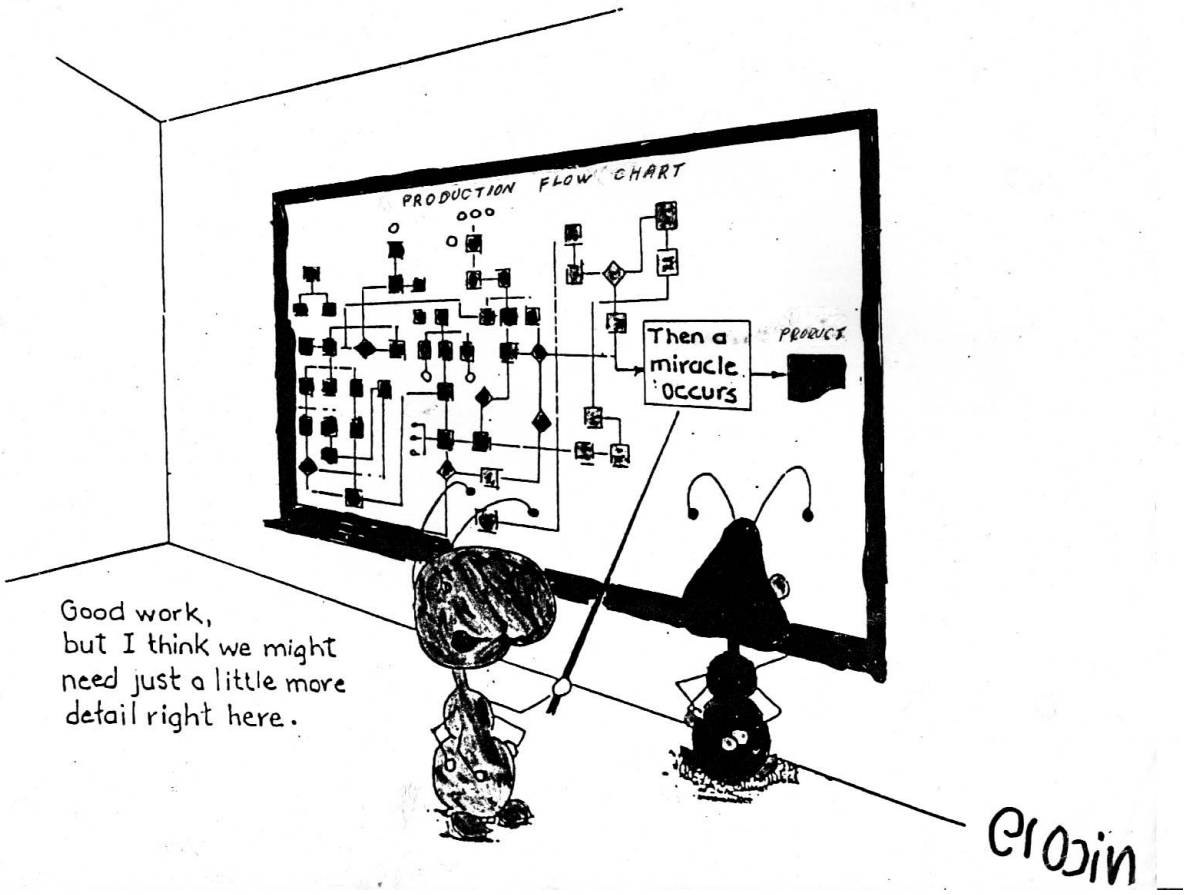


We need conversion functions for each package



Goal 2: Full project traceability, improving reproducibility.

Results should be reproducible, no 'black-boxes'

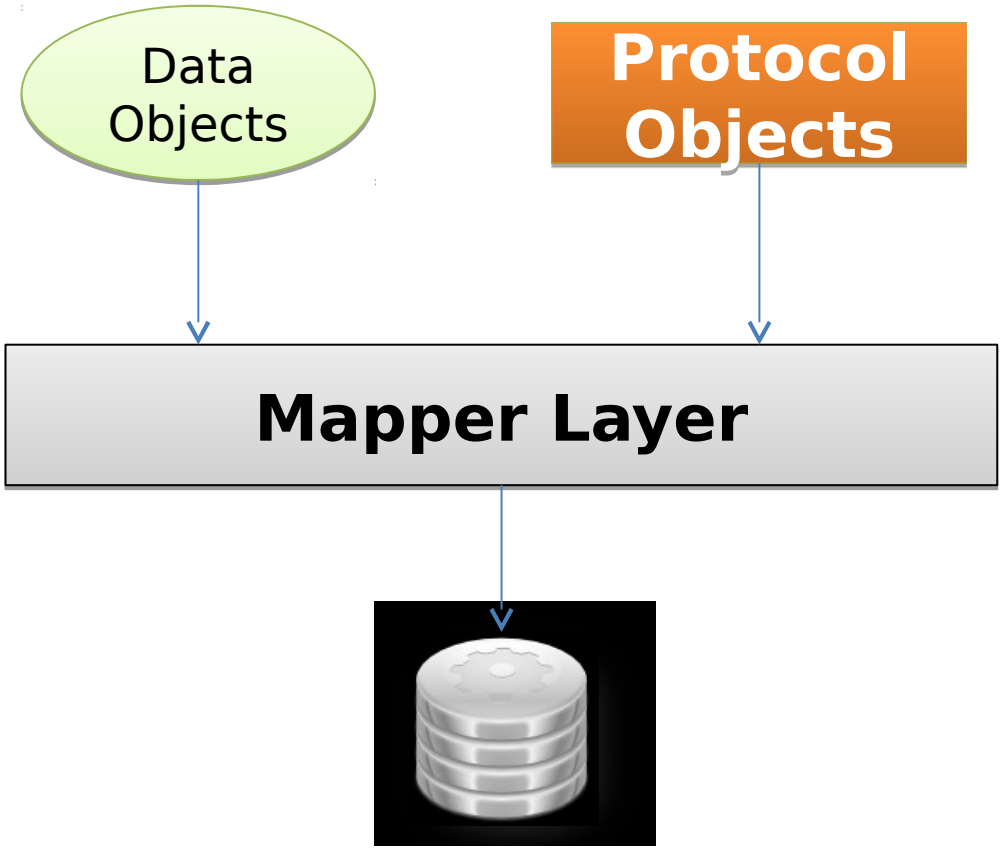


CSIC

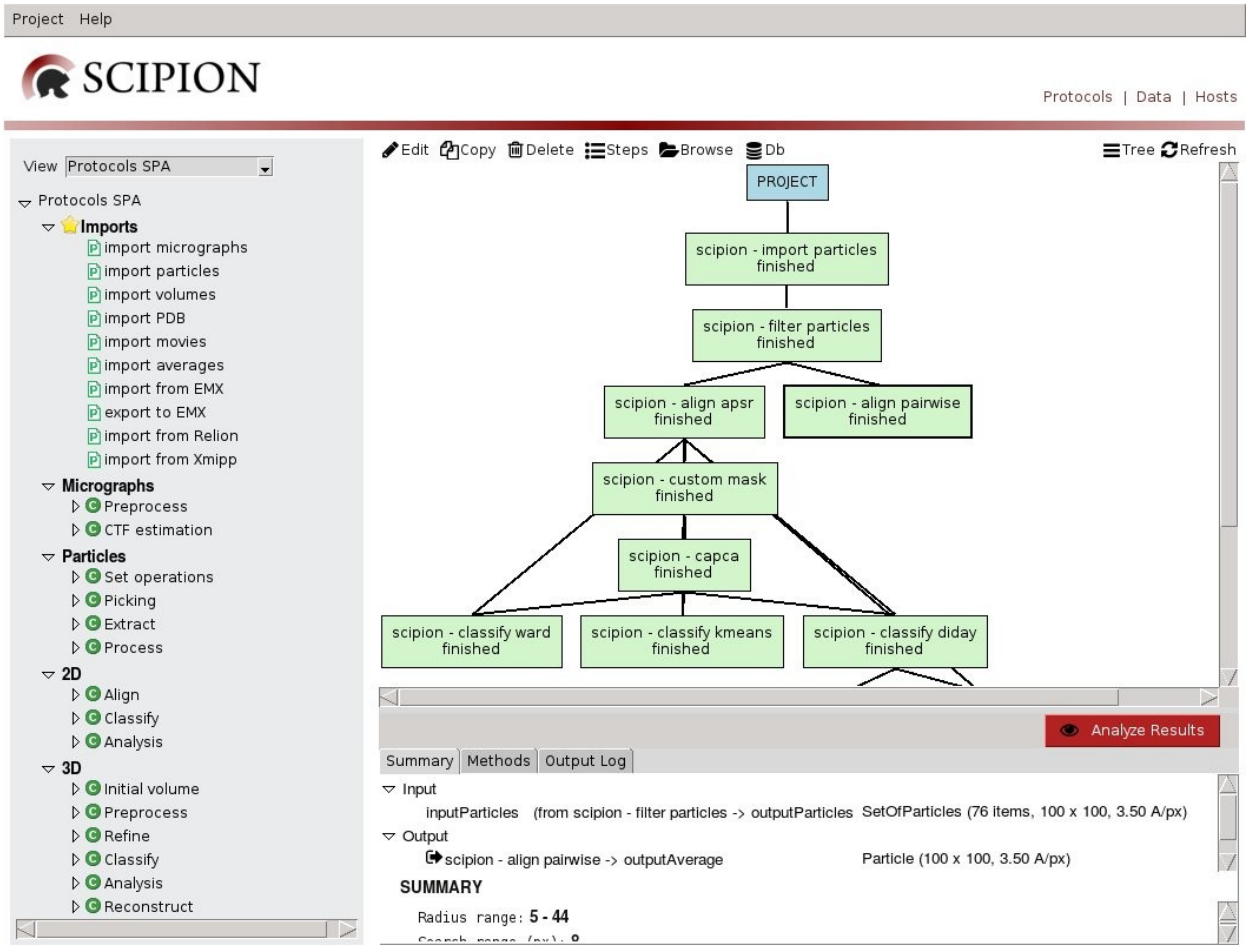
instruct Integrating Biology

instruct image Processing Center

We implemented a simple storage mechanism



We track all the steps performed in a project



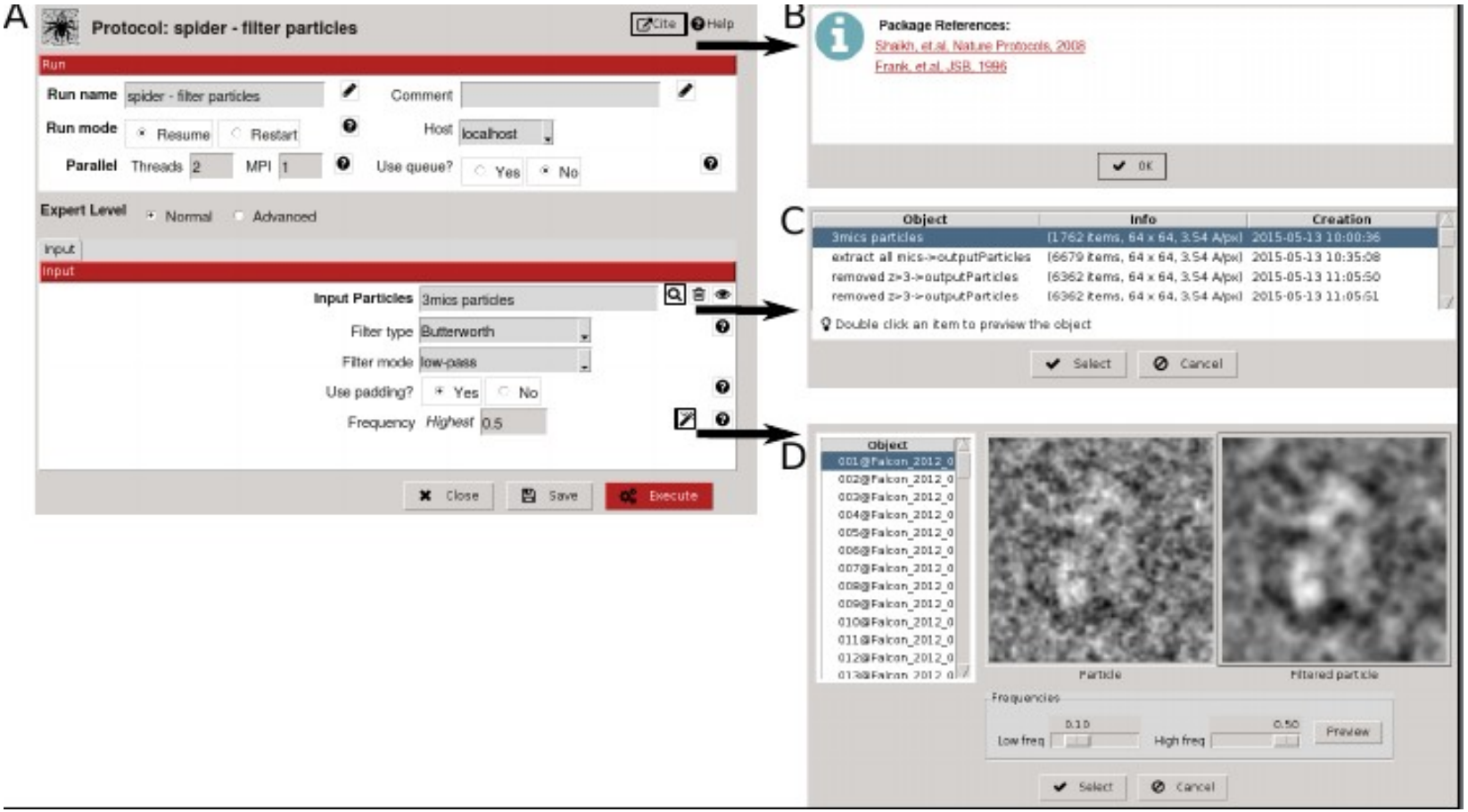
CSIC

instruct Integrating Biology

BIO COMPUTING Unit

instruct image Processing Center

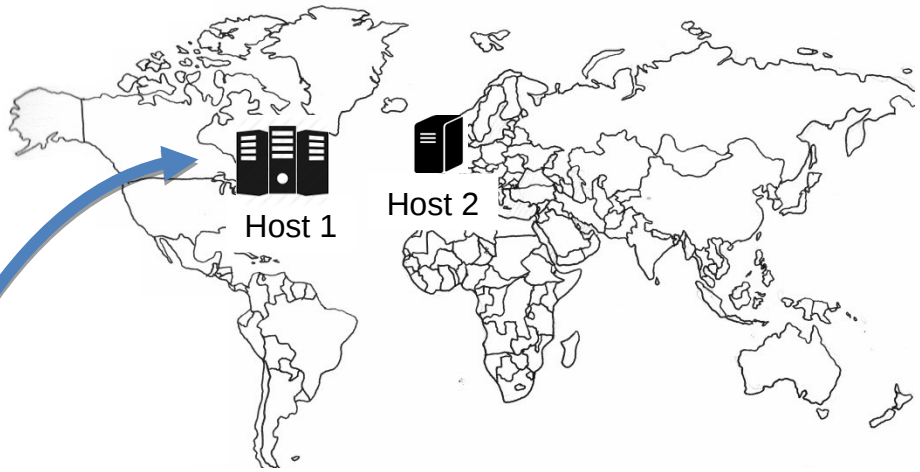
All parameters are also stored



Goal 3: Execute complete workflows in an automated manner.

Designed to perform distributed execution

Scipion Client



Big data transfers

Scipion Server

CSIC

instruct
Integrating
Biology

BIO COMPUTING Unit

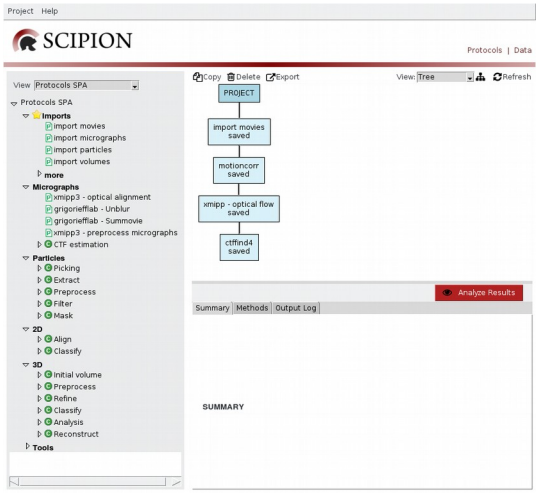
instruct
image
Processing
Center

Run workflows automatically and in streaming

Microscope

Scipion Box

- Import Movies
- Movie Alignment
- CTF Estimation
- Particle Picking ???



CSIC

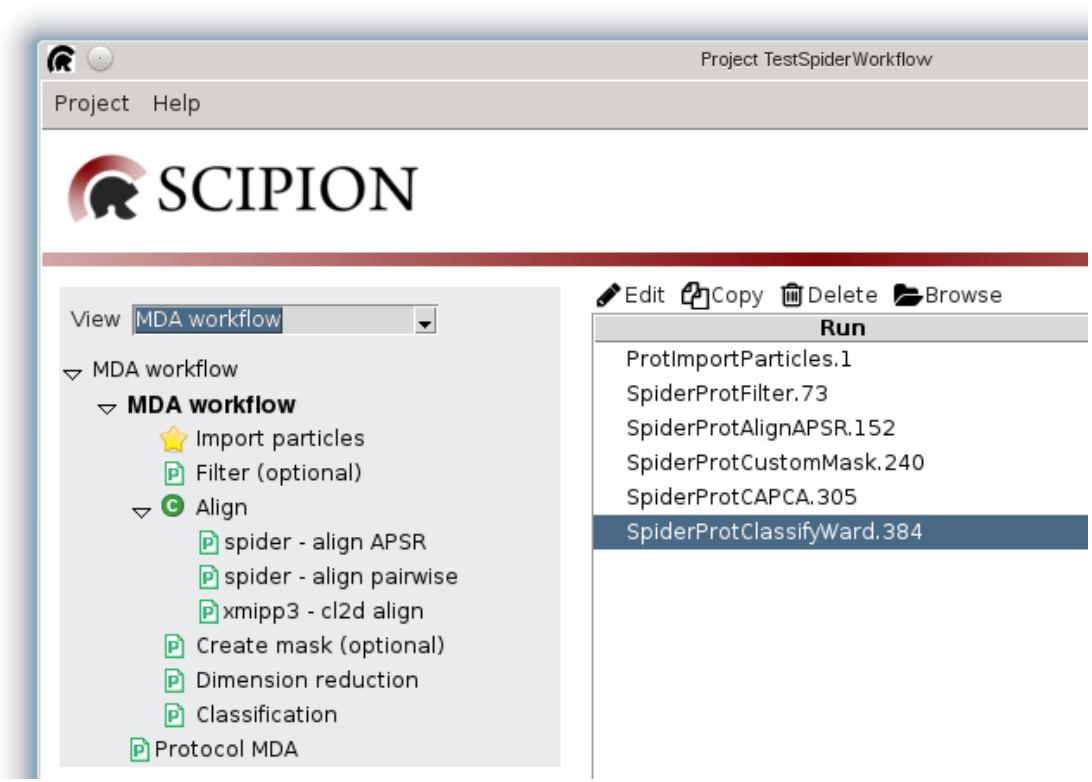
instruct Integrating Biology

BIO COMPUTING Unit

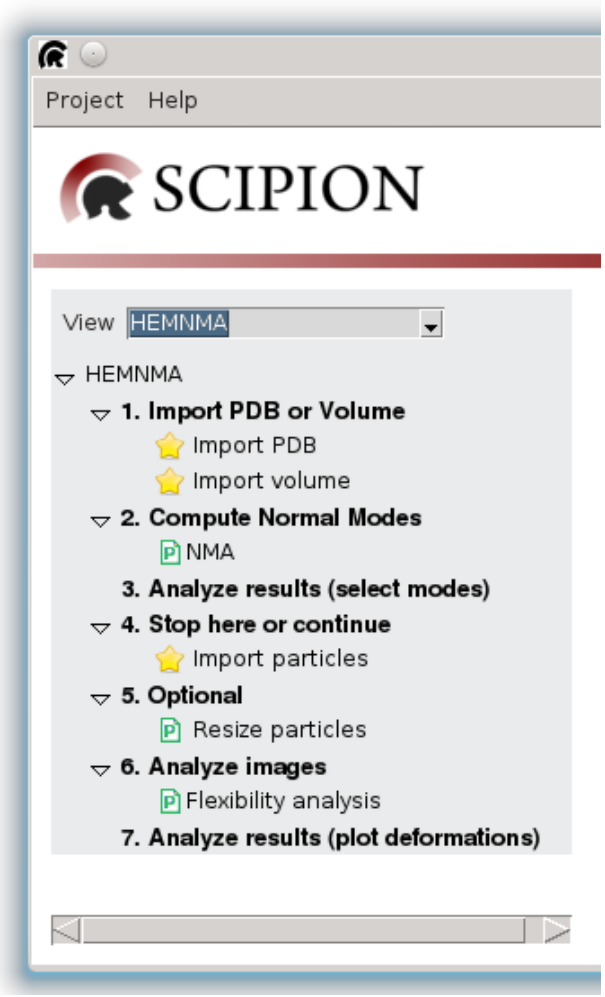
instruct image Processing Center

Goal 5: Easy to extend with new methods

Example 1: Integration of Spider-MDA (in collaboration with Tanvir Shaikh)



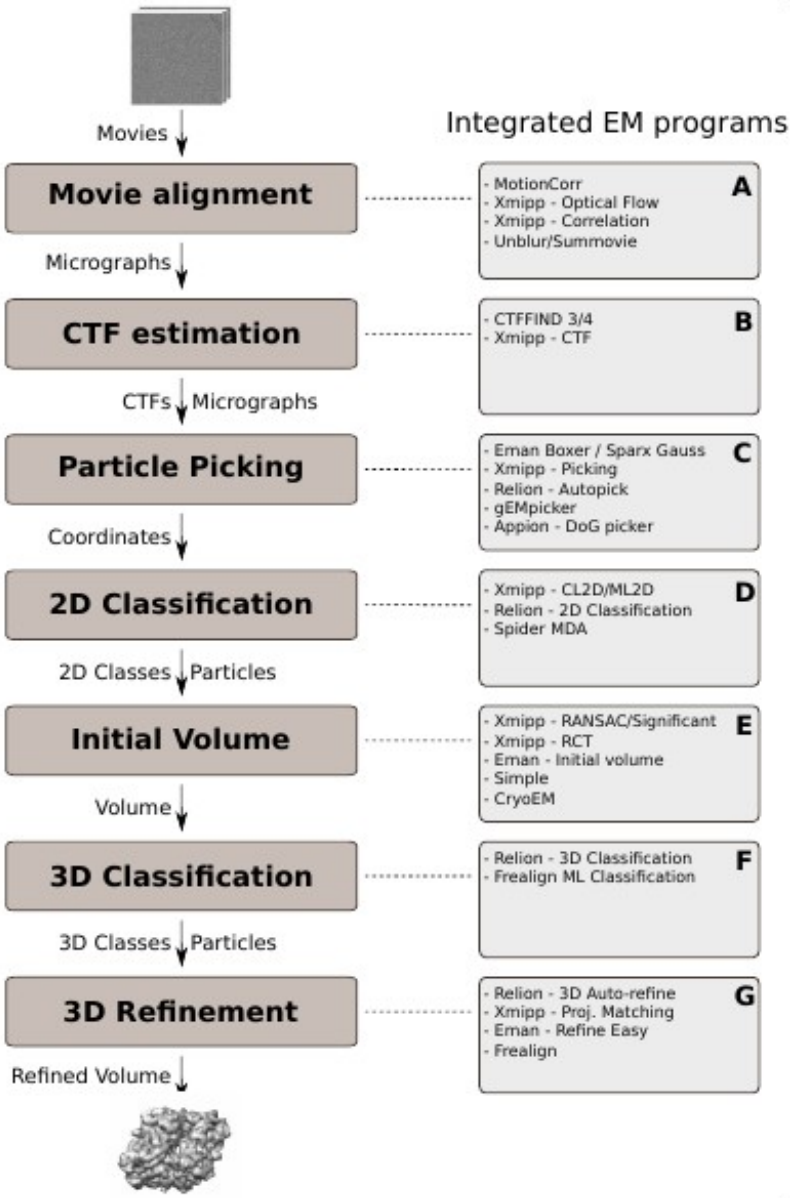
Example 2: Integration of Normal Modes analysis and flexible fitting (in collaboration with Slavica Jonic)



Example 3: Integration of ResMap

(in collaboration with Alp Kucukelbir and Hemant Tagare)

SPA workflow and Scipion integrated programs



CSIC

instruct Integrating Biology

BIO COMPUTING Unit

instruct image Processing Center

instruct
Integrating
Biology

www.structuralbiology.eu
Follow us on twitter
@instructhub

CSIC

BIO COMPUTING Unit