# CryoEM DataModels for processing

Biocomputing Unit, CNB-CSIC

Instruct Image Processing Center

# Data Models Use

- Public repositories:
  - EMDB: https://www.ebi.ac.uk/emdb/documentation#data_model
  - EMPIAR: https://www.ebi.ac.uk/empiar/faq
- Facilities and LIMS:
  - ISpyB: https://github.com/ispyb/ispyb-database-modeling
  - EMAdmin: https://github.com/I2PC/EMadmin/tree/master/EMadmin
- Data processing workflows:
  - Scipion: https://github.com/scipion-em/scipion-em/blob/devel/pwem/objects/data.py
  - (CryoSparc, Relion, …)

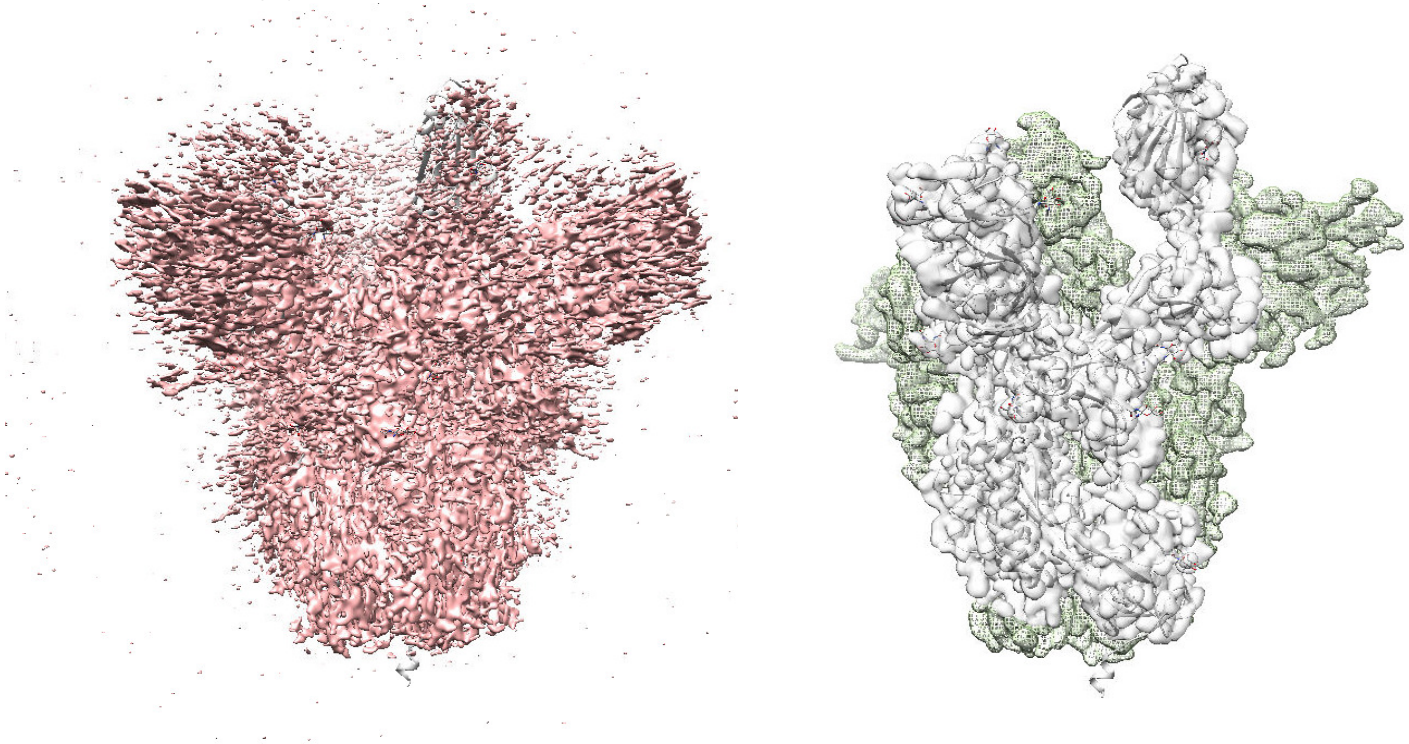
# Data models use

# Data Models Workflows

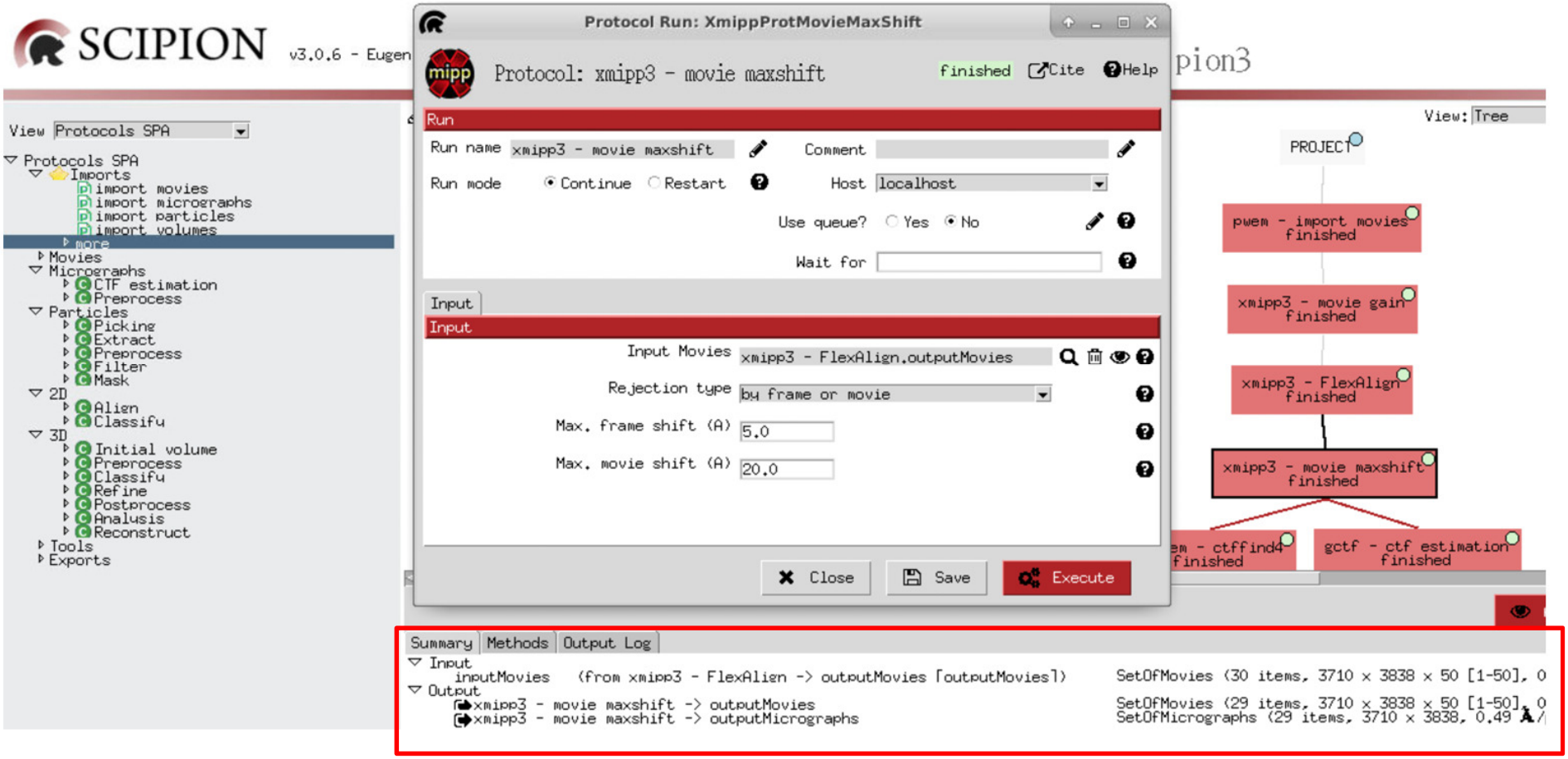
We should distinguish the purpose:

- To run on a new dataset (WorkflowHub)
- To understand a given result (EMPIAR)

**EMDB 22301**



# Workflows


SCIPION
v3.0.6 - Eugen
View Protocols SPA
Protocols SPA
Imports
import movies
import micrographs
import particles
import volumes
more
Movies
Micrographs
CTF estimation
Preprocess
Particles
Picking
Extract
Preprocess
Filter
Mask
2D
Align
Classify
3D
Initial volume
Preprocess
Classify
Refine
Postprocess
Analysis
Reconstruct
Tools
Exports
Protocol Run: XmippProtMovieMaxShift
Protocol: xmipp3 - movie maxshift
finished
Cite
Help
Run
Run name xmipp3 - movie maxshift
Comment
Run mode Continue Restart
Host localhost
Use queue? Yes No
Wait for
Input
Input
Input Movies xmipp3 - FlexAlign.outputMovies
Rejection type by frame or movie
Max. frame shift (A) 5.0
Max. movie shift (A) 20.0
Close
Save
Execute
View: Tree
PROJECT
pwem - import movies finished
xmipp3 - movie gain finished
xmipp3 - FlexAlign finished
xmipp3 - movie maxshift finished
gctf - ctf estimation finished
Summary
Methods
Output Log
Input
inputMovies (from xmipp3 - FlexAlign -> outputMovies [outputMovies])
SetOfMovies (30 items, 3710 x 3838 x 50 [1-50], 0
Output
xmipp3 - movie maxshift -> outputMovies
SetOfMovies (29 items, 3710 x 3838 x 50 [1-50], 0
xmipp3 - movie maxshift -> outputMicrographs
SetOfMicrographs (29 items, 3710 x 3838, 0.49 Å/


SetOfMovies
Movie MaxShift
SetOfMovies
SetOfMicrographs

# Data Models

scipion-em/pwem/objects/data

github.com/scipion-em/scipion-em/blob/devel/pwem/objects/data.py#L2324

Files

devel

Go to file

- .github
- pwem
  - cmd
  - convert
  - emlib
  - objects
    - __init__.py
    - data.py
    - data_tiltpairs.py
  - protocols
  - templates
  - tests
  - viewers
  - wizards
  - __init__.py
  - bibtex.py
  - constants.py
  - protocols.conf

scipion-em / pwem / objects / data.py

Top

Code Blame 2632 lines (2078 loc) · 86.5 KB

Raw

```python
class ClassVol(SetOfVolumes):
    """ Represent a Class that groups Volume objects.
    Usually the representative of the class is another Volume.
    """

    def close(self):
        # Do nothing on close, since the db will be closed by SetOfClasses
        pass


class SetOfClasses(EMSet):
    """ Store results from a classification. """
    ITEM_TYPE = None  # type of classes stored in the set
    REP_TYPE = None  # type of the representatives of each class

    def __init__(self, **kwargs):
        EMSet.__init__(self, **kwargs)
        # Store the average images of each class(SetOfParticles)
        self._representatives = Boolean(False)
        self._imagesPointer = Pointer()

    def copyInfo(self, other):
        """ Copy basic information from other set of classes to current one"""
        self.copyAttributes(other, '_representatives', '_imagesPointer')

    def iterClassItems(self, iterDisabled=False):
        """ Iterate over the images of a class.
        Params:
            iterDisabled: If True, also include the disabled items. """
        for cls in self.iterItems():
            if iterDisabled or cls.isEnabled():
```

Symbols

Find definitions and references for functions and other symbols in this file by clicking a symbol below or in the code.

Filter symbols

- class EMSet
- class SetOfImages
- class SetOfMicrographsBase
- class SetOfMicrographs
- class SetOfParticles
- class SetOfAverages
- class SetOfVolumes
- class SetOfCTF
- class SetOfDefocusGroup
- class SetOfAtomStructs
- class SetOfPDBs
- class SetOfSequences
- class Coordinate
- class SetOfCoordinates
- class Matrix
- class Class2D

# Data Models

scipion-em-xmipp / xmipp3 / protocols / protocol_movie_max_shift.py ↑ Top

Code Blame 391 lines (337 loc) · 17.3 KB Raw

```python
class XmippProtMovieMaxShift(ProtProcessMovies):
    def __init__(self, **args):

    # ------------------------- DEFINE param functions ------------------------
    def _defineParams(self, form):
        form.addSection(label=Message.LABEL_INPUT)
        form.addParam('inputMovies', PointerParam, important=True,
                      label=Message.LABEL_INPUT_MOVS,
                      pointerClass='SetOfMovies',
                      help='Select a set of previously aligned Movies.')

        form.addParam('rejType', params.EnumParam, choices=self.REJ_TYPES,
                      label='Rejection type', default=self.REJ_OR,
                      help='Rejection criteria:\n'
                           ' - *by frame*: Rejects movies with drifts between '
                                   'frames bigger than a certain maximum.\n'
                           ' - *by whole movie*: Rejects movies with a total '
                                   'travel bigger than a certain maximmum.\n'
                           ' - *by frame and movie*: Rejects movies if both '
                                   'conditions above are met.\n'
                           ' - *by frame or movie*: Rejects movies if one of '
                                   'the conditions above are met.')

        form.addParam('maxFrameShift', params.FloatParam, default=5,
                      label='Max. frame shift (A)',
                      condition='rejType==%s or rejType==%s or rejType==%s'
                                % (self.REJ_FRAME, self.REJ_AND, self.REJ_OR),
                      help='Maximum drift between consecutive frames '
                           'to evaluate the frame condition.')
        form.addParam('maxMovieShift', params.FloatParam, default=15,
                      label='Max. movie shift (A)',
                      condition='rejType==%s or rejType==%s or rejType==%s'
```

# Data Models

scipion-em-xmipp / xmipp3 / protocols / **protocol_movie_max_shift.py** ↑ Top

Code Blame 391 lines (337 loc) · 17.3 KB Raw

```
42   class XmippProtMovieMaxShift(ProtProcessMovies):
205          def fillOutput(newDoneList, firstTime, AccOrDisc='Accepted'):
248
249              if movieSet.getSize() > 0:
250                  self._updateOutputSet('outputMovies%s' % suffix, movieSet,
251                                        streamMode)
252
253              if self.inputMics is not None and micsSet.getSize() > 0:
254                  self._updateOutputSet('outputMicrographs%s' % suffix, micsSet,
255                                        streamMode)
256              if self.inputDwMics is not None and micsDwSet.getSize() > 0:
257                  self._updateOutputSet('outputMicrographsDoseWeighted%s' % suffix,
258                                        micsDwSet, streamMode)
259              if firstTime:  # define relation just the first time
260                  if movieSet.getSize() > 0:
261                      self._defineTransformRelation(self.inputMovies.get(), movieSet)
262                  if self.inputMics is not None and micsSet.getSize() > 0:
263                      self._defineTransformRelation(self.inputMics, micsSet)
264                  if self.inputDwMics is not None and micsDwSet.getSize() > 0:
265                      self._defineTransformRelation(self.inputDwMics, micsDwSet)
266
267              movieSet.close()
268              if self.inputMics is not None:
269                  micsSet.close()
270              if self.inputDwMics is not None:
271                  micsDwSet.close()
```

# Scipion data model

DB Browser for SQLite - /home/coss/data/Dropbox/H/ScipionUserData/projects/CoursePasteur/Runs/007895_ProtMotionCorr/microgr...

File Edit View Tools Help

New Database | Open Database | Write Changes | Revert Changes | Open Project | Save Project | Attach Database

Database Structure | Browse Data | Edit Pragmas | Execute SQL

Table: Classes

| | id | label_property | column_name | class_name |
|---|---|---|---|---|
| 1 | 1 | self | c00 | Micrograph |
| 2 | 2 | _index | c01 | Integer |
| 3 | 3 | _filename | c02 | String |
| 4 | 4 | _samplingRate | c03 | Float |
| 5 | 5 | _acquisition | c04 | Acquisition |
| 6 | 6 | _acquisition._magnification | c05 | Float |
| 7 | 7 | _acquisition._voltage | c06 | Float |
| 8 | 8 | _acquisition._sphericalAberration | c07 | Float |
| 9 | 9 | _acquisition._amplitudeContrast | c08 | Float |
| 10 | 10 | _acquisition._doseInitial | c09 | Float |
| 11 | 11 | _acquisition._dosePerFrame | c10 | Float |
| 12 | 12 | _acquisition.opticsGroupInfo | c11 | String |
| 13 | 13 | _micName | c12 | String |
| 14 | 14 | plotGlobal | c13 | Image |
| 15 | 15 | plotGlobal._index | c14 | Integer |
| 16 | 16 | plotGlobal._filename | c15 | String |
| 17 | 17 | plotGlobal._samplingRate | c16 | Float |
| 18 | 18 | _rlnAccumMotionTotal | c17 | Float |
| 19 | 19 | _rlnAccumMotionEarly | c18 | Float |
| 20 | 20 | _rlnAccumMotionLate | c19 | Float |

1 - 20 of 20 Go to: 1

Edit Database Cell

Mode: Text

Type of data currently in cell: Text / Numeric
1 character(s)

Apply

Remote

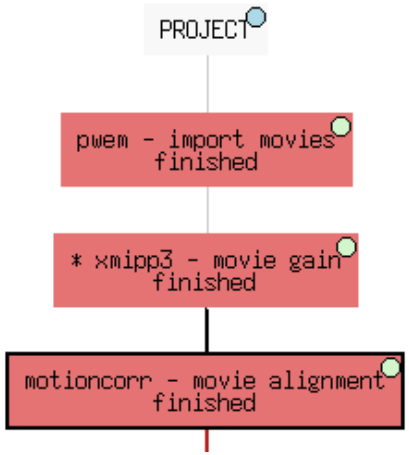
Identity Select an identity to connect

DBHub.io | Local | Current Database

Name | Last modified | Size

SQL Log | Plot | DB Schema | Remote

UTF-8




Summary | Methods | Output Log

Input
inputMovies (from xmipp3 - movie gain -> outputMovies [outputMovies]) Movies (30 items, 3710 x 3838 x 50 [1-50], 0.49 Å/px)

Output
motioncorr - movie alignment -> outputMovies Movies (30 items, 3710 x 3838 x 50 [1-50], 0.49 Å/px)
motioncorr - movie alignment -> outputMicrographsDoseWeighted Micrographs (30 items, 3710 x 3838, 0.49 Å/px)

# Extensible data model

- SetOfMicrographs:
  - Compulsory fields
  - Extended fields

# CryoEM ontology model

Search CRYOEM...

☐ Exact match ☐ Include obsolete terms ☑ Include imported terms

Tree | Graph

- EMThing (110)
  - EMObject (44)
    - Acquisition
    - AtomStruct
    - Coordinate
    - CTFModel
    - DefocusGroup
    - EMSet (23)
      - SetOfAtomStructs
      - SetOfClasses (3)
      - SetOfCoordinates
      - SetOfCTF
      - SetOfDefocusGroup
      - SetOfFSCs
      - SetOfImages (11)
        - SetOfImages2D (7)
          - SetOfAverages
          - SetOfMicrographs
          - SetOfMovies
          - SetOfParticles (3)
        - SetOfImages3D (2)
      - SetOfNormalModes
      - SetOfSequences
    - FSC
    - Image (10)
    - NormalMode
    - Sequence
    - Transform
  - EMProtocol (64)

☑ Show counts

☐ Show obsolete terms

☑ Show all siblings

▼ Class Information

**IsASetOf**
Micrograph

▼ Class Relations

**Subclass of**
SetOfImages2D

https://www.ebi.ac.uk/ols4/ontologies/cryoem

# Ontology

**Actions achieved within the project: Workflow FAIRness**

- CryoEM ontology
  - Ontology Lookup Service: https://www.ebi.ac.uk/ols/ontologies/cryoem
  - BioPortal: https://bioportal.bioontology.org/ontologies/CRYOEM
  - FAIRsharing: https://fairsharing.org/bsg-s001477/
- RO-Crate describing the image processing process
  - JSON and CWL workflow + diagram + metadata

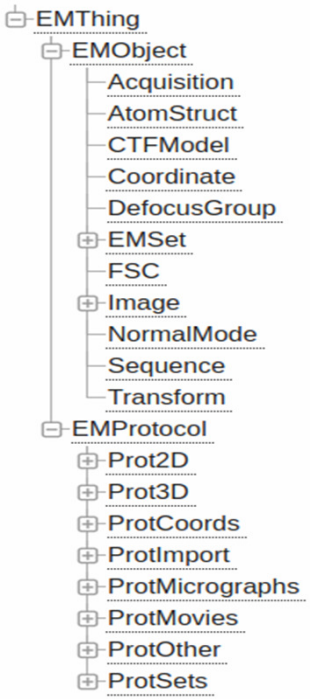



Figure 1. CryoEM ontology view from OLS

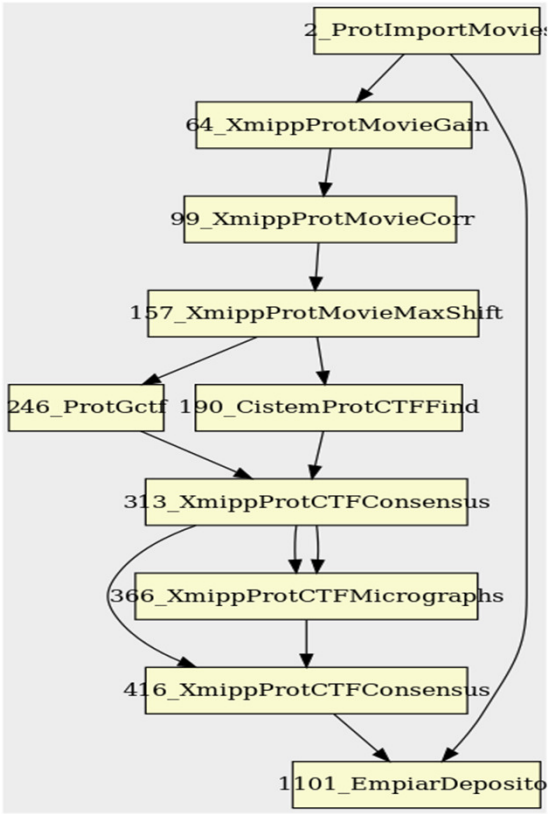



Figure 2. RO-Crate diagram

# Mathematical meaning

Geometrical transformations can be represented by matrix operations between homogeneous coordinates:

$$\tilde{\mathbf{r}}_{\tilde{A}} = \tilde{A}\tilde{\mathbf{r}}, \tag{1.2}$$

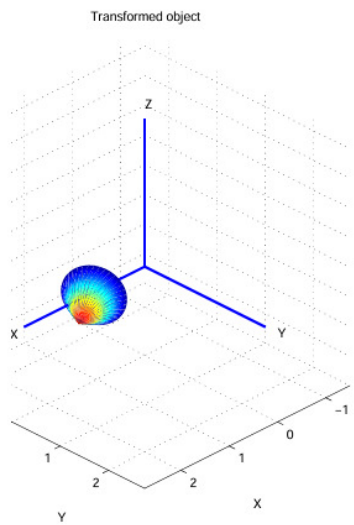
where $\tilde{\mathbf{r}} \in \mathbb{R}^3 \times \{1\}$ is the homogeneous coordinate of the point to transform, $\tilde{\mathbf{r}}_{\tilde{A}} \in \mathbb{R}^3 \times \{1\}$ is its transformed point in homogeneous coordinates, and $\tilde{A}$ is a $4 \times 4$ invertible, real matrix of the form

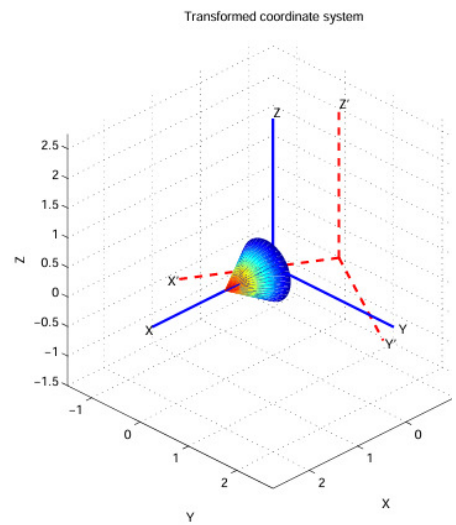
$$\tilde{A} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} R & \mathbf{t} \\ \mathbf{0}^T & 1 \end{pmatrix}. \tag{1.3}$$

$$V_{\tilde{A}}(\tilde{\mathbf{r}}) = V(\tilde{A}^{-1}\tilde{\mathbf{r}})$$

$$I_{\tilde{A}}(\tilde{\mathbf{s}}) = \int_{-\infty}^{\infty} V_{\tilde{A}}(\tilde{H}^T\tilde{\mathbf{s}})dt$$

$$\tilde{H}^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & t \\ 0 & 0 & 1 \end{pmatrix}.$$

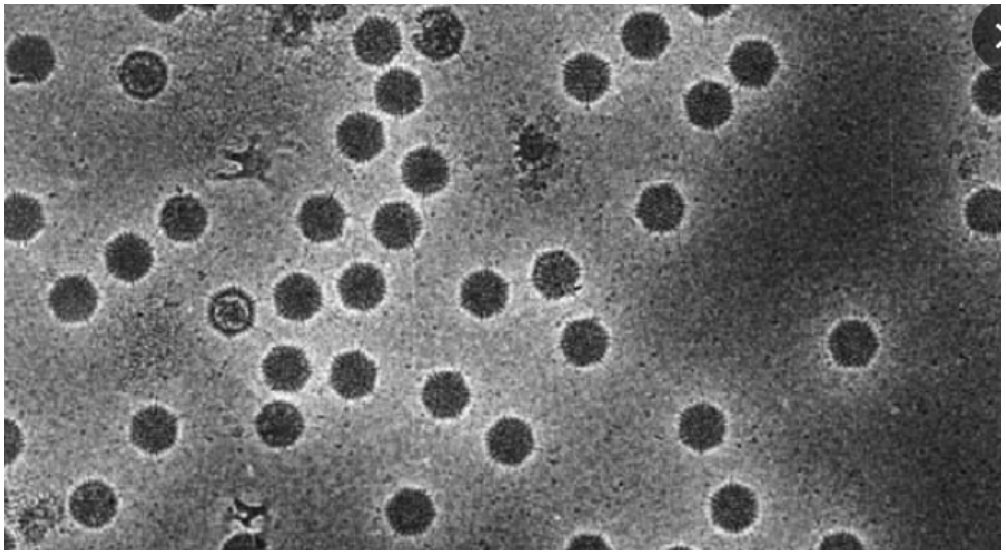


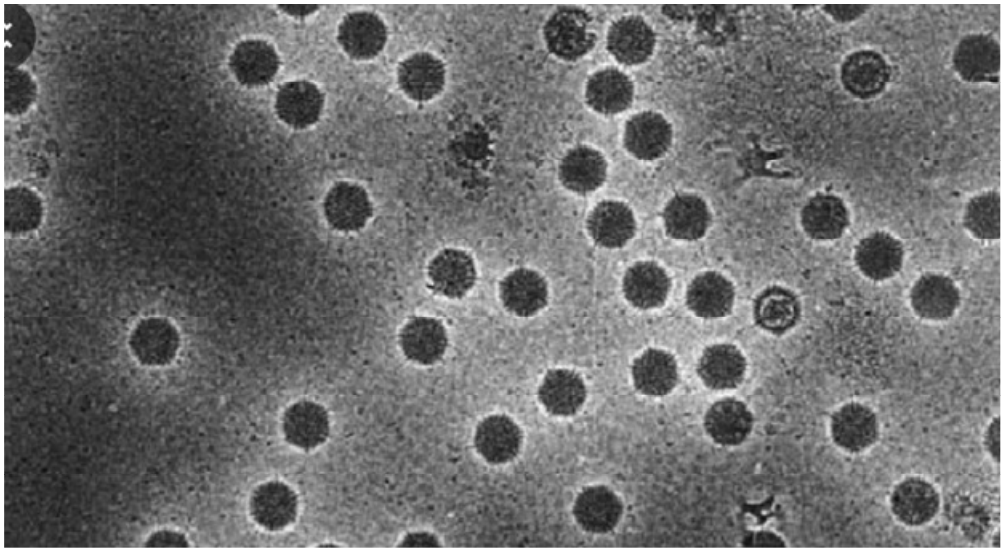




**C.O.S. Sorzano**, R. Marabini, J. Vargas, J. Oton, J. Cuenca-Alba, A. Quintana, J.M. de la Rosa-Trevin, J.M. Carazo. *Interchanging geometry information in electron microscopy single particle analysis: mathematical context for the development of a standard*. Computational Methods for Three-Dimensional Microscopy Reconstruction: 7-42 (2014) (preprint)

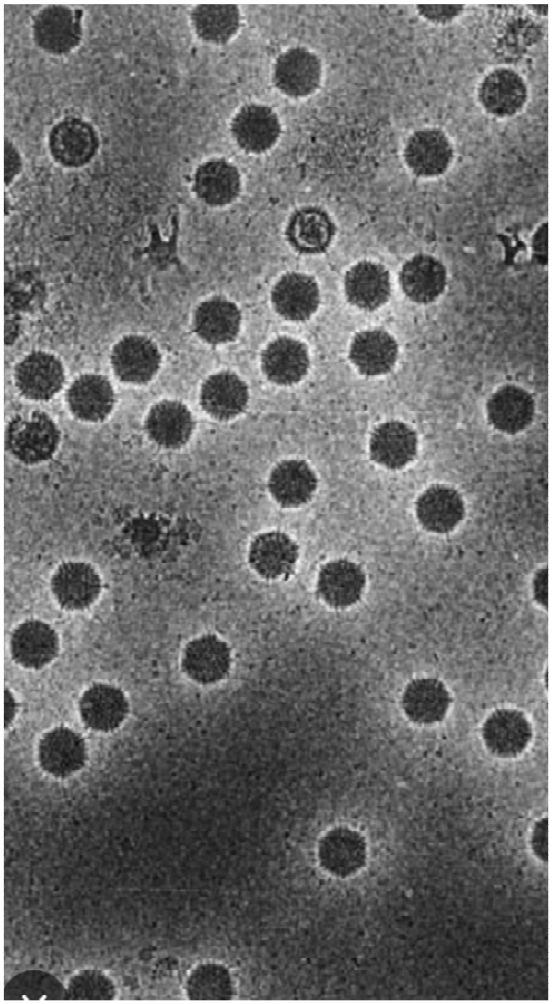
# Mathematical meaning

myMicrograph.mrc

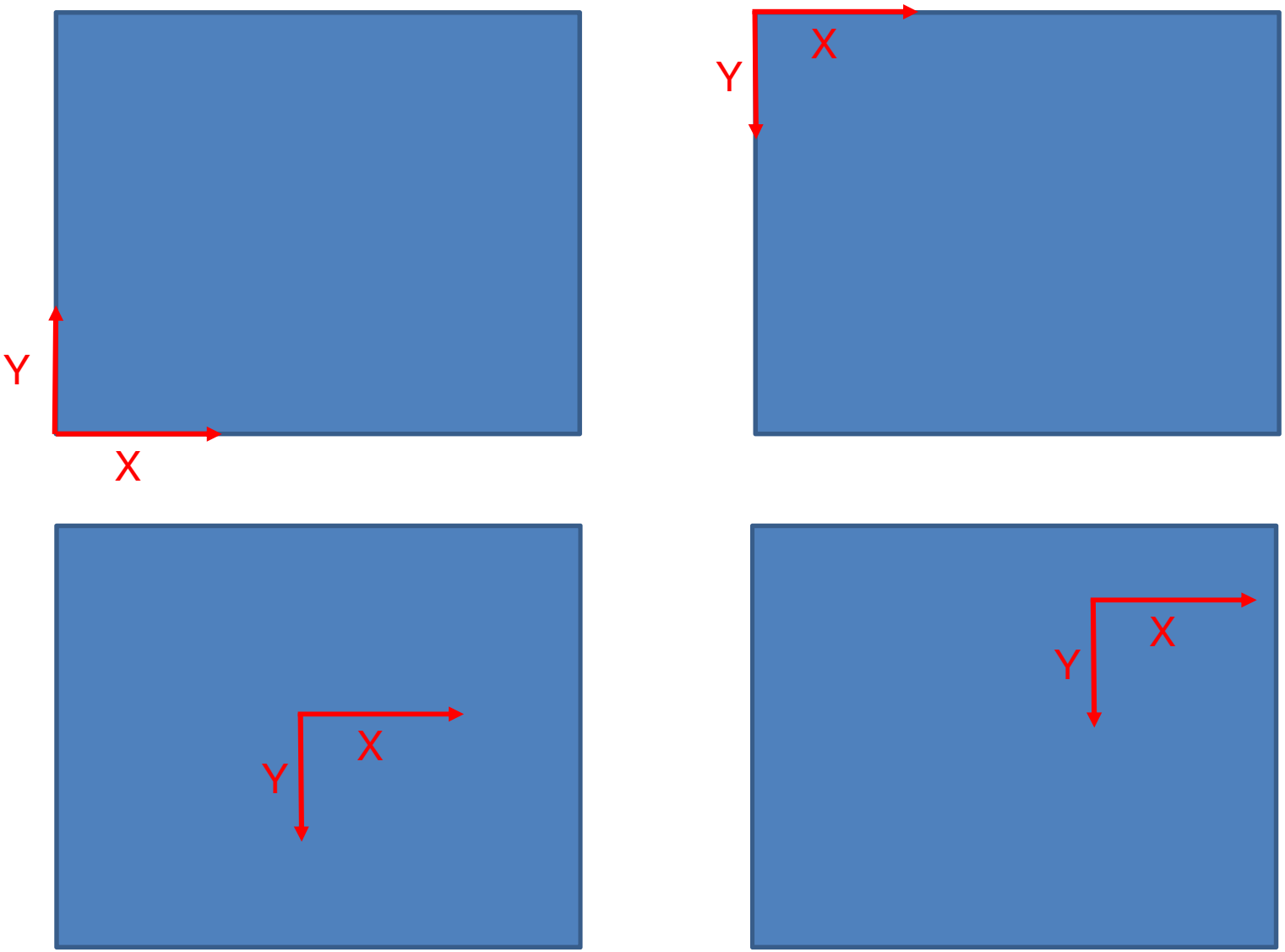




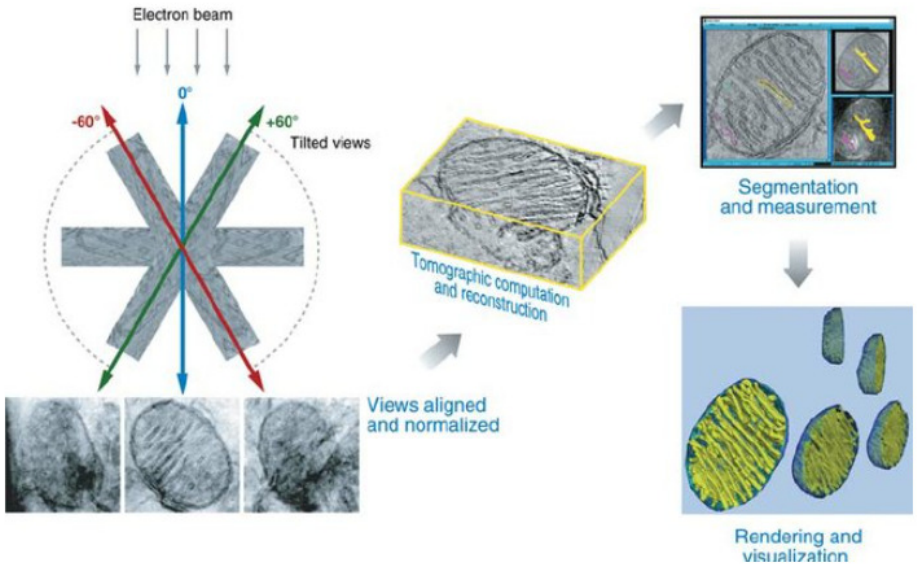
myMicrograph.tif

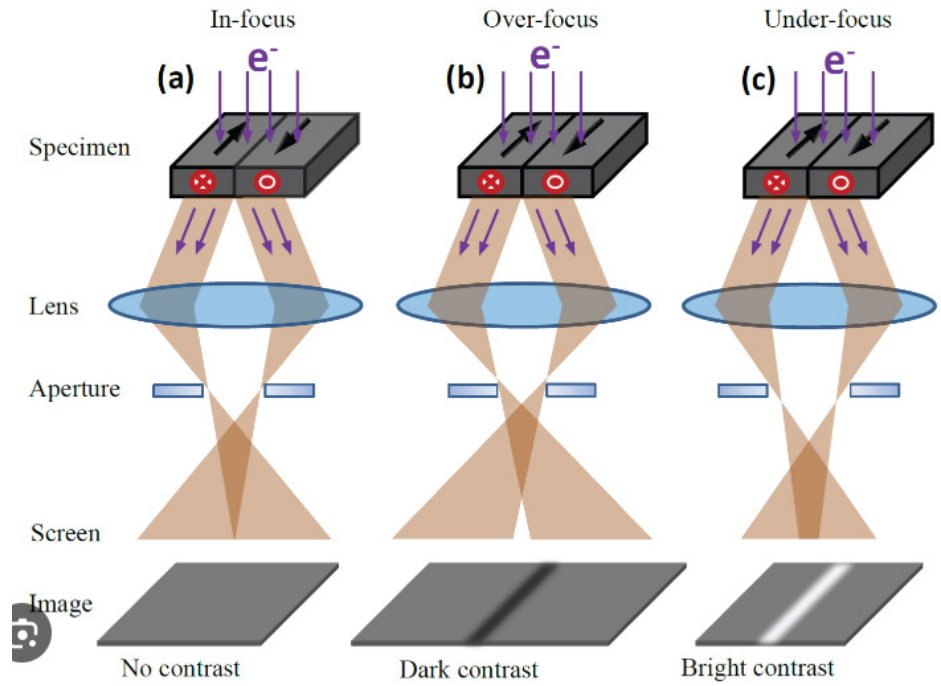


# Mathematical meaning

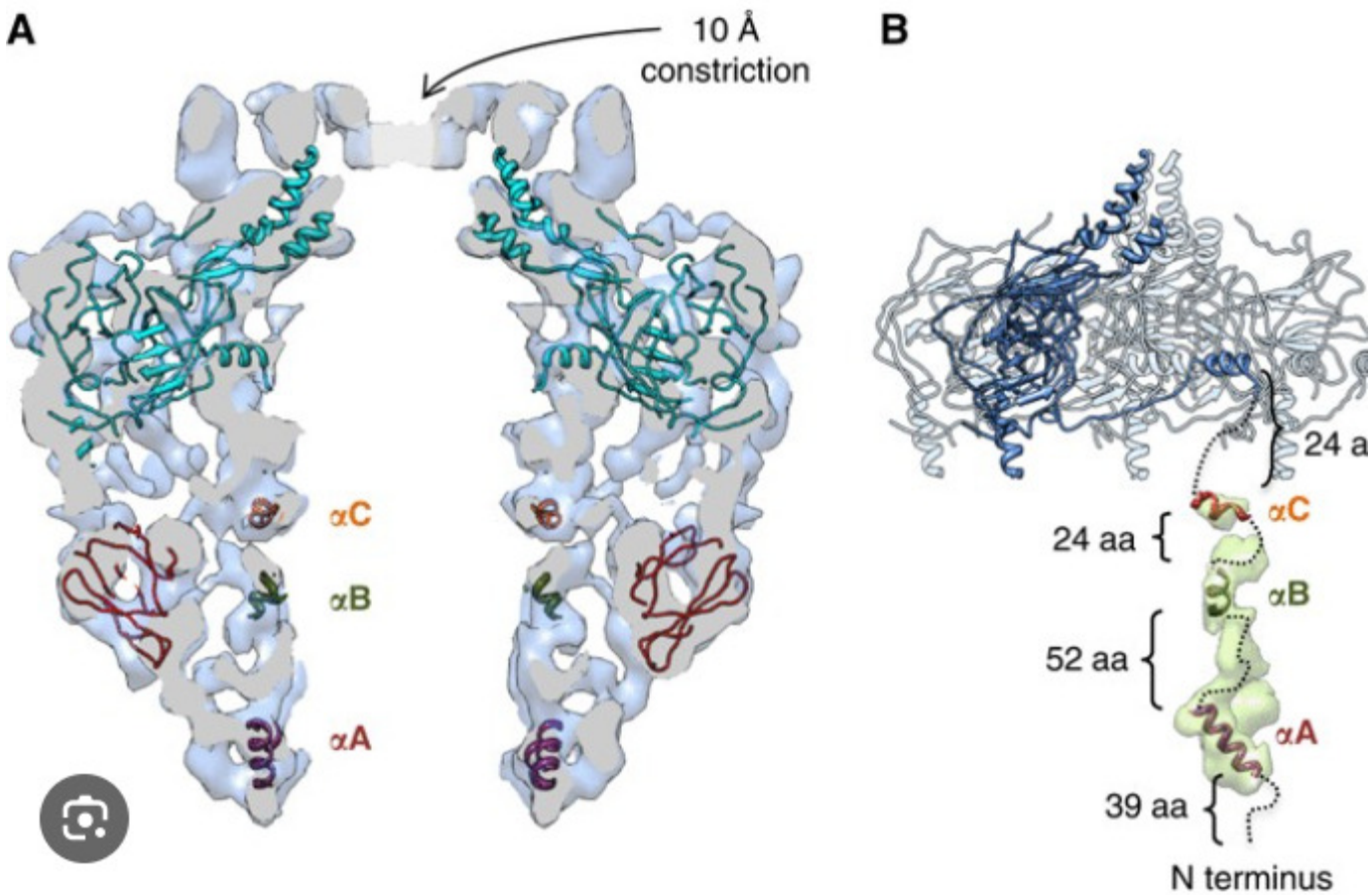


# Mathematical meaning



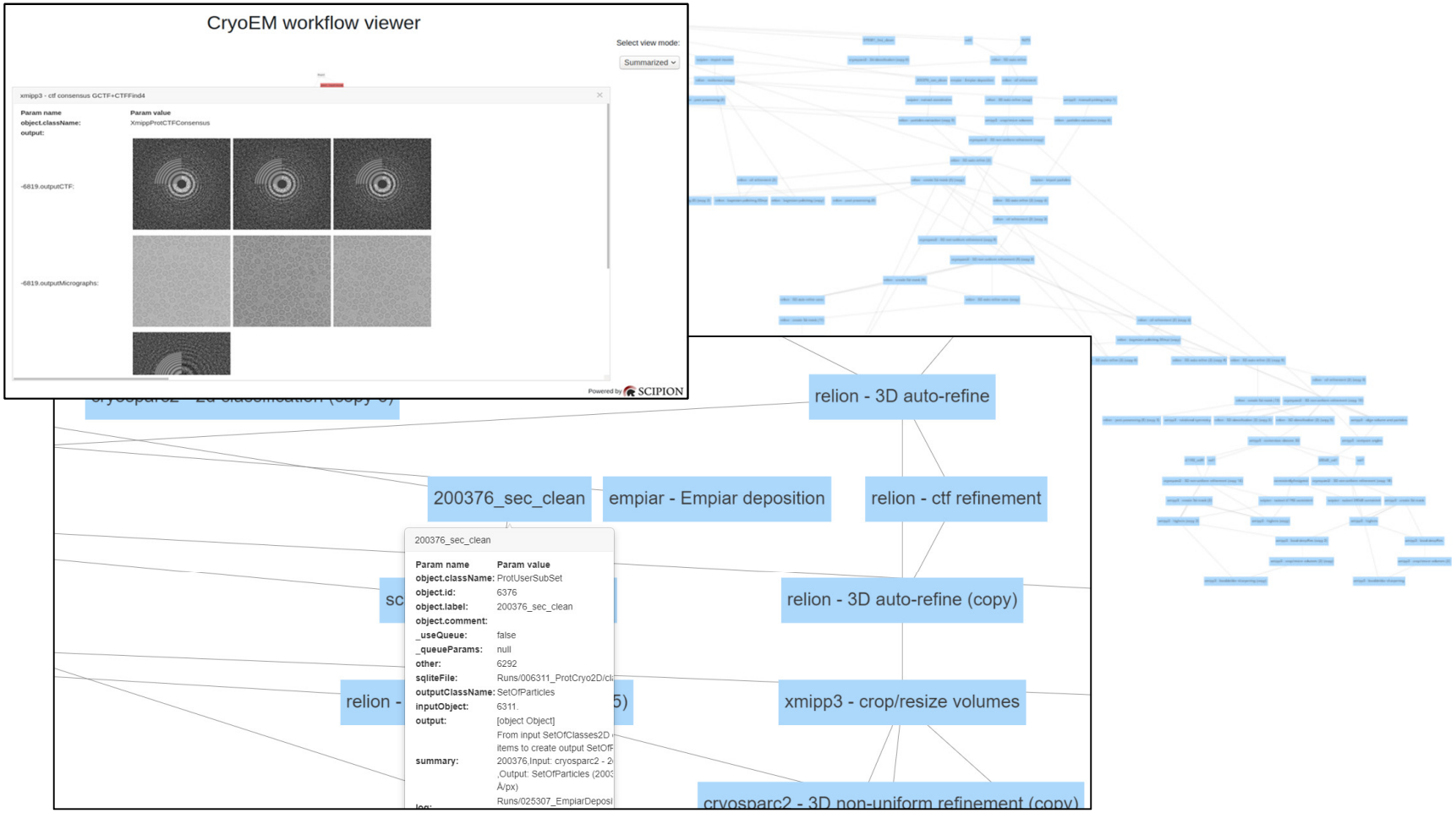
# Mathematical modelling



# EMPIAR Submission

ebi.ac.uk/pdbe/emdb/empiar/entry/10516/scipion_workflow/data/SARS-CoV-2-spike/workflow.json

## Scipion workflow viewer - EMPIAR-10516



# EMPIAR Submission


```
{
    "object.className": "ProtImportMovies",
    "object.id": "7262",
    "object.label": "pwem - import movies",
    "object.comment": "",
    "_useQueue": false,
    "_prerequisites": "",
    "_queueParams": null,
    "runName": null,
    "runMode": 0,
    "importFrom": 0,
    "filesPath": "/home/coss/ScipionUserData/projects/Example_10248_Scipion3/EMPIAR/",
    "filesPattern": "*.tiff",
    "copyFiles": false,
    "haveDataBeenPhaseFlipped": false,
    "acquisitionWizard": null,
    "voltage": 300.0,
    "sphericalAberration": 2.7,
    "amplitudeContrast": 0.1,
    "magnification": 50000,
    "samplingRateMode": 0,
    "samplingRate": 0.495,
    "scannedPixelSize": 7.0,
    "doseInitial": 0.0,
    "dosePerFrame": 1.0,
    "gainFile": "/home/coss/ScipionUserData/projects/Example_10248_Scipion3/EMPIAR//gain.mrc",
    "darkFile": null,
    "dataStreaming": false,
    "timeout": 43200,
    "fileTimeout": 30,
    "blacklistDateFrom": null,
    "blacklistDateTo": null,
    "useRegexps": true,
    "blacklistFile": null,
    "inputIndividualFrames": false,
    "numberOfIndividualFrames": null,
    "stackFrames": false,
    "writeMoviesInProject": false,
    "movieSuffix": "_frames.mrcs",
    "deleteFrames": false
},
{
    "object.className": "XmippProtMovieGain",
    "object.id": "7393",
    "object.label": "xmipp3 - movie gain",
```

# Conclusions

- The internal data model of each processing package is very detailed.
- They are incompatible among programs.
- Purpose is important: future or past
- There are multiple levels of detail
- The standard should be compatible with both (shared minimum and individual extensions)

# Thanks

I²PC

Laura del Caño

Carolina Simón

Irene Sánchez