Image Processing in cryoEM:

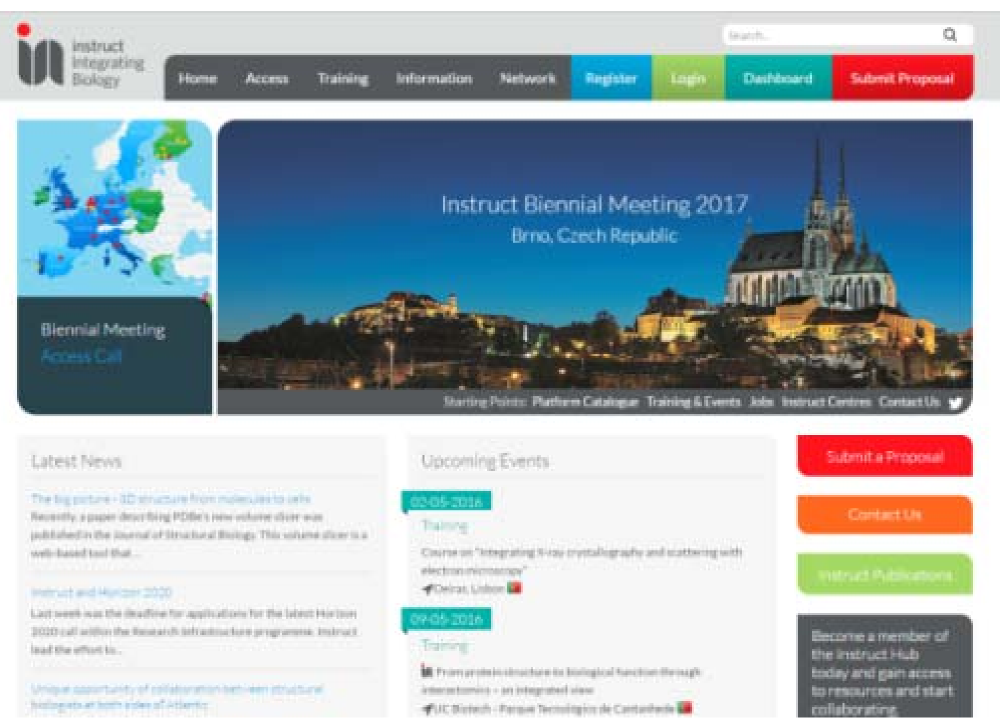
Quality checks!

José-Maria CARAZO lab

Centro Nacional de Biotecnología-CSIC

Instruct Image Processing Center

carazo@cnb.csic.es



Who are we?: The Instruct cryoEM Image Processing Center

Instruct: The European Research Infrastructure for Structural Biology

Providing access to state of the art structural biology infrastructure for researchers

CSIC

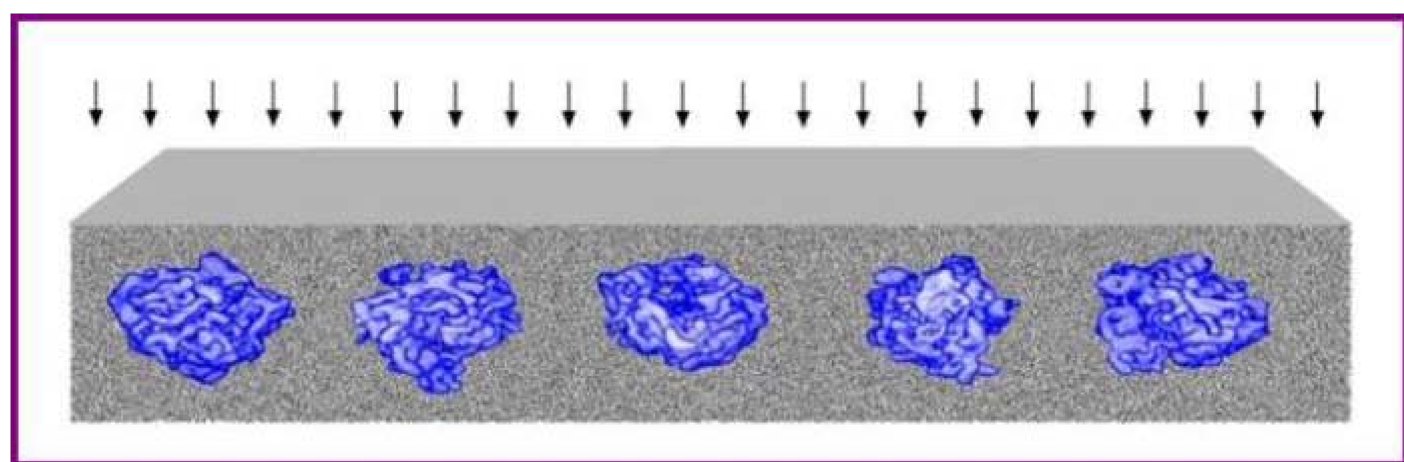
instruct Integrating Biology

BIO computing Unit

Instruct image Processing Center

Instruct copyright 2013

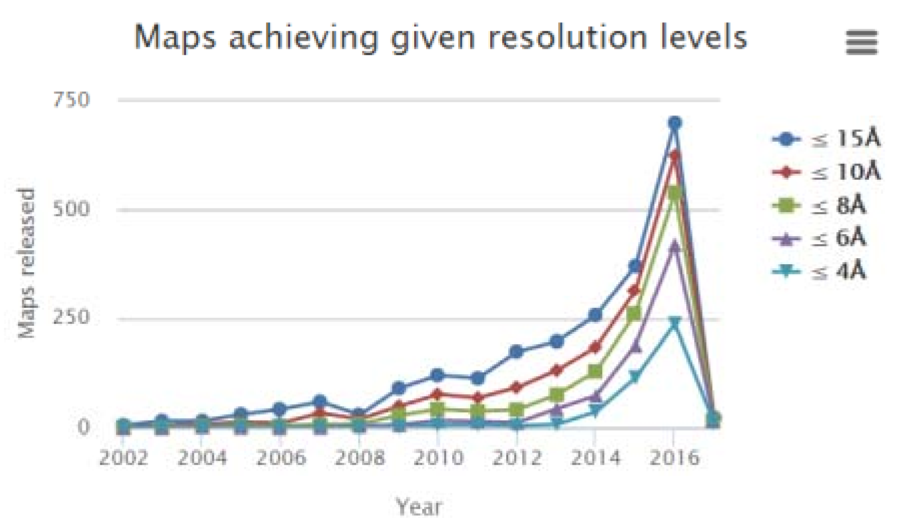
The conceptual base of cryoEM

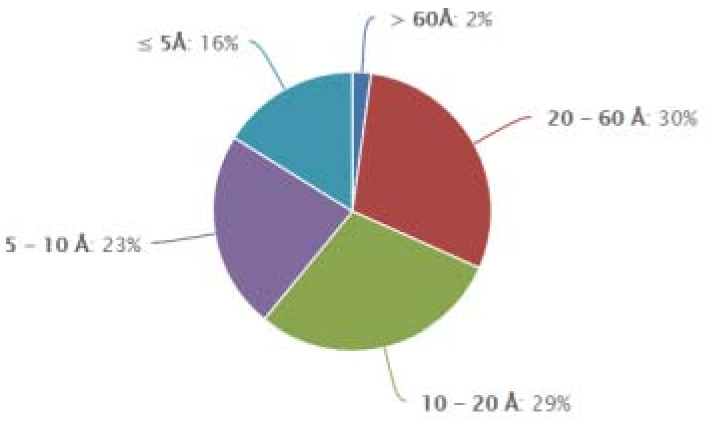


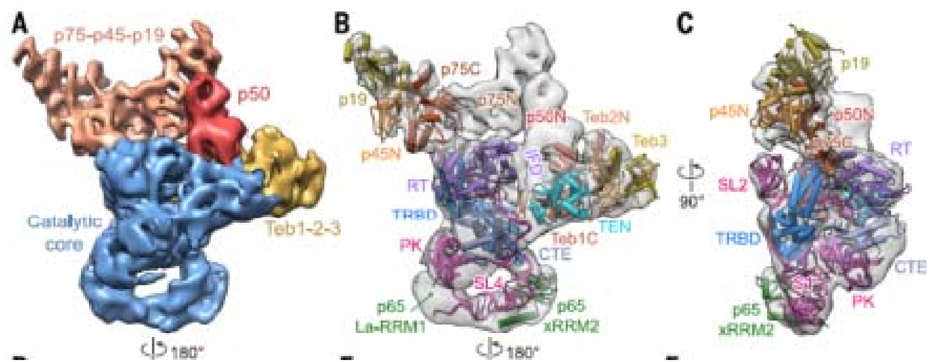
Note the "geometrical differences"!!

The cryoEM "resolution revolution": A few numbers

In 2012, FOUR structures below 4 A were solved. Now…..

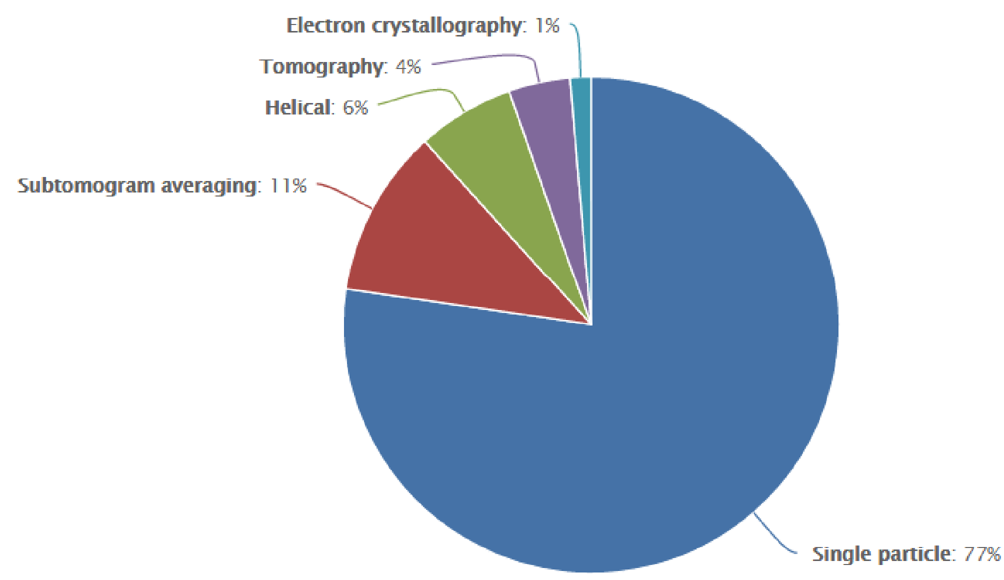






www.ebi.ac.uk/pdbe/emdb/

The cryoEM "resolution revolution": A few numbers



www.ebi.ac.uk/pdbe/emdb/

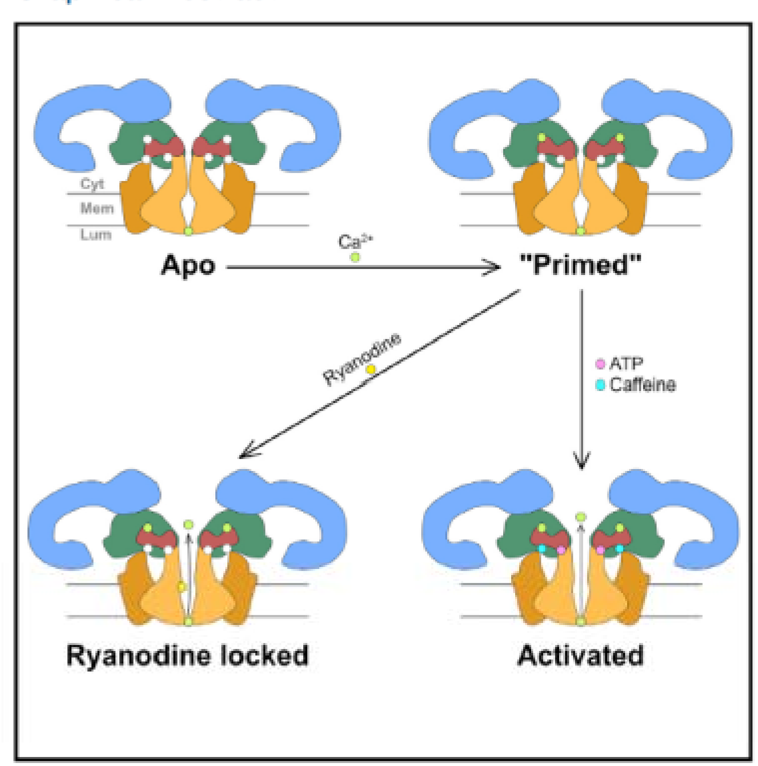
With a huge impact in the study of macromolecular flexibility

Cell

Article

Structural Basis for Gating and Activation of RyR1

Graphical Abstract



Authors

Amédée des Georges, Oliver B. Clarke, Ran Zalk, ..., Wayne A. Hendrickson, Andrew R. Marks, Joachim Frank

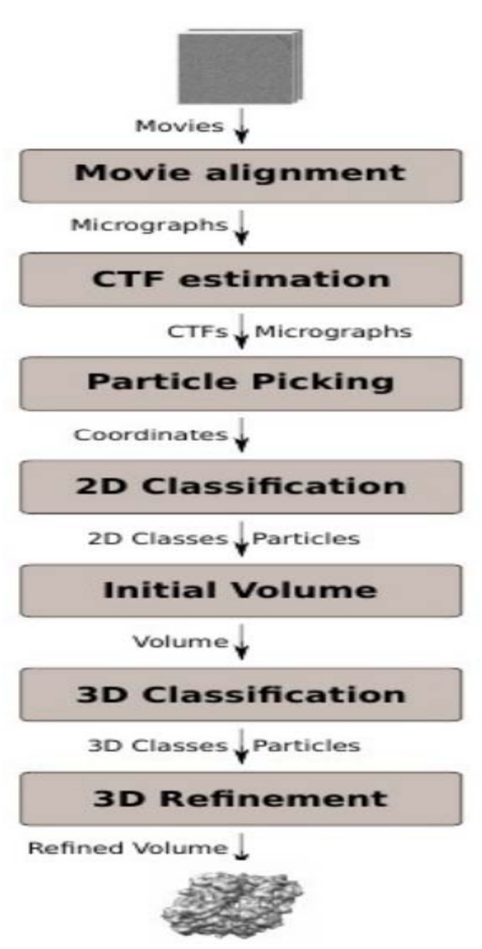
Correspondence

wayne@xtl.cumc.columbia.edu (W.A.H.),
arm42@cumc.columbia.edu (A.R.M.),
jf2192@cumc.columbia.edu (J.F.)

In Brief

Cryo-EM studies of the open and closed forms of the ryanodine receptor RyR1 in response to multiple activators provide the structural basis for its gating and activation

The cryoEM “resolution revolution”: Why now?



1) Excellent specimen preservation in cryogenic conditions

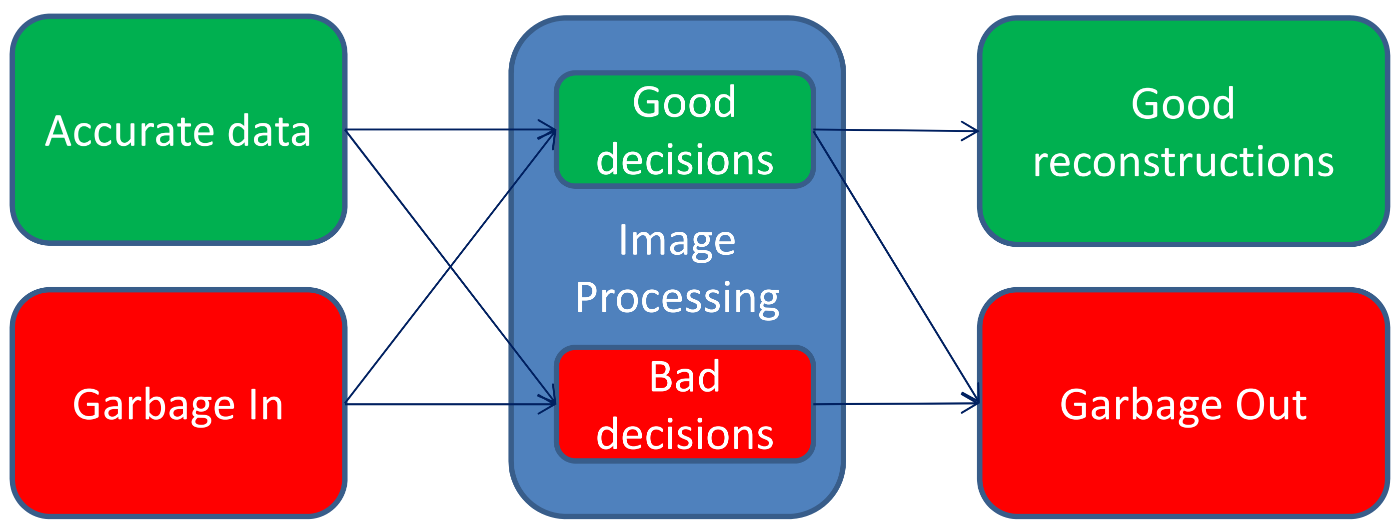
2) Excellent microscope stability and automation

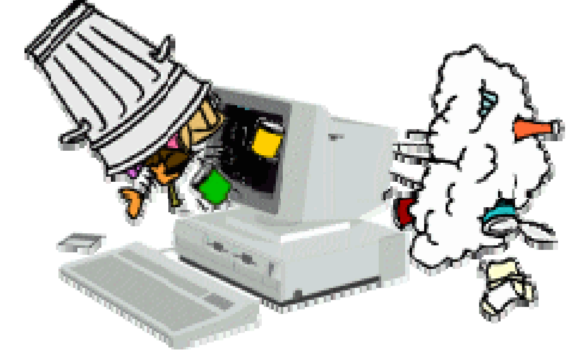
3) New Direct Electron Detectors

4) New Methods for Image Processing and Software

EM is a GIGO system

- The good thing about EM is that it always gives a volume
- The bad thing about EM is that it always gives a volume





Many steps in cryo EM involve local optimizers: Goal function landscape

f(x)

local minima

global minimum

x

c(x)

x0

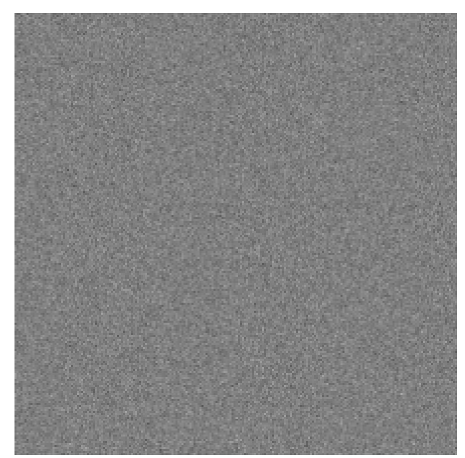
x*

S

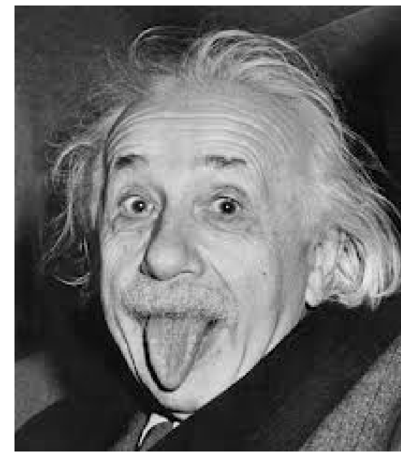
Begin

Local Minima

Model bias



1000 images



=

Model bias=local minimum!!

Shatsky, M.; Hall, R. J.; Brenner, S. E. & Glaeser, R. M. A method for the alignment of heterogeneous macromolecules from electron microscopy. J Struct Biol, 2009, 166, 67-78

New Image Processing Methods: Changing the landscape of solutions

f(x)

local minima

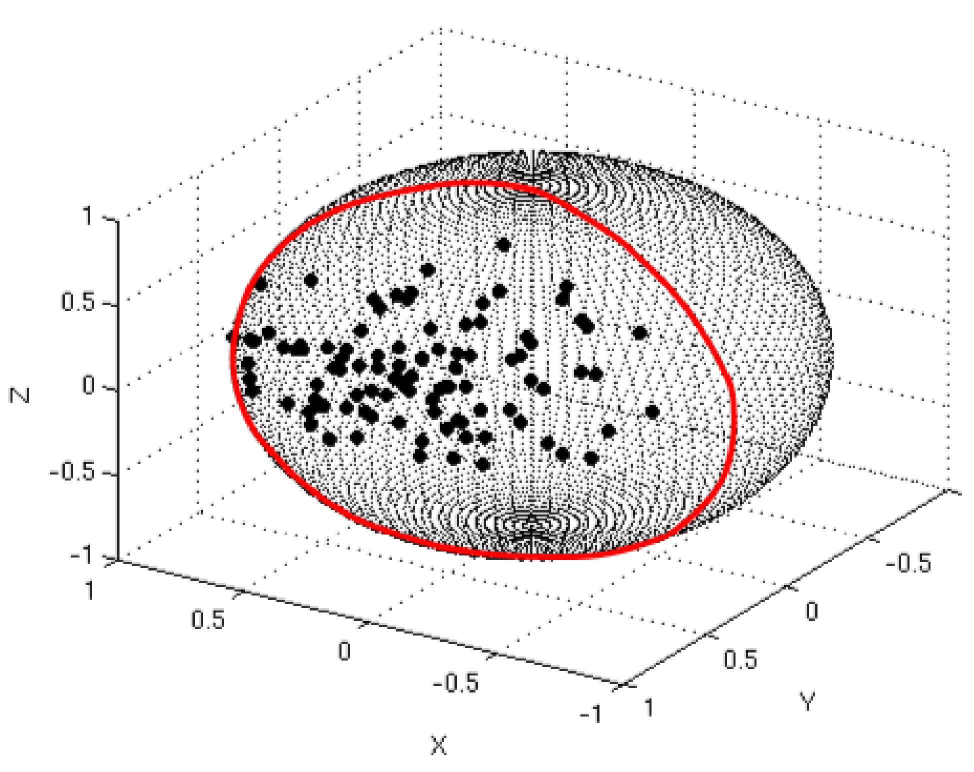
global minimum

x

f(x) f_s (x)

x

Statistical Methods for Image Processing: Conceptual bases



Statistical Methods for Image Processing

Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization

Sjors H W Scheres[1], Haixiao Gao[2], Mikel Valle[1,5], Gabor T Herman[3], Paul P B Eggermont[4], Joachim Frank[2] & Jose-Maria Carazo[1]

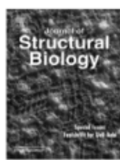
www.nature.com/naturemethods

Journal of Structural Biology 180 (2012) 519–530

Contents lists available at SciVerse ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi



RELION: Implementation of a Bayesian approach to cryo-EM structure determination

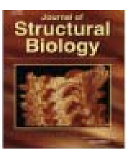
Sjors H.W. Scheres

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

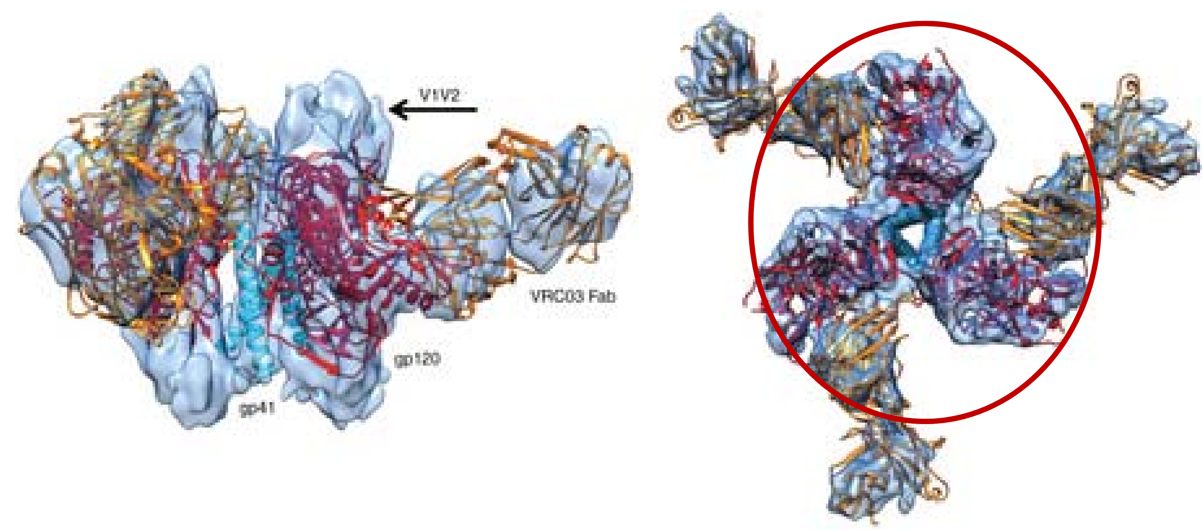


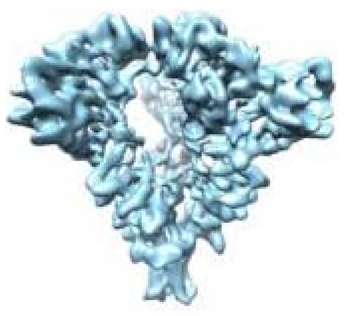
Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy

J.M. de la Rosa-Trevín [a,b,*], A. Quintana [c], L. del Cano [a], A. Zaldívar [a], I. Foche [d], J. Gutiérrez [a], J. Gómez-Blanco [a], J. Burguet-Castell [a], J. Cuenca-Alba [a], V. Abrishami [a], J. Vargas [a], J. Otón [a], G. Sharov [e], J.L. Vilas [a], J. Navas [a], P. Conesa [a], M. Kazemi [a], R. Marabini [b], C.O.S. Sorzano [a,f], J.M. Carazo [a]

but errors happen....

Which cryo EM map do you believe?





Subramanian map:EMDB_2848

Mao map: EMDB_5447

Map validation

PERSPECTIVE

Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise

Richard Henderson[1]

Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 0QH, United Kingdom

Edited by Michael G. Rossmann, Purdue University, West Lafayette, IN, and approved September 19, 2013 (received for review August 5, 2013)

Single particle cryo-electron microscopy is currently poised to produce high-resolution structures of many biological assemblies, but several pitfalls can trap the unwary. This critique highlights one problem that is particularly relevant when smaller structures are being studied. It is known as "Einstein from noise," in which the experimenter honestly believes they have recorded images of their particles, whereas in reality, most if not all of their data consist of pure noise. Selection of particles using cross-correlation methods can then lead to 3D maps that resemble the model used in the initial selection and provide the illusion of progress. Suggestions are given about how to circumvent the problem.

cryo-EM | SPEM | 3DEM | EMDB | validation

LETTER

Finding trimeric HIV-1 envelope glycoproteins in random noise

In two recent papers, Mao et al. (1, 2) present the trimeric structure of the HIV-1 gp160 trimer at 11- and 6-Å resolutions, respectively. The authors repeatedly emphasize their "reference-free," "gold standard" methodology, but in neither of their papers do they state how the particles were selected from the electron micrographs. They do state (figure S3 in ref. 1): "Each row shows a sequence of average images from the same projection class, progressing from the initial average to the converged average." and "refinements started without an external or prior reference images and did not assume any symmetry." These projection classes thus must have existed prior to the refinements, as the result of an explicit search for specific molecular views, namely projection images generated from an earlier 3D structure. Because the reference bias is introduced at this earlier stage, it becomes irrelevant that the projection classes can be averaged without using further references or symmetry assumptions. In contrast to local-variance particle detection (3), correlation function picking looks for specific views in specific orientations and may introduce reference bias (4), as do all forms of reference-based alignment (4, 5). My best

overall procedure lies in the correlation-based particle picking. The resulting 3D reconstruction, compared with the 3D search structure, yields a 13-Å cross-resolution, surprisingly better than the ~18 Å found between the two published structures (6) by Mao et al. (1, 2), although the latter is supposed to be a refined version of the first. The two maps are entirely uncorrelated at ~11-Å resolution [Fourier shell correlation (FSC) ~ 0], showing that the particles for the 6-Å map were picked/aligned with other reference images and not with those derived from the 11-Å map, as stated.

The facts that trimeric envelope glycoprotein alignments and Einstein's face can be recovered from the same pure noise images and that one can recover the full 3D map used to pick particles from pure random noise illustrate that the results of Mao et al. (1, 2) are invalid due to the use of an undisclosed reference-based methodology. Moreover, the well-defined secondary structure in the final map of Mao et al. (2) appears to have come from reiterating their reference-based particle-picking procedures after modeling in secondary structures, thereby adding new bias of the "Einstein" kind. These authors must make their datasets,

published in PNAS. This procedure was not described in the original *Materials and Methods* sections, and is prone to introduce serious bias. The authors have ignored the request to make their full data set of raw images and 670,000 selected particles publically available and have not described their particle-picking references. This failure to disclose their data for public scrutiny is disappointing and deepens the concern about the results described in their PNAS paper.

Note Added in Proof. This letter is accompanied by a related Letter from Sriram Subramaniam and Perspective article by Richard Henderson.

ACKNOWLEDGMENTS. This study was financed by grants from the Brazilian science foundation, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (Grants CNPq-[illegible] and CNPq-[illegible]), and the Dutch Ministry of Economic Affairs (Cyttron [illegible]). The computations for fig. 2 were performed by Pavel Afanasyev, on equipment financed by BBSRC Grant [illegible].

Marin van Heel[1]

Leiden Institute of Chemistry, Leiden University, 2333 CC Leiden, The Netherlands

PNAS

LETTER

Structure of trimeric HIV-1 envelope glycoproteins

Mao et al. (1) published a paper describing a 3D structure of uncleaved, trimeric HIV-1 envelope glycoprotein (Env) at ~6-Å resolution, following a similar paper last year on the same structure at ~11-Å resolution (2). Examination of these two papers leads me to doubt that the structures presented are results of a reliable cryo-electron microscopic analysis of purified HIV-1 Env trimers.

An underlying assumption in structure determination by cryo-electron microscopy is that there is demonstrable evidence that the micrographs contain images of the relevant proteins, although their mere existence in no way guarantees that a structure at 6 Å, or even at 11 Å, can be obtained. The raw electron micrographs presented by Mao et al. (1, 2) do not provide convincing evidence for the presence of molecular images of HIV-1 Env trimers [compare fig. 1A from Mao et al. (1) with fig. 1B from Harris et al. (3), where trimers are clearly visible in the images]. I am therefore concerned that the authors have fallen into the well-known reference-bias trap in image processing by recovering what looks like a real structure starting from images of random noise.

Support for this concern comes from inspection of the density maps deposited in the Electron Microscopy Data Bank (EMD-5447 and EMD-5418 from refs. 1 and 2, respectively). EMD-5418 closely matches a related map determined earlier using cryo-electron tomography (ref. 4, uncleaved trimeric HIV-1 Env as presented in the work of Mao et al. (1, 2) is in agreement with the structure of native, cleaved trimeric HIV-1 Env presented in Liu et al. (4) is incorrect. EMD-5019 represents only the ectodomain of the Env trimer; it does not include the transmembrane region, which was deposited separately (EMD-5022). Assignment of the end regions of the maps presented by Mao et al. to the gp41 transmembrane helices (1, 2) (Fig. 1 *D* and *G*) thus contradicts the earlier cryo-electron tomography results (4).

Quantitative comparison of the maps deposited by Mao et al. at resolutions of ~11 (2) and ~6 Å (1) shows that the Fourier shell coefficient falls to zero at ~11 Å (Fig. 2*A*). If the latter map was derived from the former, as the authors write, the correlation at ~11 Å would be nearly 100% and not zero. Further, filtering the 6-Å map to a resolution of 11 Å yields a map that is markedly different from EMD-5418 (Fig. 2*A*, *Inset*). The most ordered regions of EMD-5447 are the gp41 transmembrane helices, which stand out clearly against the background, and the V4 loop, which is disordered in crystal structures of monomeric gp120 (5) (Fig. 2 *B*–*D*). A plausible explanation for these surprising map features is that the authors used a molecular model to select particles from their micrographs; in this case, the refinement process would be circular, resulting in a final map that recapitulates the starting template.

original micrographs, list of locations where particles were selected, and a complete description of the protocols used for image processing.

Note Added in Proof. This letter is accompanied by a related Letter from Marin van Heel (6) and Perspective article by Richard Henderson (7).

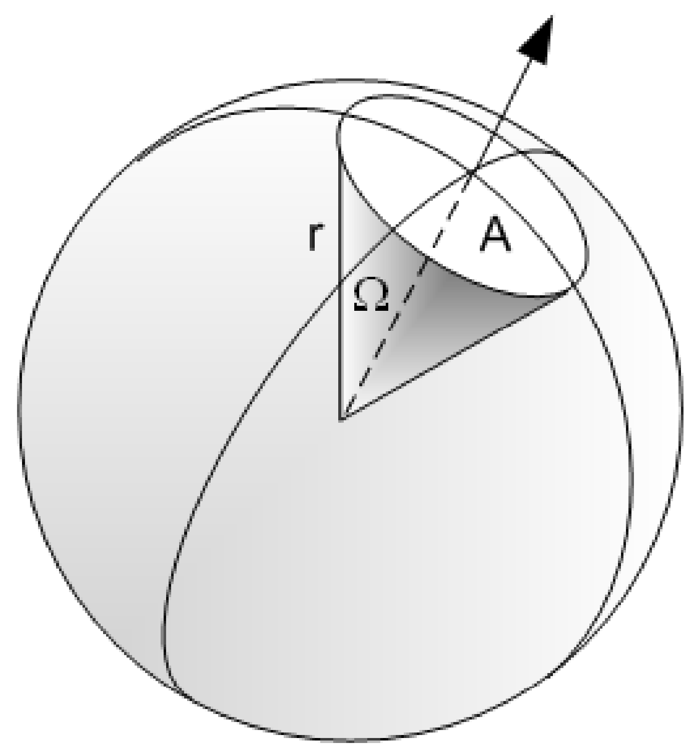
Sriram Subramaniam[1]

Laboratory of Cell Biology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892

1 Mao Y, et al. (2013) Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer. Proc Natl Acad Sci USA 110(30):12438–12443.
2 Mao Y, et al. (2012) Subunit organization of the membrane-bound HIV-1 envelope glycoprotein trimer. Nat Struct Mol Biol 19(9):893–899.
3 Harris A, et al. (2011) Trimeric HIV-1 glycoprotein gp140 immunogens and native HIV-1 envelope glycoproteins display the same closed and open quaternary molecular architectures. Proc Natl Acad Sci USA 108(28):11440–11445.
4 Liu J, Bartesaghi A, Borgnia MJ, Sapiro G, Subramaniam S (2008) Molecular architecture of native HIV-1 gp120 trimers. Nature 455(7209):109–113.
5 Kwong PD, et al. (2000) Structures of HIV-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. Structure 8(12):1329–1339.
6 van Heel M (2013) Finding trimeric HIV-1 envelope glycoproteins in random noise. Proc Natl Acad Sci USA [illegible].
7 Henderson R (2013) Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. Proc Natl Acad Sci USA 110:18037–18041.

How can we analyze the quality of a 3D map?

- **Let go to “first principles”**
  - **Consider “Basic” characteristics of a 3D map: Soft validation**

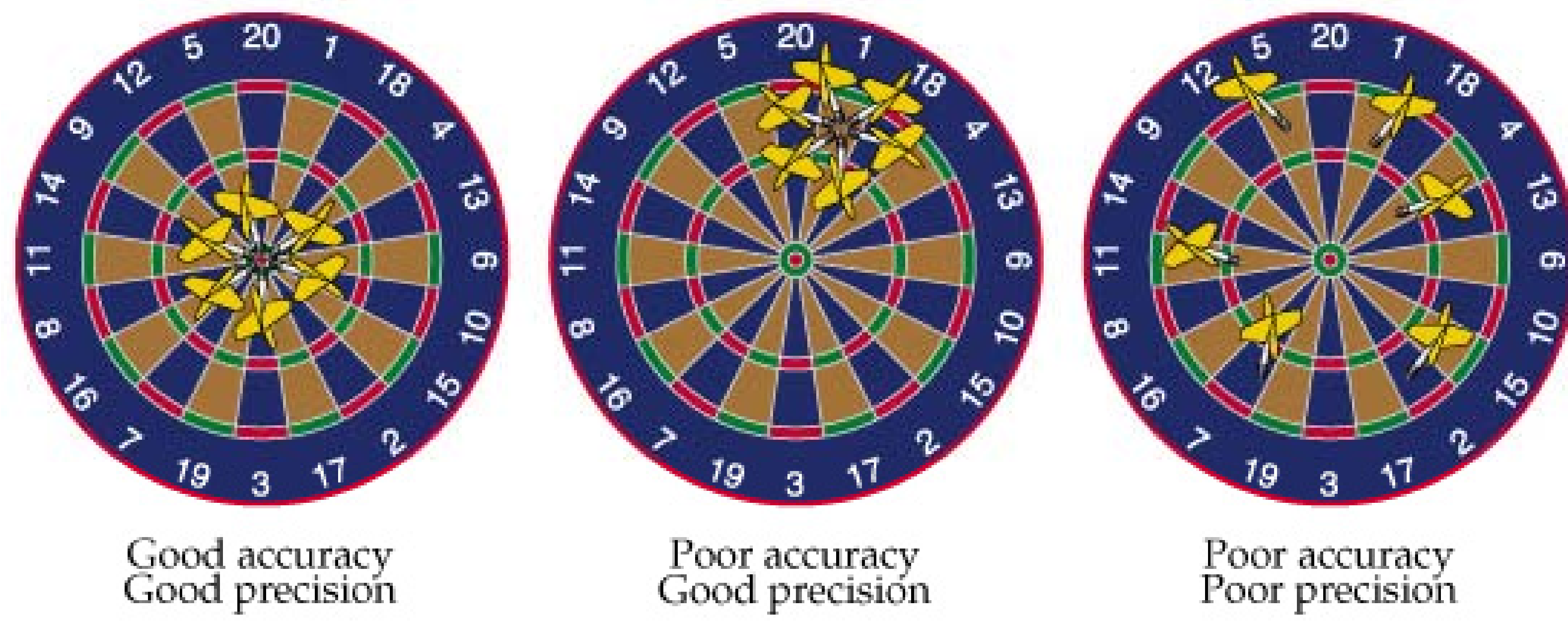


Soft alignment validation

With Soft validation we refer to NECESSARY conditions that a valid 3DEM map must verify.

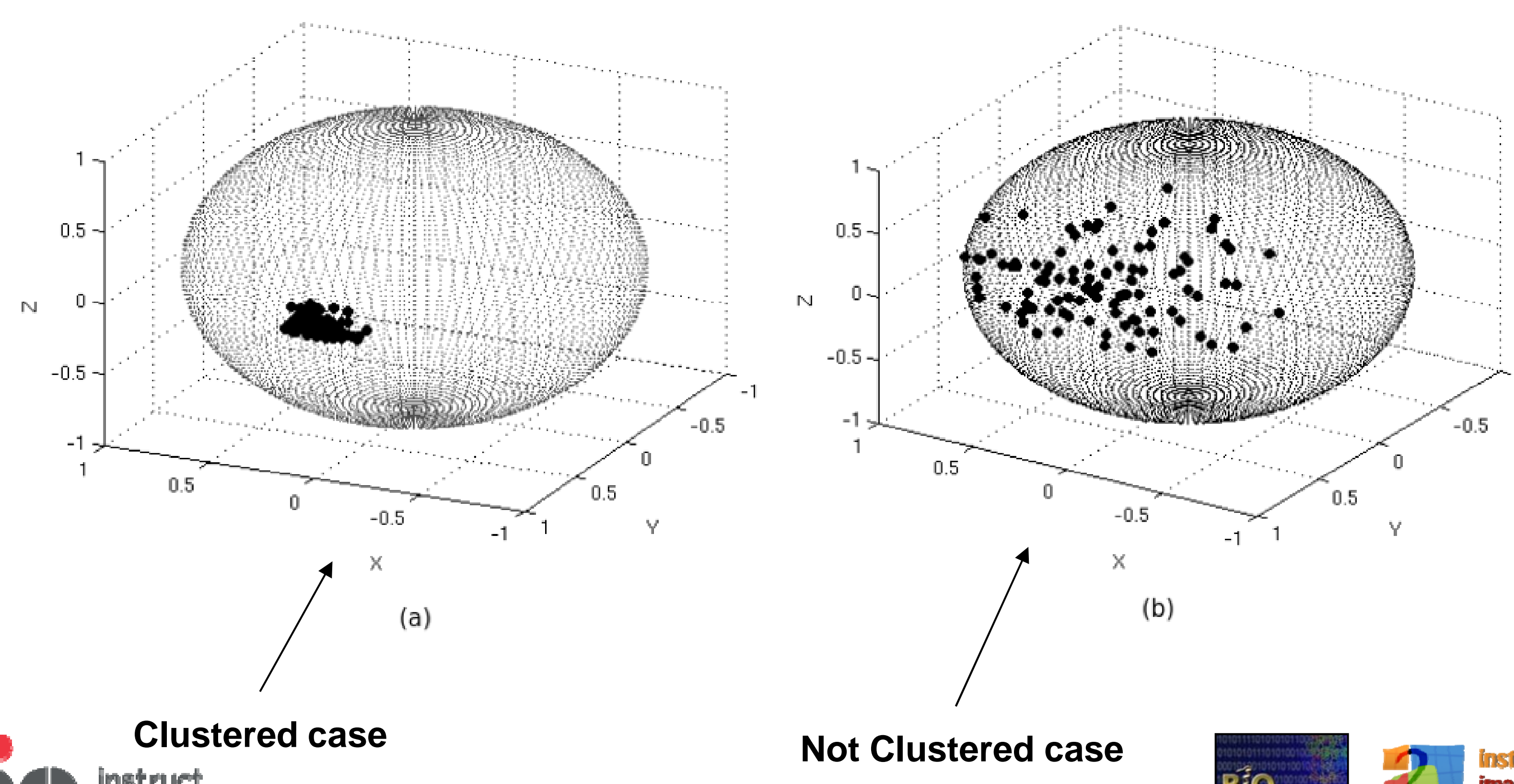
With Soft validation we DO NOT mean SUFFICIENT conditions that a valid 3DEM map must verify

Accuracy and precision



Checking for precission

We can project the 3DEM map into a regular grid of orientations and for each projection image determine the set of "best" orientations



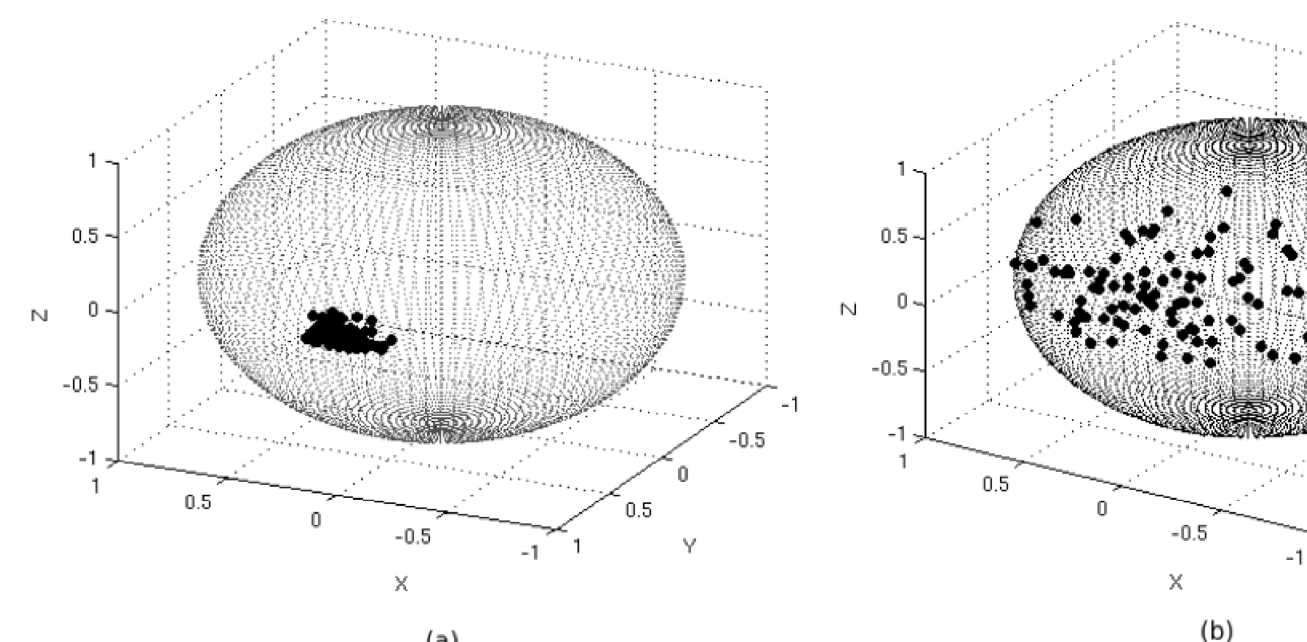
CSIC
instruct Integrating Biology
instruct image Processing Center

Checking for precission

The cluster tendency of the orientations can be quantified by the Hopkins clustering tendency

For each particle *k* we define:

$$H_k = \frac{\sum_{i=1}^{M} \alpha_{ik}}{\sum_{i=1}^{M} \alpha_{ik} + \sum_{i=1}^{M} u_{ik}}$$



Checking for precision

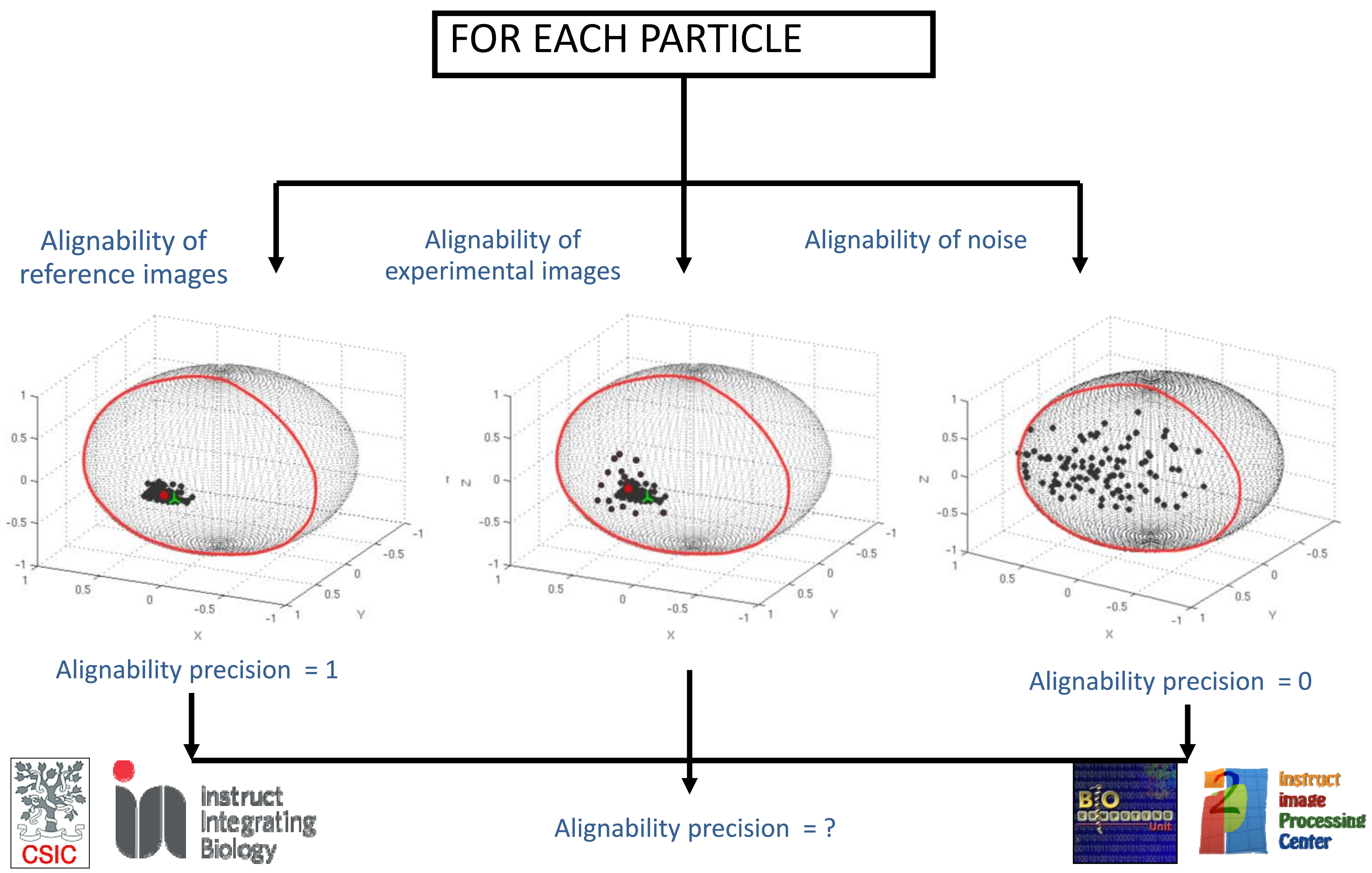
FOR EACH PARTICLE

Alignability of reference images

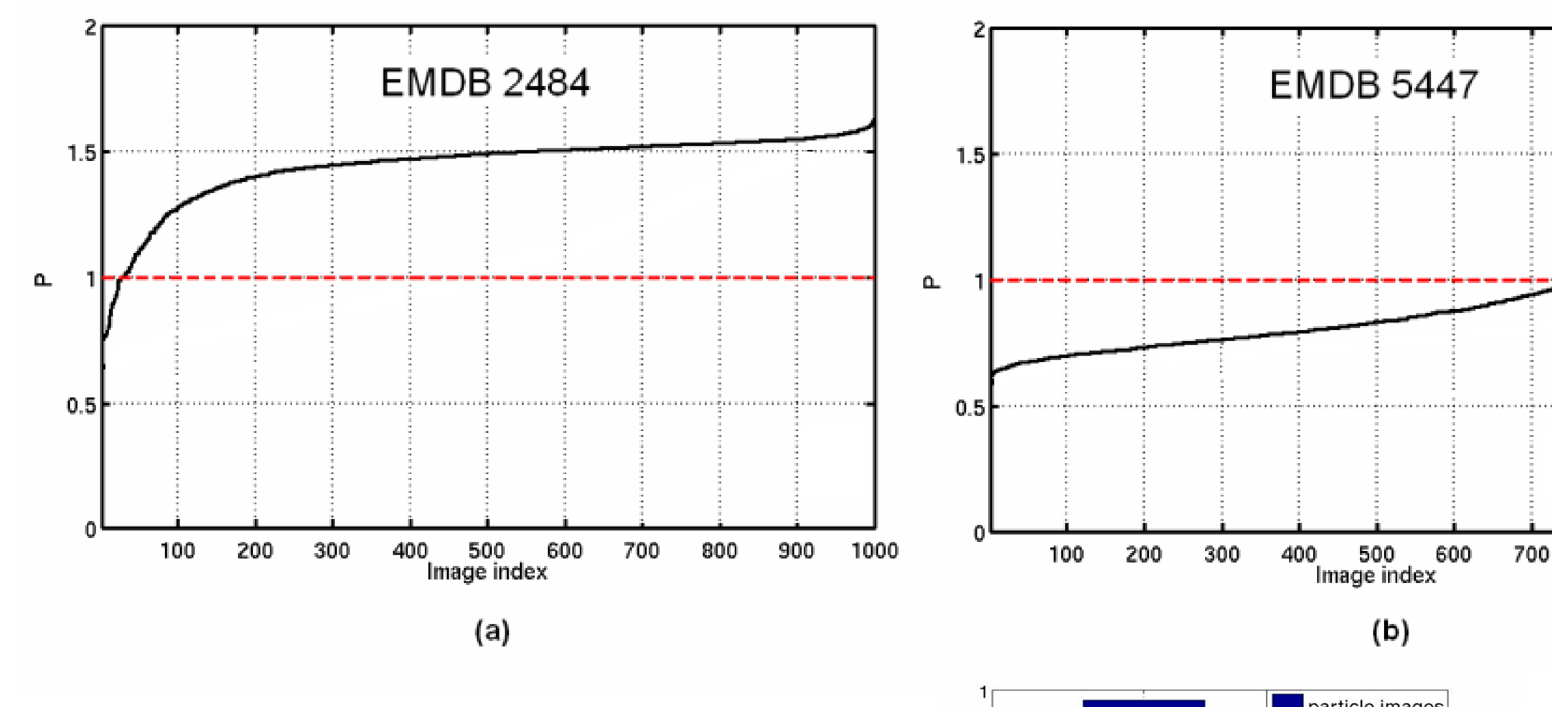
Alignability of experimental images

Alignability of noise

Checking for precission



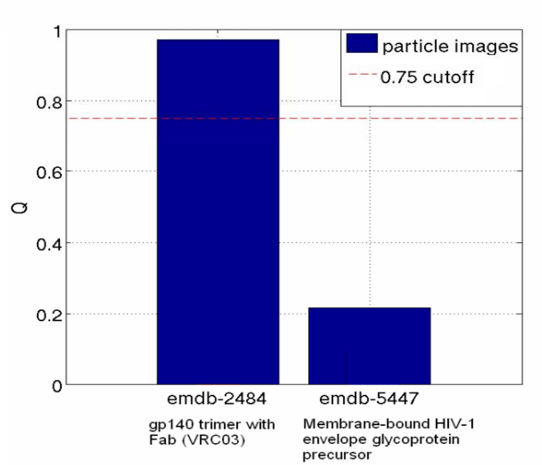
Checking for precission



$$P_k > 1$$

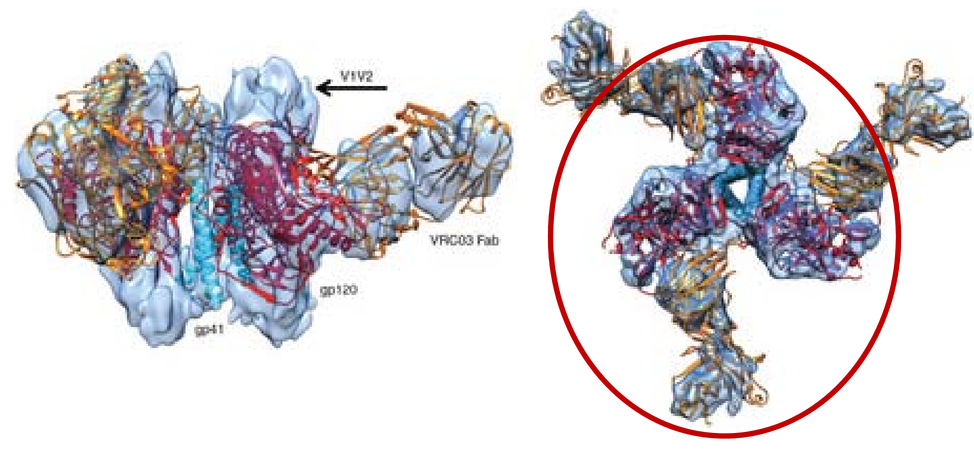
$$P_k \leq 1$$

Alignment better than noise with a significance of 95%

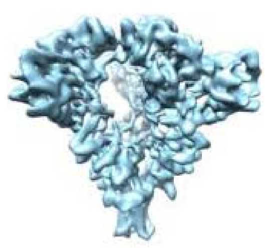


but errors happen.... and can be sorted!

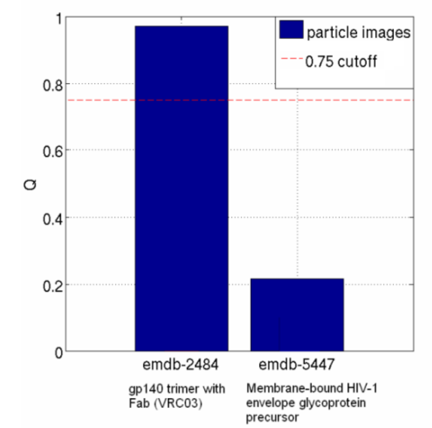
Which cryo EM map do you believe?



Subramanian map:EMDB_2848



Mao map: EMDB_5447



Checking for accuracy

Tilt direction 90º

Tilt direction 0º

(a) (b) (c) (d)

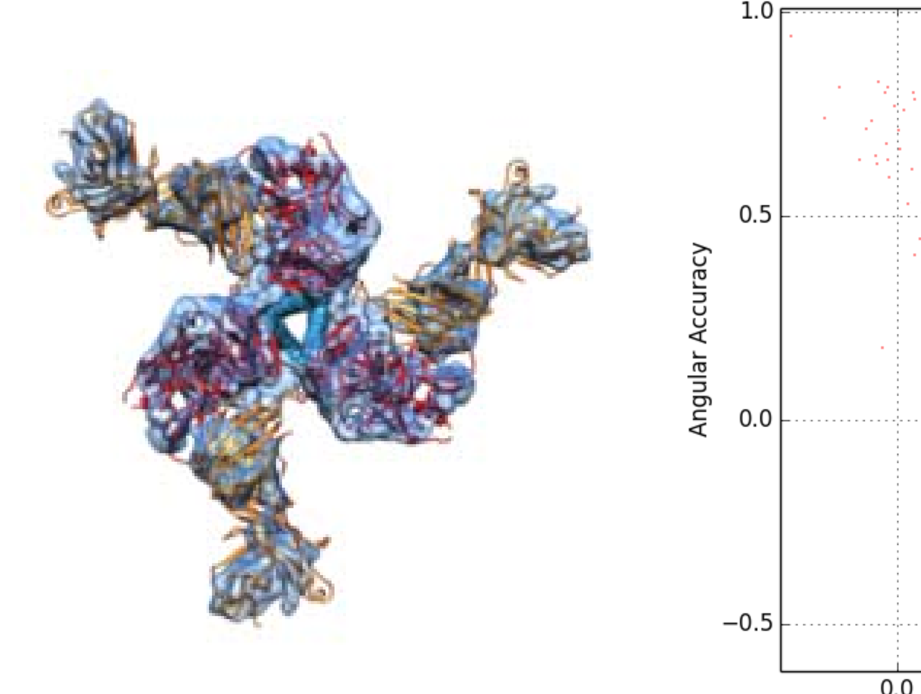
Optimization

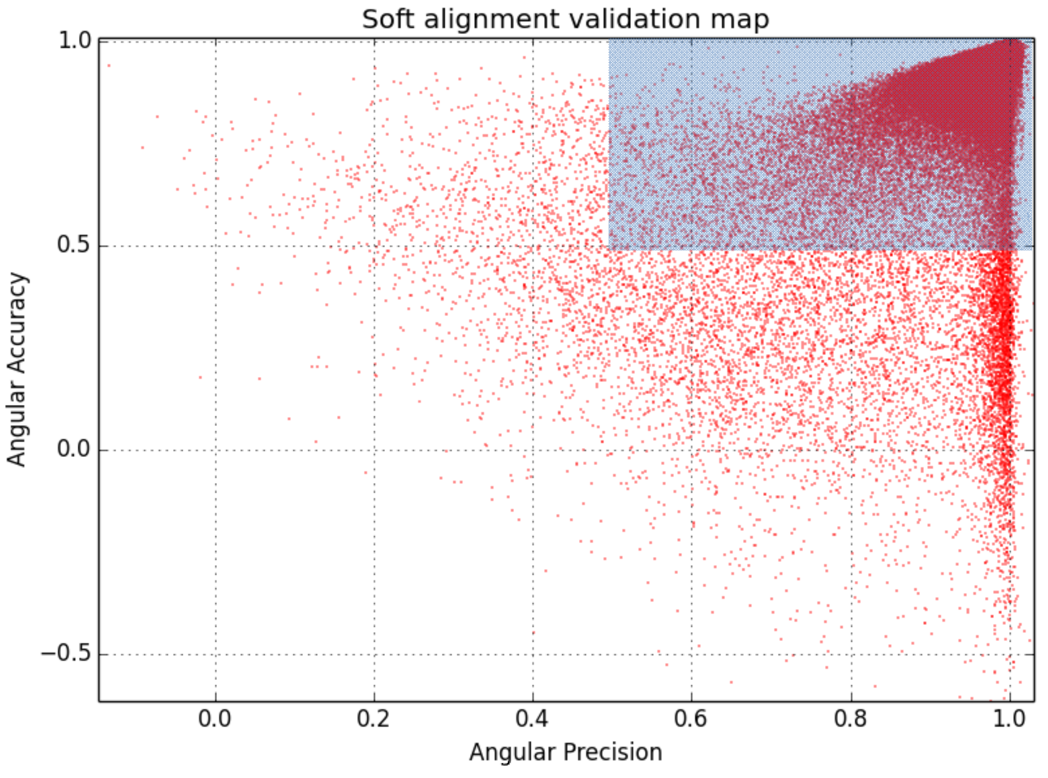
Soft alignment validation

(e)

Soft-alignment validation

88,125 particles deposited in EMPIAR





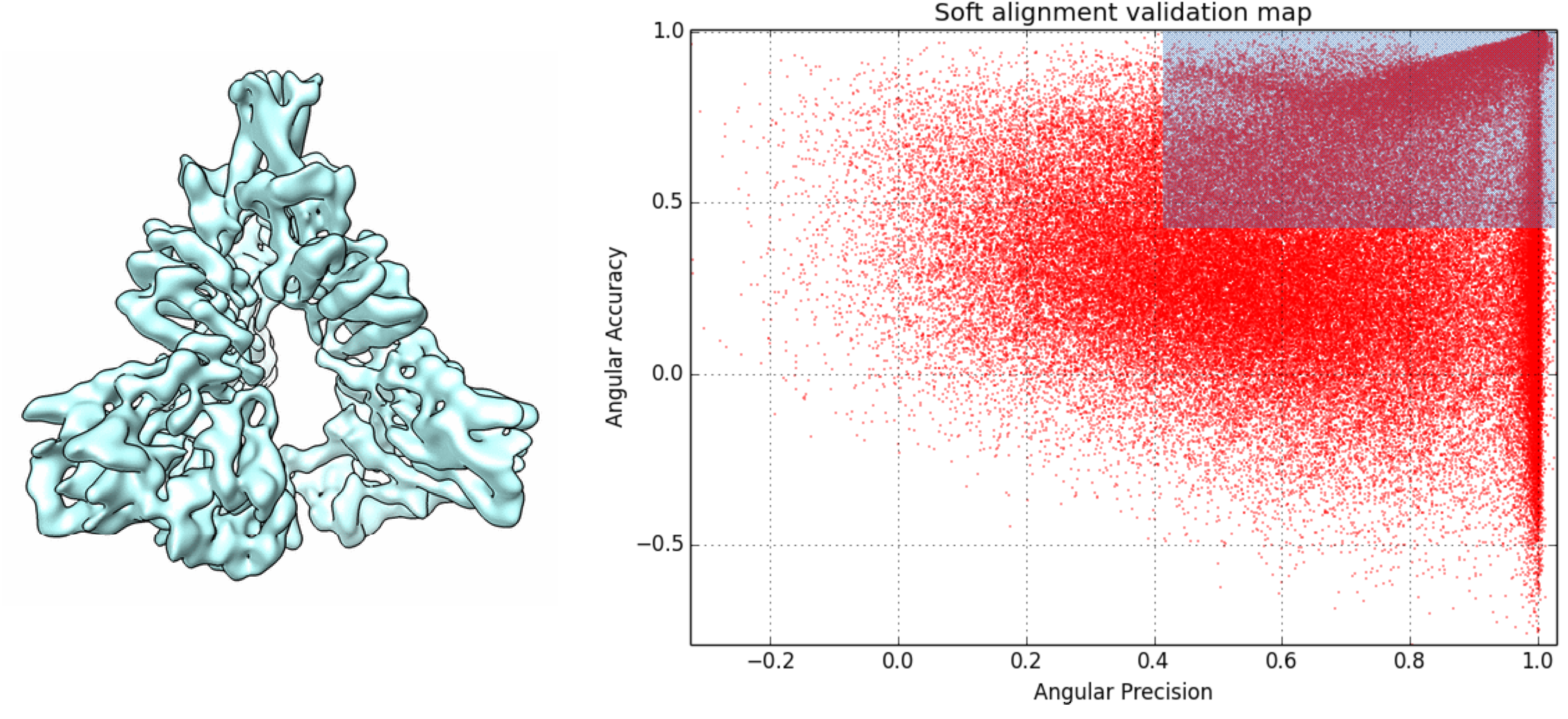
91% particles align with precision

98% particles align with accuracy

90% particles align with accuracy and precision

Soft-alignment validation

124,478 particles deposited in EMPIAR



75% particles align with precision

45% particles align with accuracy

38% particles align with accuracy and precision

Soft-alignment validation

	EMDB 5447	EMDB 2484
Precision (%)	75	91
Accuracy (%)	45	98
Precision & Accuracy (%)	38	90

Test on CHALLENGE datasets

EMDataBank Validation Challenges

Login

ABOUT | MAP CHALLENGE | MODEL CHALLENGE | NEWS

Home » Map Challenge

2015/2016 Map Challenge Goals

- Establish a benchmark set of single particle raw image datasets suitable for high resolution cryoEM, suitable for both software developers and beginners
- Encourage developers of 3DEM software packages and biological end users to analyze these datasets and present results with the best practice
- Evolve criteria for evaluation and validation of the results of the reconstruction and analysis
- Compare and contrast the various reconstruction approaches in a positive spirit, to achieve high efficiency and accuracy

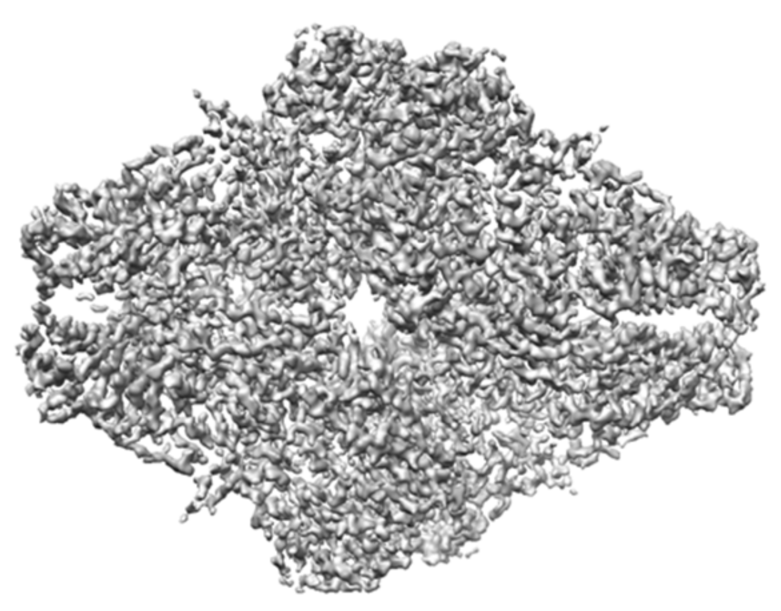
Participation Links

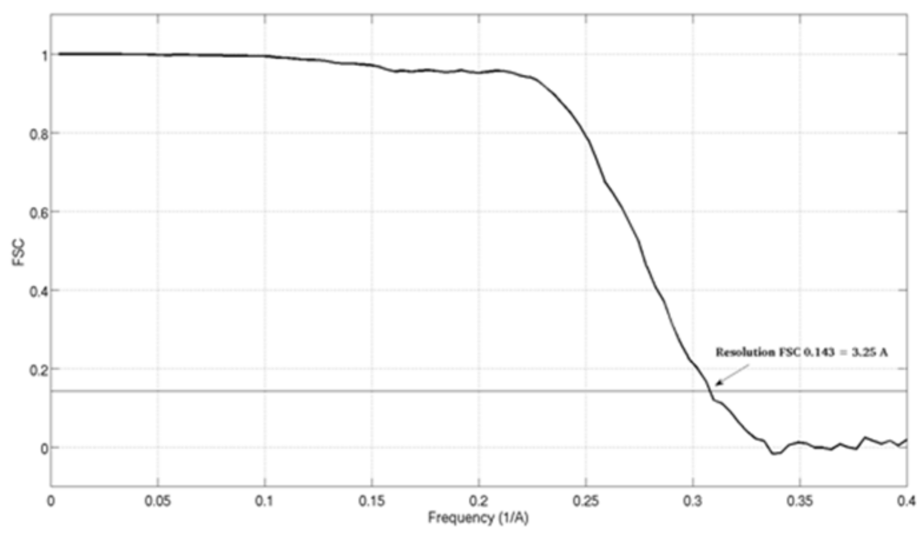
- View/Download the map challenge submissions
- Sign up to participate as an assessor
- Contribute a question, comment, or suggestion
- Register your assessment plan
- FAQ

Map Committee

Bridget Carragher (Chair), Jose-Maria Carazo, Wen Jiang, John Rubinstein, Peter Rosenthal, Fei Sun, Janet Vonck, Wah Chiu, Cathy Lawson, Ardan Patwardhan

Soft-alignment validation for pruning





Soft-alignment validation for pruning

Half 1

Soft alignment validation map

Angular Accuracy

Angular Precision

(a)

Half 2

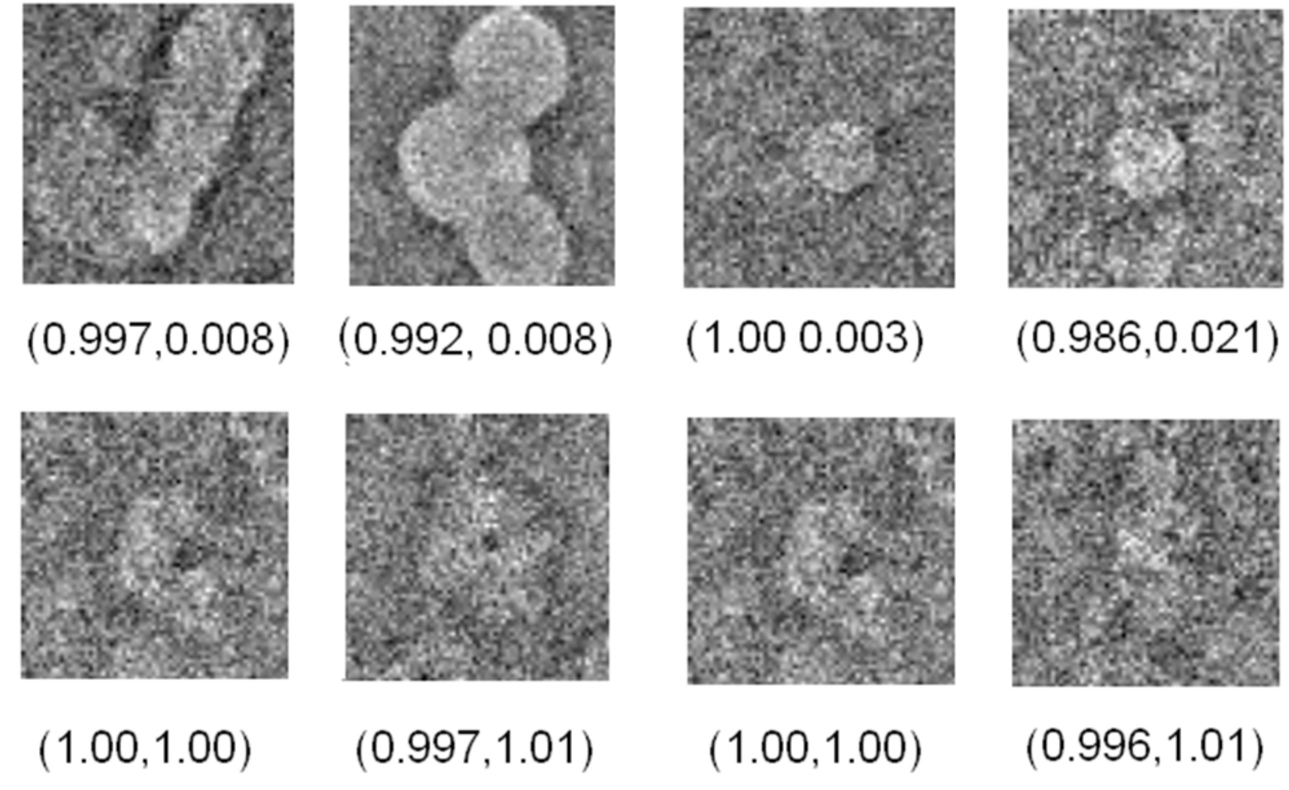
Soft alignment validation map

Angular Accuracy

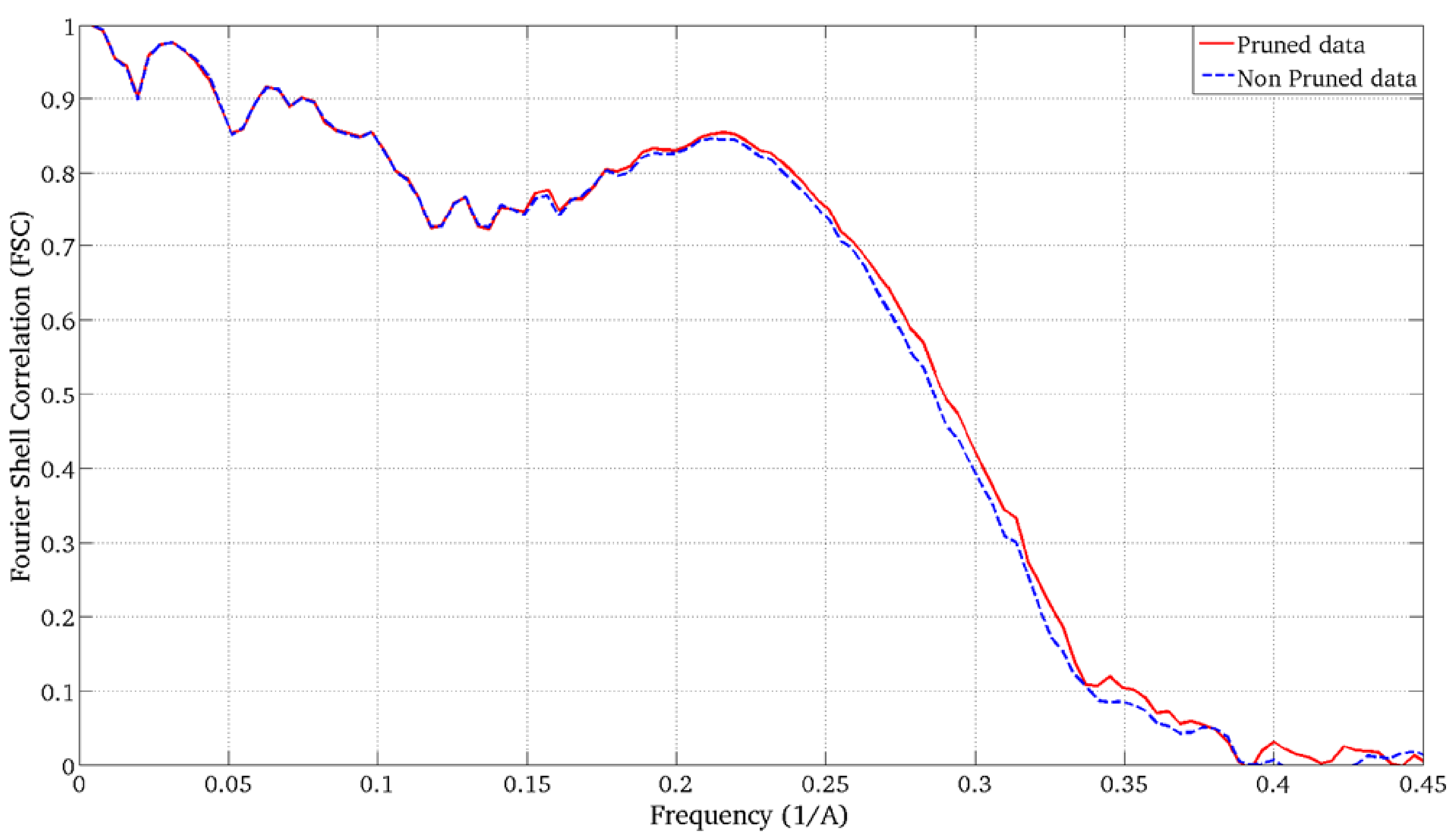
Angular Precision

(b)

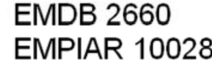
Soft-alignment validation for pruning

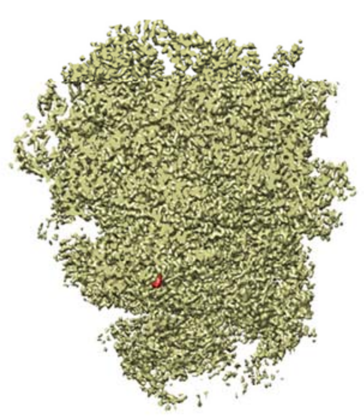


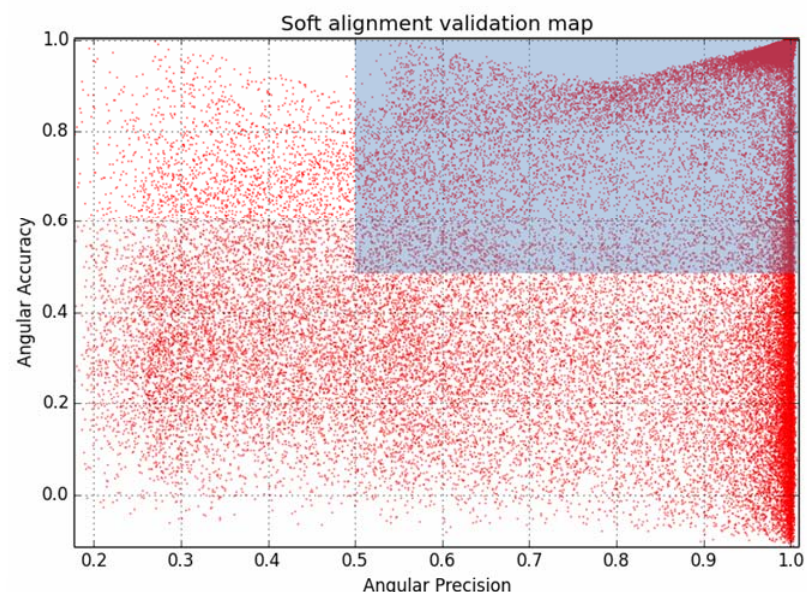
Soft-alignment validation for pruning



Soft-alignment validation for pruning







Precision (%)	97
Accuracy (%)	79
Precision & Accuracy (%)	79

Soft-alignment validation for pruning

Non Pruned data
Pruned data

Fourier Shell Correlation (FSC)

Frequency (1/A)

… and making everything simple to use!

Scipion Web Tools - West-Life

Single particle analysis tools

Align your movies — My movie alignment

Create your initial volume — My first map

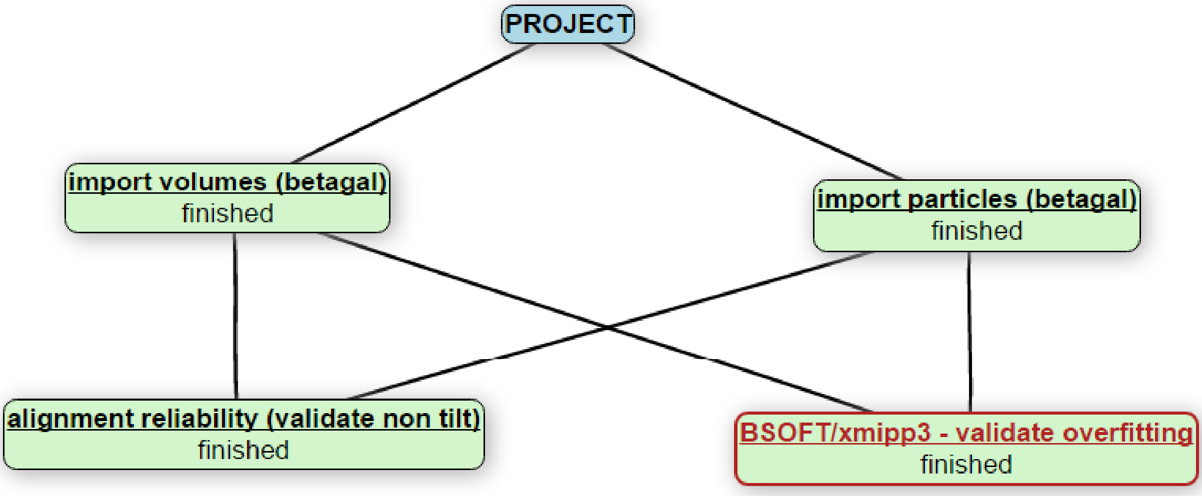
Analyze your maps — My resolution map

Reliability tools

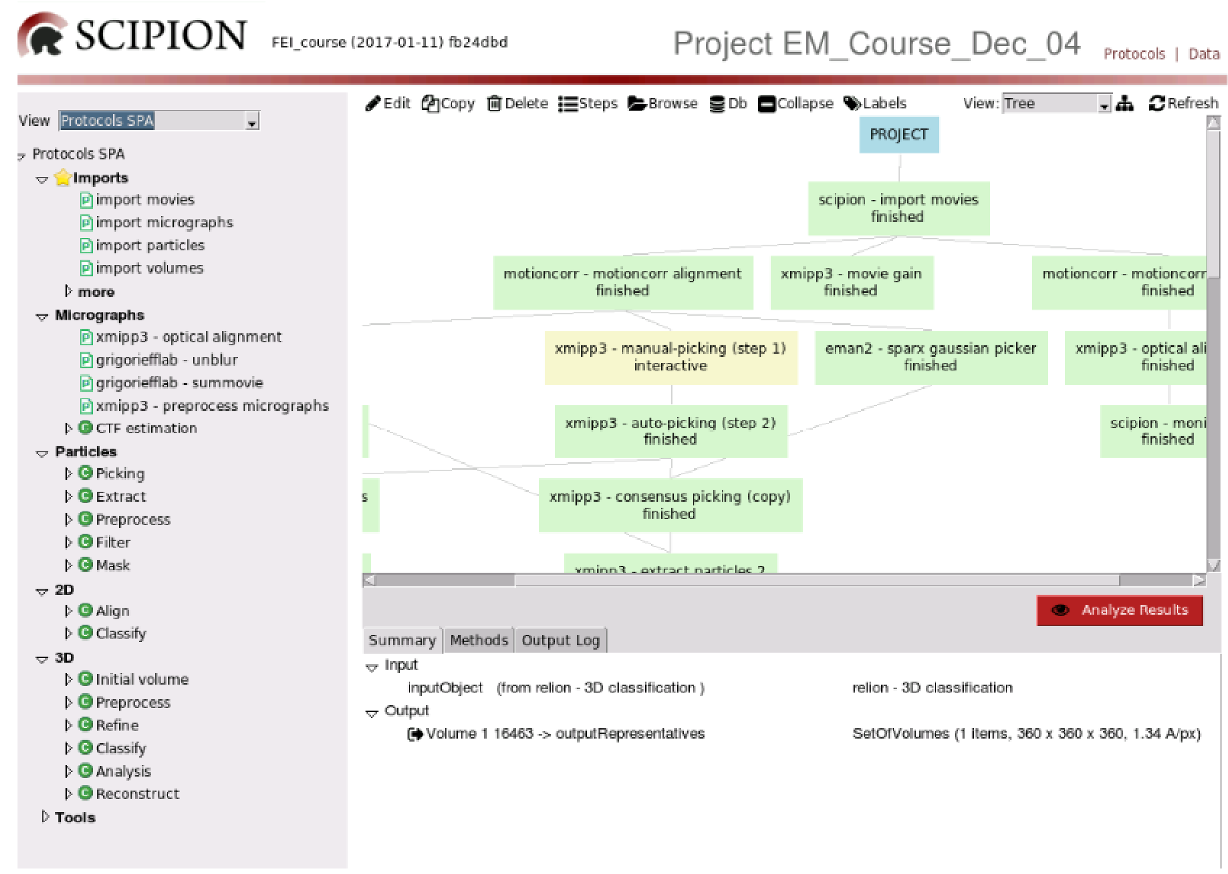
Explore protein interactions

Check the reliability of your data — My reliability tool

… and making everything simple to use!



… and making everything simple to use!



… and what about resolution? (a very "final" quality indicator)

- An ideal map should have the same "amount of detail" (resolution) in all directions, but reality is different…
  - Because of molecular flexibility
  - Because of alignement errors
- So, resolution is "local"

State of the Art

Currently, there are thre methods for estimating local resolution:

- Blocres: [Cardone2013] It was the piooner in the resolution methods and belongs to the bsoft package.

- ResMap: [Kulcukelbir2014] Released in 2014, currently it is the most popular method.

- MonoRes: [submitted] Recently developed and added into the Scipion image processing framework.

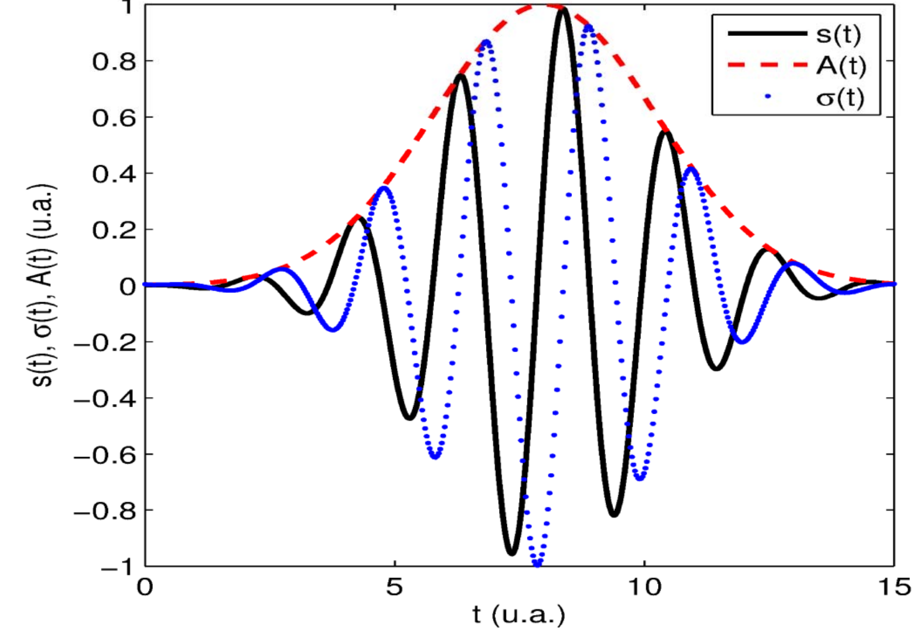
MonoRes

MonoRes: A new and ***fully automatic method*** for determining the local resolution of 3D maps

The root of this new algorithm is an extension of the concept of analytic signal, called monogenic signal. Thus, the map is filtered at different frequencies and the amplitude of monogenic signal is calculated, then a criteria for determining the resolution is computed.

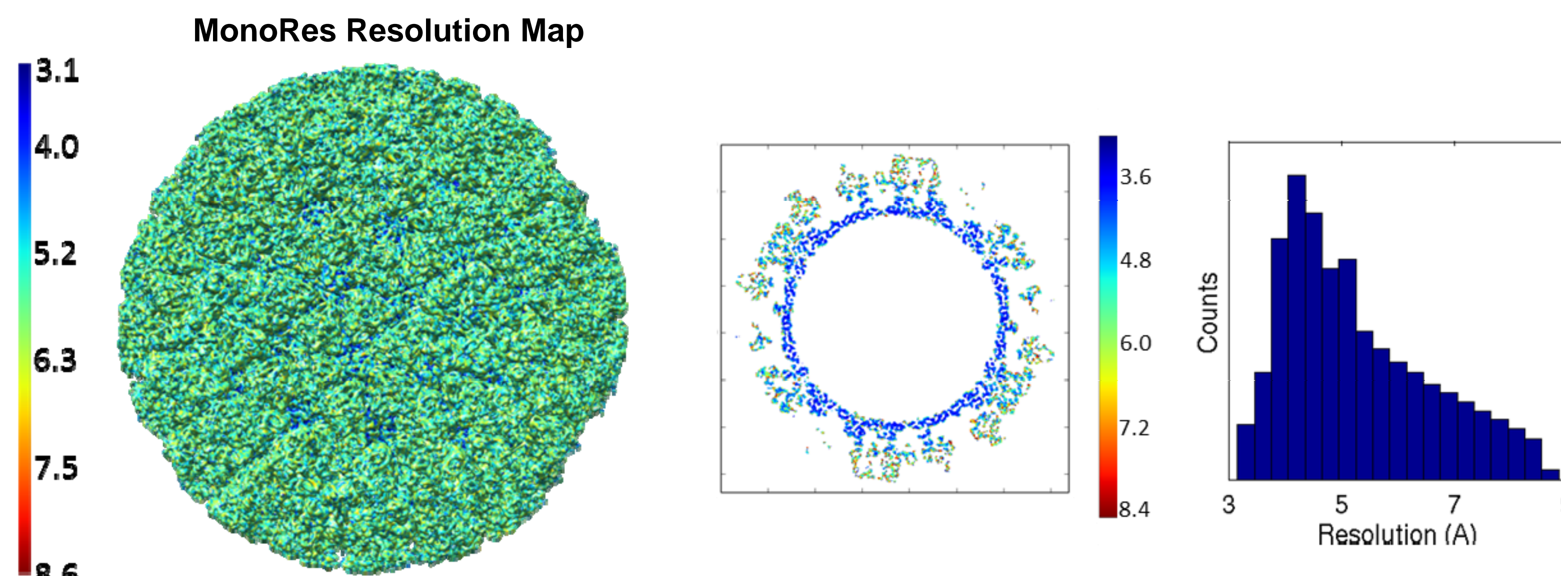
MonoRes

- The local energy is obtained from the transformed map (monogenic signal), which allows decomposing a function/map into envelop and phase terms.
- Local resolution is defined as the highest frequency at which local energy is significantly higher than noise



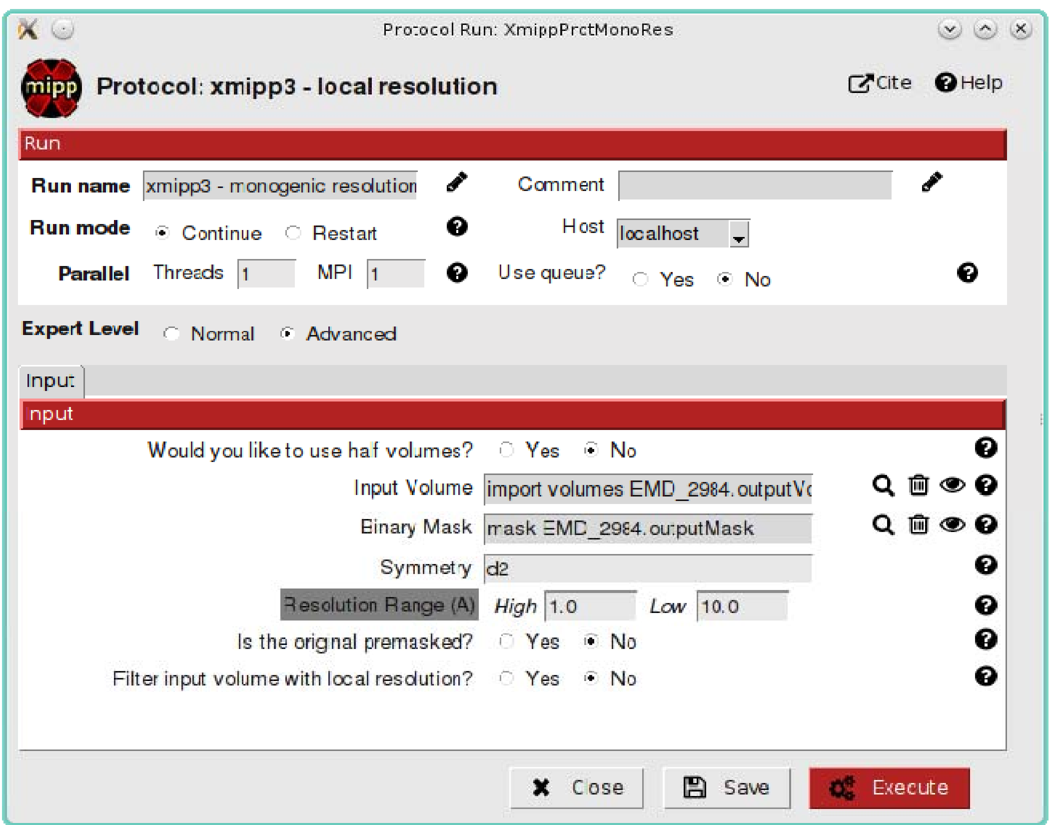
MonoRes

Aquaerovirus at 3.2 A (FSC at 0.143)
Size 760x760x760   MonoRes took ~4h



… and making everything simple to use!

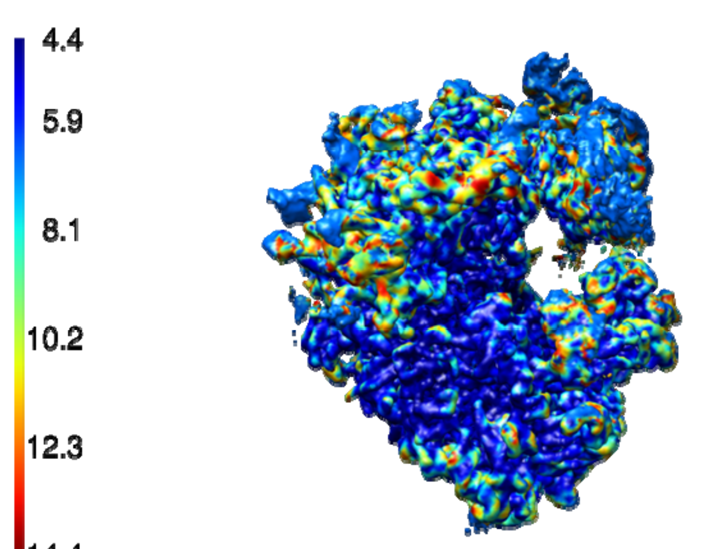
MonoRes is fully integrated in Scipion v1.1. Its interface is very simple

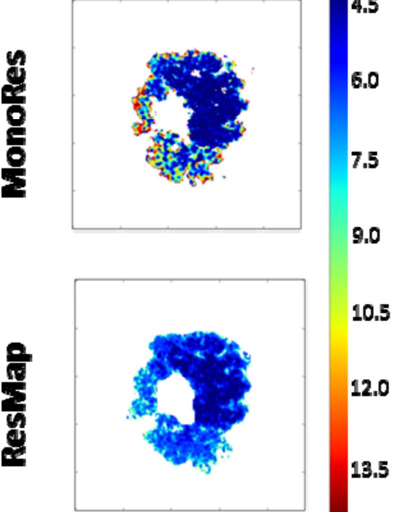


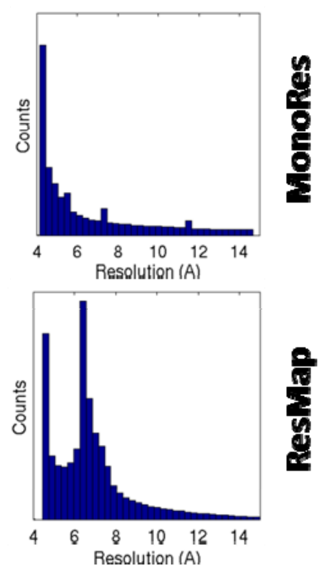
… and making everything simple to use!

MonoRes is also highly connected with chimera, generating a colored resolution map

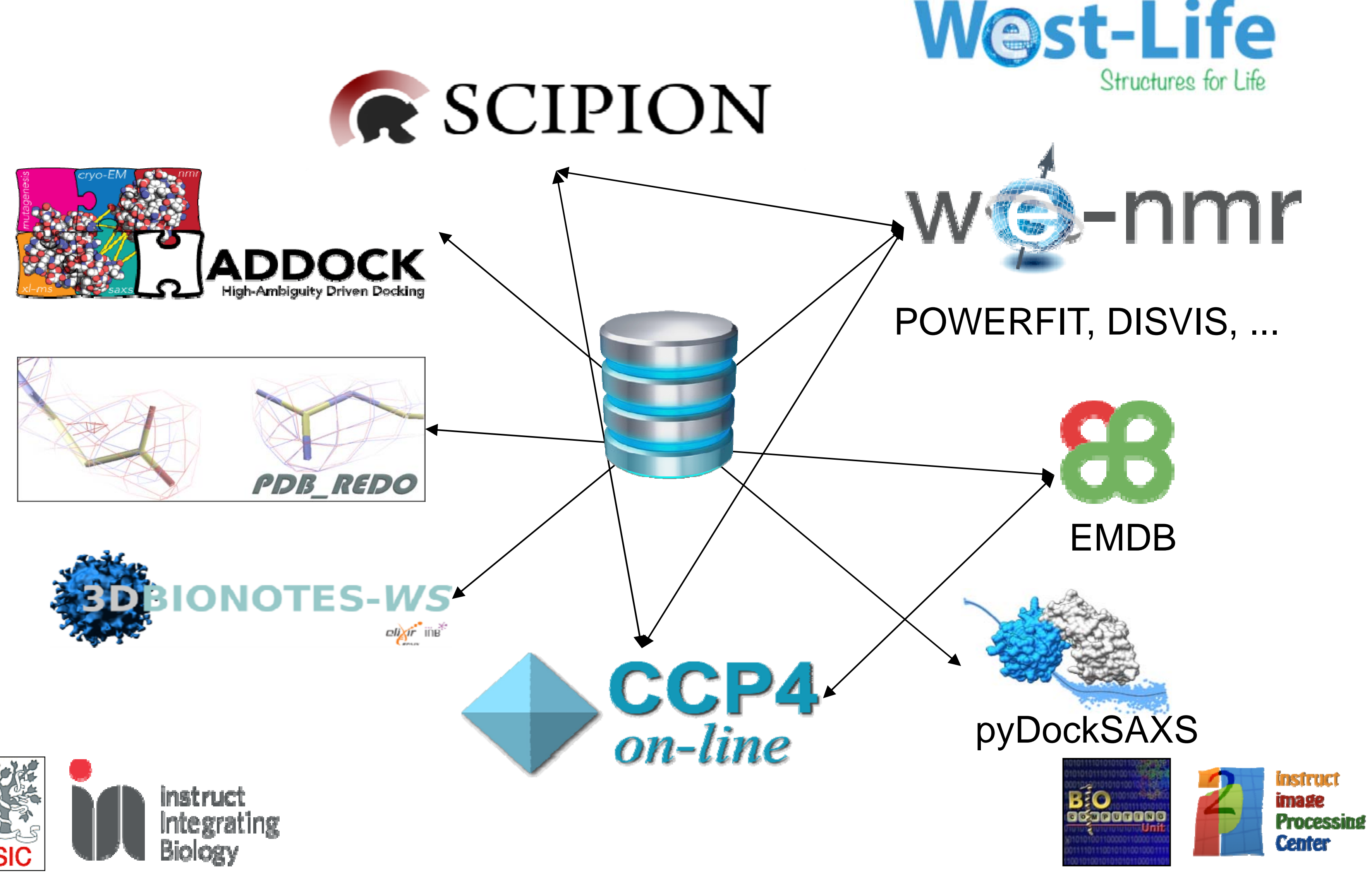
Ribosome at 5.5 A (FSC at 0.143)







Interconnecting Structural Biology resources: West-Life



An what about biological significance?: 3DBIONOTES, Enriching macromolecular data

EMBL-EBI | Services | Research | Training | About us

Protein Data Bank in Europe
Bringing Structure to Biology

Search
Examples: hemoglobin, BRCA1_HUMAN
Search EMDB

Feedback

EMDB › EMD-8576

Single particle reconstruction
3.2Å resolution

Structure of the Plasmodium falciparum 80S ribosome bound to the antimalarial drug mefloquine

Source organism: *Plasmodium falciparum 3D7* [36329]

Fitted atomic model: 5umd

3Dbionotes: available for this entry

To be published

Map released: 2017-03-01

Quick links

- **EMD-8576 overview**
- Function and Biology
- Experiments and Validation
- View
- Downloads
- Volume viewer
- Volume slicer
- Visual analysis

Related entries

Function and Biology — Details

Experimental Information — Details

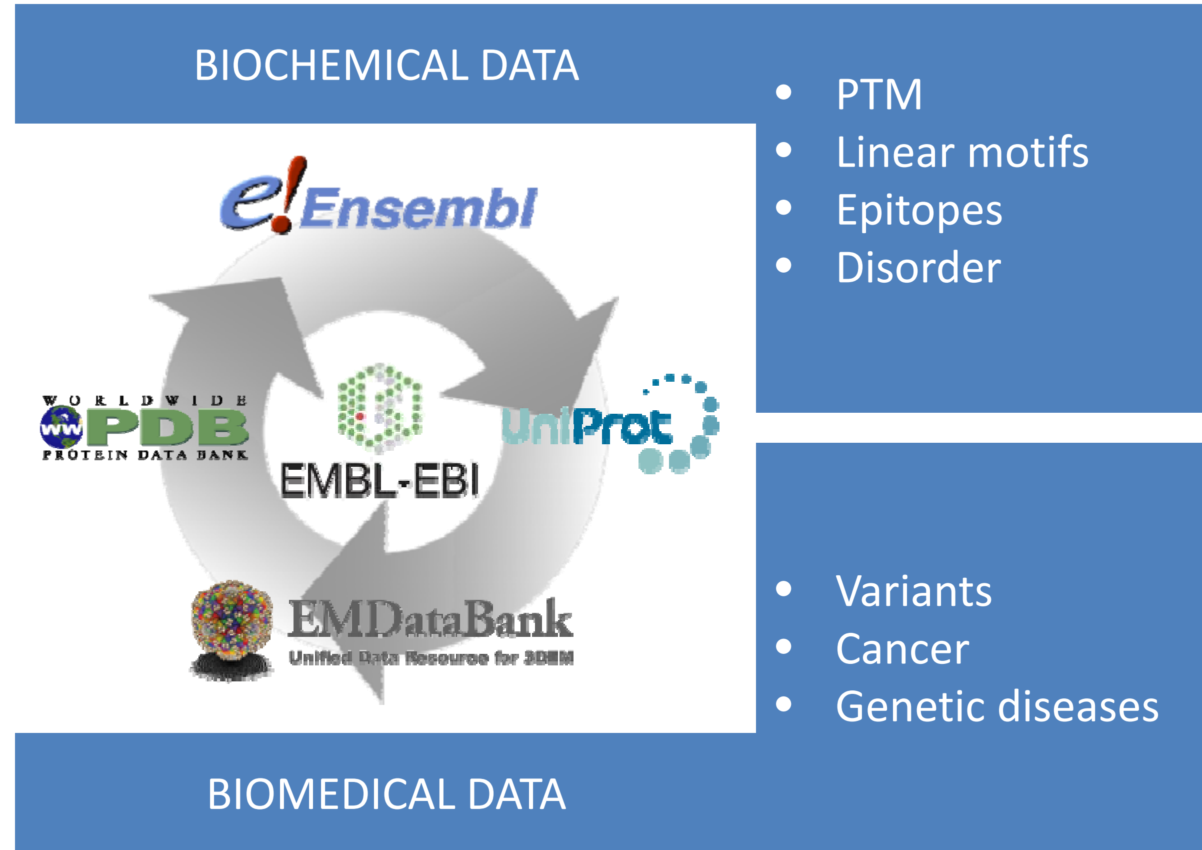
An what about biological significance?: 3DBIONOTES, Enriching macromolecular data

STRUCTURAL MODEL

DATABASE ENTRY

BIOCHEMICAL DATA

- PTM
- Linear motifs
- Epitopes
- Disorder



- Variants
- Cancer
- Genetic diseases

BIOMEDICAL DATA

3DBIONOTES: Enriching macromolecular data

- Single interface to display EMDB, PDB and UniProt entries
- Annotating structural models
- Protein annotations displayed at sequence and structural level
- Interactive annotation viewer
- Enriching data with other protein resources
- http://3dbionotes.cnb.csic.es

3DBIONOTES GUI

3DBIONOTES-WS HOME - SUBMIT - QUERY - HELP

CRYO-EM STRUCTURE OF THE HUMAN APC/C-CDH1-EMI1 TERNARY COMPLEX AT 3.6 ANGSTROM RESOLUTION.

EMD-2924 PROTEINS IN THIS MODEL: PDB:4UI9 CH:R Uniprot:Q9UM11 FIZZY-RELATED P

ATOMIC STRUCTURE

PROTEIN ANNOTATIONS

Domains & sites
Molecule processing
PTM
Sequence information
Mutagenesis
Domain families
Structure coverage
Proteomics
Antigenic sequences
Variants

PROTEIN SEQUENCE

Biomedical data

3DBIONOTES-WS HOME - SUBMIT - QUERY - HELP

CRYO-EM STRUCTURE OF THE HUMAN APC/C-CDH1-EMI1 TERNARY COMPLEX AT 3.6 ANGSTROM RESOLUTION.

EMD-2924 PROTEINS IN THIS MODEL: PDB:4UI9 CH:R Uniprot:Q9UM11 FIZZY-RELATED P

- Domains & sites
- Molecule processing
- PTM
- Sequence information
- Mutagenesis
- Domain families
- Structure coverage
- Proteomics
- Antigenic sequences
- Variants

Diseases

- Colon cancer
- Hematologic cancer
- Liver cancer
- Meningioma
- Oral cavity cancer
- Peripheral nervous syst...
- Skin cancer

Biochemical data

3DBIONOTES-WS HOME - SUBMIT - QUERY - HELP

elixir INB SPAIN

CRYO-EM STRUCTURE OF THE HUMAN APC/C-CDH1-EMI1 TERNARY COMPLEX AT 3.6 ANGSTROM RESOLUTION.

EMD-2924 PROTEINS IN THIS MODEL: PDB:4UI9 CH:R Uniprot:Q9UM11 FIZZY-RELATED P

MOD_RES_PHO 300:R

MOD_RES_PHO 343:R

MOD_RES_PHO 346:R

1 MDQDYERRLLRQIVIQNENTMPRVTEMRRTLTPASSPVSSPSKHGDRFIPSRAGANWSVNFHRINENEKS

SKHGDRFIPSRAGANWSVNFHRINEN

71 PSQNRKAKDATSDNGKDGLAYSALLKNELLGAGIEKVQDPQTEDRRLQPSTPEKKGLFTYSLSTKRSSPD

GLAYSALLKNELLGAGIEKVQ KGLFTYSL

141 DGNDVSPYSLSPVSNKSQKLLRSPRKPTRKISKIPFKVLDAPELQDDFYLNLVDWSSLNVLSVGLGTCVY

SPYSLSPVSNKSQKLLRS SKIPFKVLDAPELQD

50 100 150 200 250 300 350 400 450

283 366

290 300 310 320 330 340 350 360

- Domains & sites
- Molecule processing
- PTM
- Phosphorylation
- Acetylation
- SUMOylation
- Ubiquitination
- Glycosylation
- Sequence information
- Mutagenesis
- Domain families
- Structure coverage
- Proteomics
- Antigenic sequences
- Variants

ProtVista

290 300 310 320 330 340 350 360

… and making everything simple to use!

SCIPION

Scipion Web Tools - West-Life

Single particle analysis tools

Align your movies	My movie alignment
Create your initial volume	My first map
Analyze your maps	My resolution map

Reliability tools

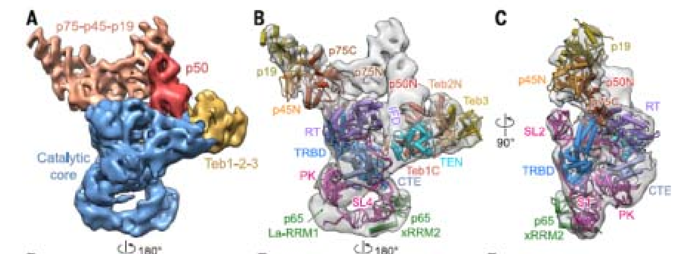
Explore protein interactions	DIANA WS
Check the reliability of your data	My reliability tool

… and what about biological significance?

- **Hybrid Models**:

  – 3DIANA:

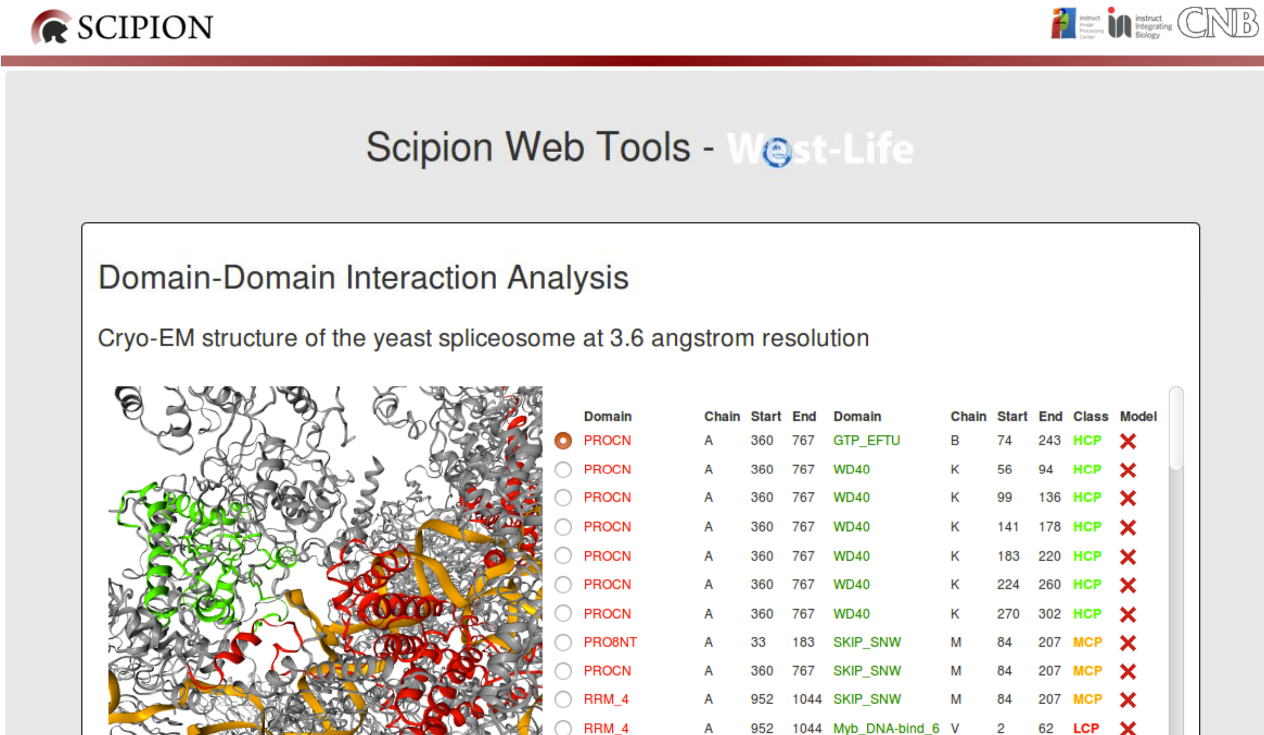
  Fitting of atomic resolution data medium_resolution EM maps



- A given hybrid model is assigned a reliability score by 3DIANA evaluating its likelihood in terms of the Domain-Domain-Interactions implied in the proposed hybrid model

Segura et al, BioPhysicis J., 2016

… and what about biological significance?



Domain	Chain	Start	End	Domain	Chain	Start	End	Class	Model
PROCN	A	360	767	GTP_EFTU	B	74	243	HCP	✗
PROCN	A	360	767	WD40	K	56	94	HCP	✗
PROCN	A	360	767	WD40	K	99	136	HCP	✗
PROCN	A	360	767	WD40	K	141	178	HCP	✗
PROCN	A	360	767	WD40	K	183	220	HCP	✗
PROCN	A	360	767	WD40	K	224	260	HCP	✗
PROCN	A	360	767	WD40	K	270	302	HCP	✗
PRO8NT	A	33	183	SKIP_SNW	M	84	207	MCP	✗
PROCN	A	360	767	SKIP_SNW	M	84	207	MCP	✗
RRM_4	A	952	1044	SKIP_SNW	M	84	207	MCP	✗
RRM_4	A	952	1044	Myb_DNA-bind_6	V	2	62	LCP	✗
RRM_4	A	952	1044	HTH_AsnC-type	V	71	94	NS	✗
U5_2-snRNA_bdg	A	1177	1310	CwfJ_C_2	Y	199	300	MCP	✗
U6-snRNA_bdg	A	1409	1562	CwfJ_C_2	Y	199	300	MCP	✗

And the people behind…….

- Carlos Oscar Sanchez-Sorzano and Roberto Marabini
- Javier Vargas
- Joan Segura
- Jose Luis Vilas and Ruben Sanchez
- Pablo Conesa, Jose Miguel de la Rosa and Scipion team
- And all the talented predocs, postdocs and engineers along the years!

And the people to come…

Always looking for talented engineers, predocs and postdoc fellows!

Instruct Biennial Structural Biology Meeting

WEST-LIFE Instruct Meeting

Brno, Czech Republic

24th to 26th May 2017

Let meet at Brno!

#IBSBM2017

Instruct Biennial Meeting

24th - 26th May, 2017 Brno, Czech

Showcasing Integrative Structural Biology

The third Instruct Biennial Structural Biology Meeting will be held at the International Best Western Hotel, Brno, Czech Republic. The meeting will showcase integrative structural biology and its impact on biological research. The program will include sessions that represent recent advances in structural biology towards cellular biology, emerging methods and technologies and results of biomedical importance.
Young scientists who want to know what structural biology can offer for their research and what are the future trends, are especially encouraged to attend.

Register today at www.structuralbiology.eu/ibsbm2017