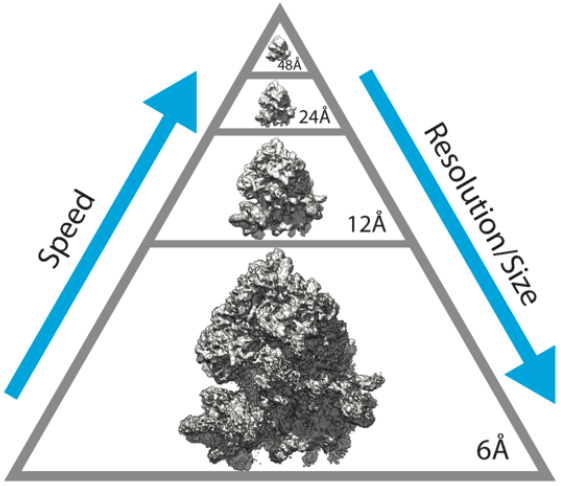
DisVis and PowerFit: Explorative and Integrative Modeling of Biomolecular Complexes harvesting EGI GPGPU resources

Alexandre Bonvin
Utrecht University

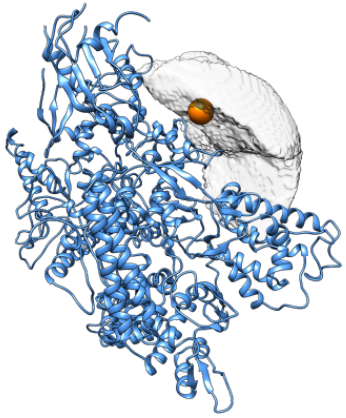


PowerFit: fast and sensitive rigid body fitting in lower-resolution densities

Van Zundert & Bonvin,J. Proteome. Res. 2015

DisVis: quantification and visualization of the information content of distance restraints

Van Zundert & Bonvin, Bioinformatics 2015



Cryo-electron microscopy: The rising star in structural biology

West-Life
Structures for Life

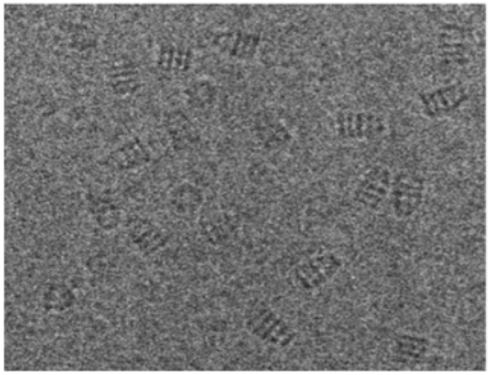
Transmission electron microscopy

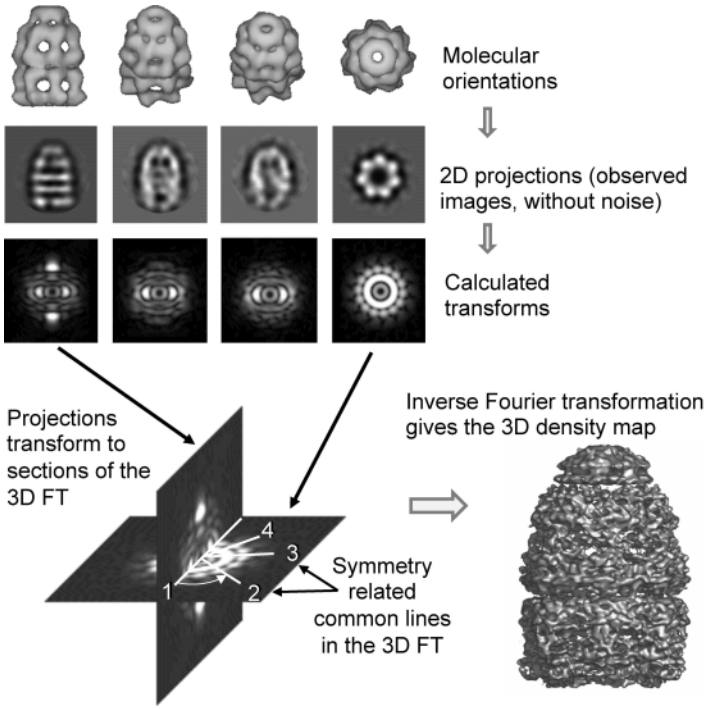
Sample is vitrified in liquid ethane

Images noisy 2D projections

Class averages are constructed for each orientation

Ultimately results in 3D density maps



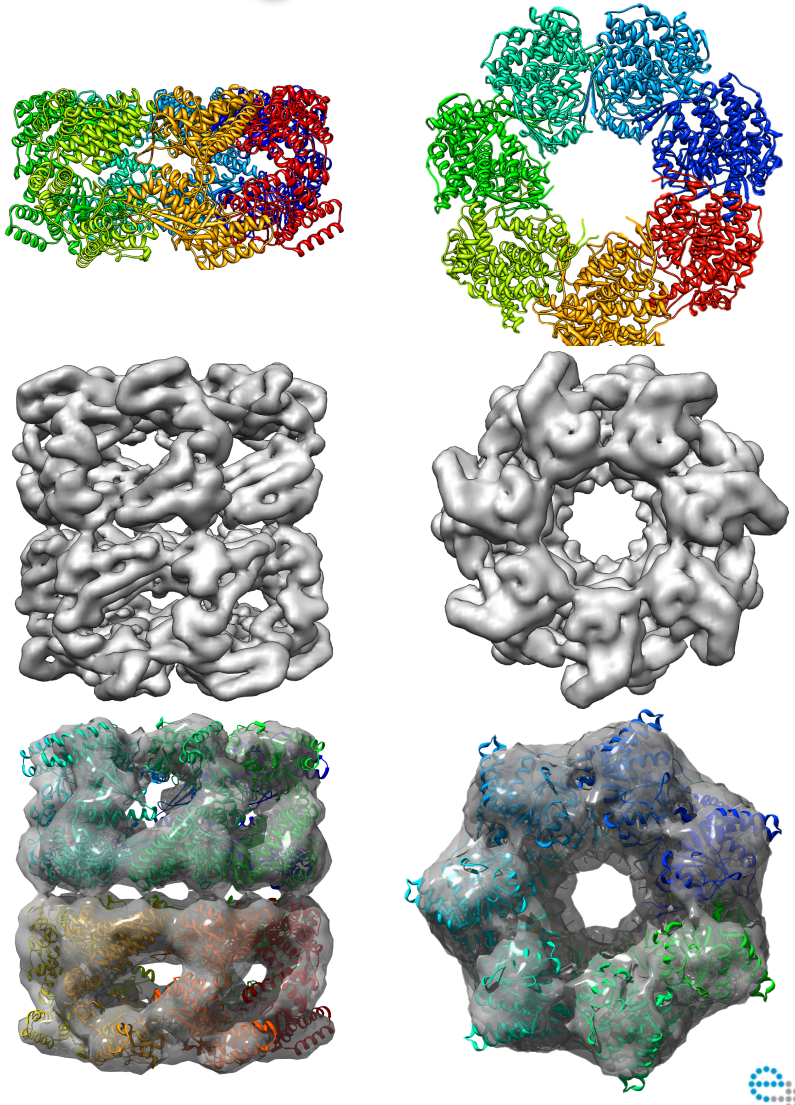


INDIGO - DataCloud
EGI MoBrain
e-infrastructure

Cryo-EM data: "high" resolution modelling

- **Rigid body fitting**
  - Manual fitting (UCSF Chimera)
  - Automatic fitting software (CoLoRes, PowerFit, Mod-EM)
  - In most cases, does not take into account the **flexibility and energetics of the interface**

- **Flexible fitting**
  - Requires an **unambiguous fit** of the subunits
  - Various approaches (e.g. normal modes, flexible refinement (Flex-EM))
  - The applicable resolution extend is debatable
  - **Overfitting** is an issue
  - Often does not take into account **other sources of data** (mutagenesis, etc.)



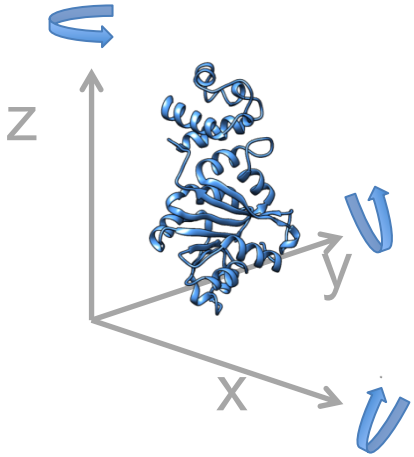
PowerFit:

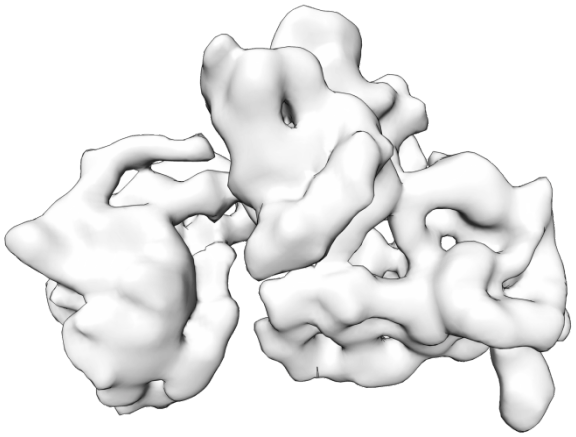
Cross-correlation based rigid body fitting

Automatic rigid body fitting of biomolecular structures in cryo-EM densities

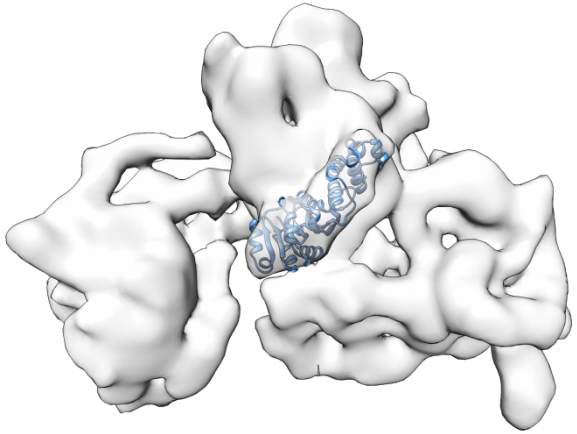
Exhaustive 6D search of 3 translational and 3 rotational degrees of freedom

Calculate cross-correlation at every scanned position





6D cross-correlation search



Issues in the field:

- **Sensitivity**
- **Speed**

Van Zundert & Bonvin
J. Proteome. Res. (2015)
J. Struct. Biol. (2016)

DisVis: Defining the information content and consistency of distance restraints

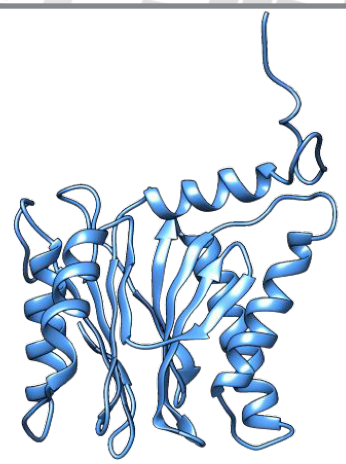
Given 2 interacting structures and a set of distance restraints between them, are there any solutions that satisfy *N* restraints?

A solution is a complex that satisfies all N distance restraints

A complex is a conformation where:
The subunits are interacting
The subunits are not clashing

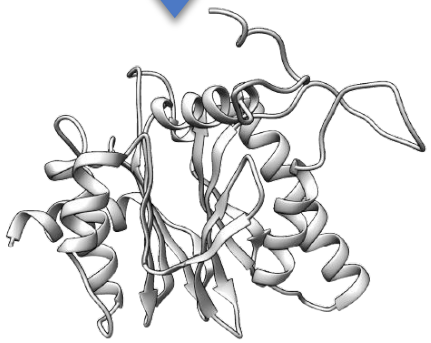
*The **accessible interaction space** is the set of all solutions satisfying at least N restraints*

DisVis



fixed chain

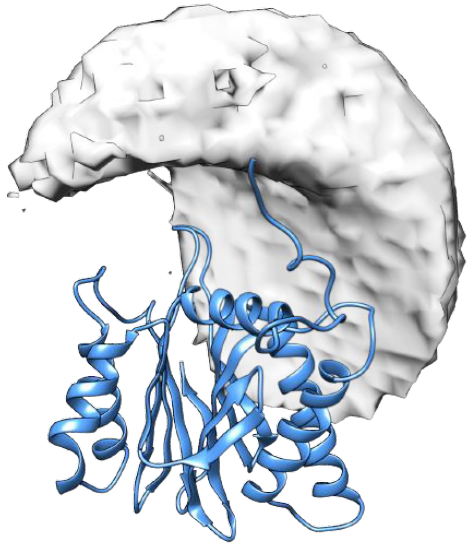
Distance Restraints



scanning chain

6D search of accessible interaction space

At every grid position, save the maximum number of consistent restraints found during the 6D search

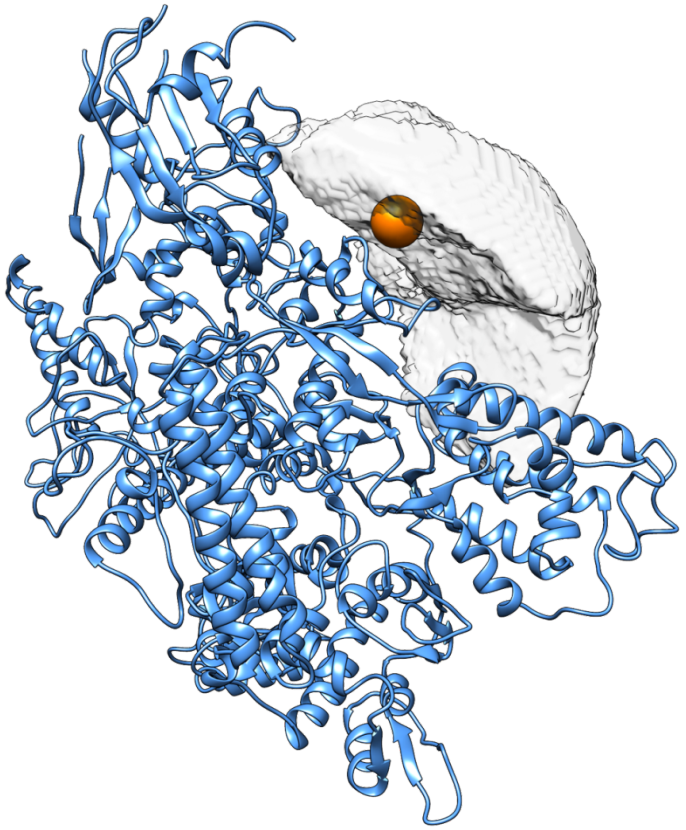


accessible interaction space + restraint evaluation

Van Zundert & Bonvin
Bioinformatics 2015

RNA-polymerase II: Accessible interaction space

Number of consistent restraints (N)	Number of accessible complexes consistent with at least N restraints	Fraction of accessible complexes consistent with at least N restraints
0	18940752204	1.0000
1	2370295166	0.1251
2	977410985	0.0516
3	298922038	0.0158
4	92651659	0.0049
5	17687776	0.0009
6	5172437	0.0003
7	9716	0.0000
8	0	0.0000



DisVis 6D systematic search with a 1Å grid size and 5.27° interval

Speeding up the search

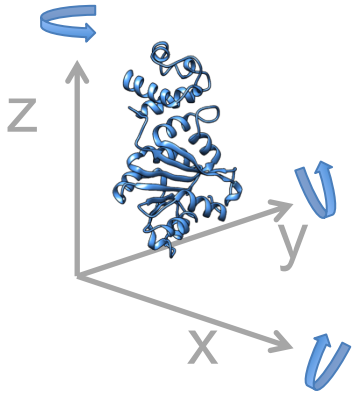
$$\text{CW-LCC} = \frac{1}{N} \frac{\sum_i^N \rho_c^n \cdot w_i \rho_o}{\sqrt{\left(\overline{\rho_o^w}\right)^2 - \overline{\left(\rho_o^w\right)^2}}}$$

$$\text{CW-GCC} = \mathcal{F}^{-1}\left[\mathcal{F}\left(w\rho_c^n\right)^* \times \mathcal{F}\left(\rho_o\right)\right]$$

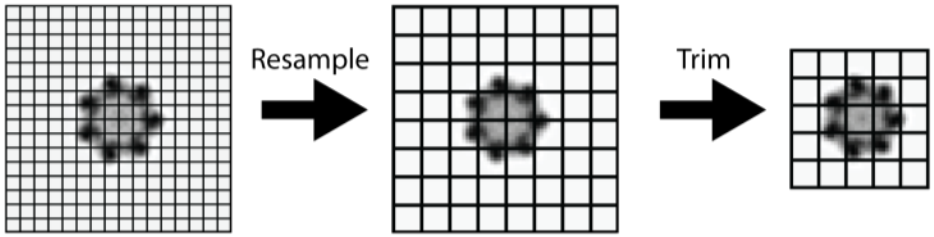
$$\left(\overline{\rho_o^w}\right)^2 = \mathcal{F}^{-1}\left[\mathcal{F}\left(w\right)^* \times \mathcal{F}\left(\rho_o\right)\right]^2$$

$$\overline{\left(\rho_o^w\right)^2} = \mathcal{F}^{-1}\left[\mathcal{F}\left(w^2\right)^* \times \mathcal{F}\left(\rho_o^2\right)\right]$$

Fast Fourier Transform for fast translational scans



Optimized rotation sets

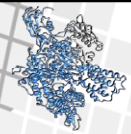
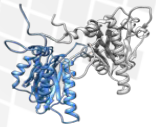


Resampling and trimming target

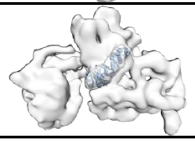
GPU acceleration

GPU-acceleration

Disvis

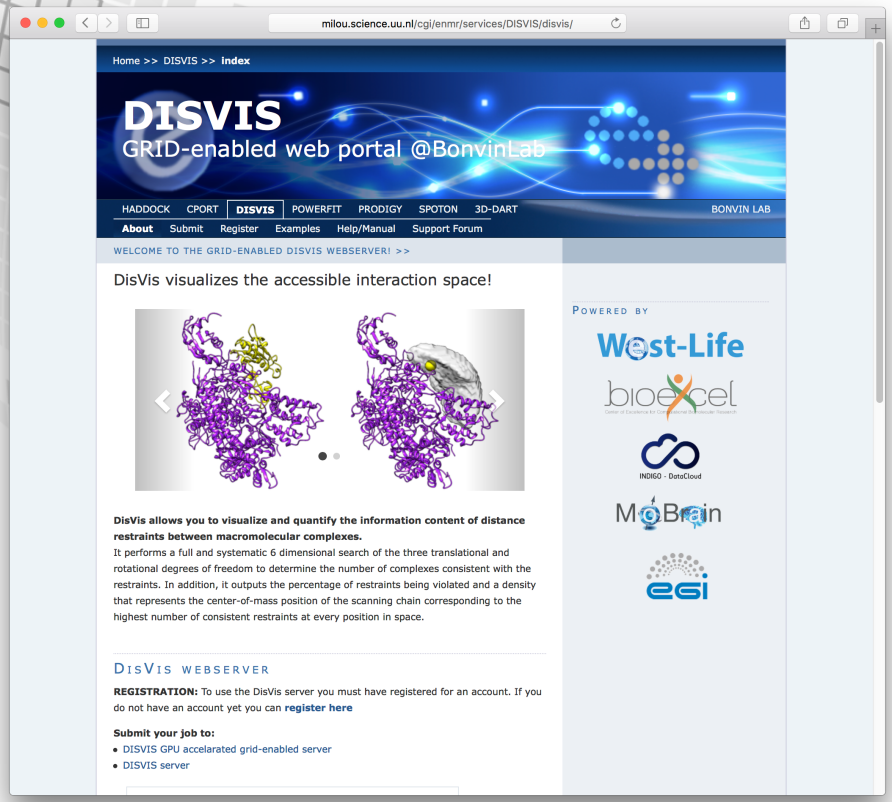
System		Number of complexes sampled	Time CPU	Time GPU	Speedup
	RNA-polymerase II	19×10^9	19h 44m	56m	**21x**
	PRE5-PUP2	7×10^9	7h 12m	15m	**29x**

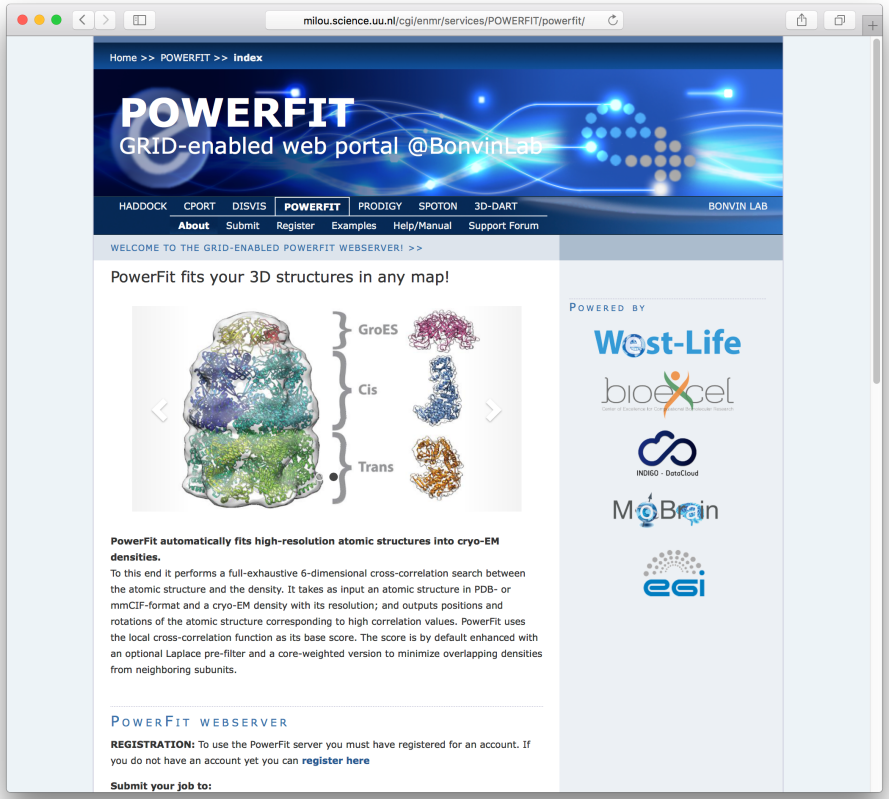
PowerFit

System		Map size (voxels)	Rotations sampled	Time CPU	Time GPU	Speedup
	GroEL-GroES	90 x 72 x 72	70728	1h 29m	4m 9s	**21x**
	RsgA into ribosome	72 x 80 x 72	70728	1h 16m	4m 2s	**19x**

(in house GTX680 card vs single core)

GRID-enabled web portals





http://milou.science.uu.nl/enmr/services/DISVIS/

http://milou.science.uu.nl/enmr/services/POWERFIT/

Architecture behind the portals

User DB

User not found

Input error

Validation

Pre-processing
+
Input files packaging

Submission to local nodes

CPU-calculation

OR

Submission to grid node

Output files packaging
+
submission of image generation

Chimera image generation

Post-processing
+
Results formatting

WEB CLIENT

WEB SERVER

MASTER NODE

WORKING NODE

Software Provisioning

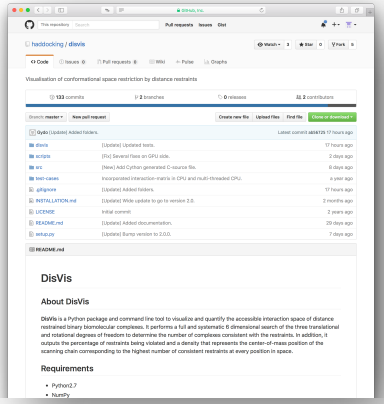
Because of complex software dependencies we use docker containers

- Python2.7
- NumPy 1.8+
- SciPy
- FFTW3
- pyFFTW 0.10+
- OpenCL1.1+
- pyopencl
- clFFT
- gpyfft

And to avoid security issues on the grid side, udocker from INDIGO

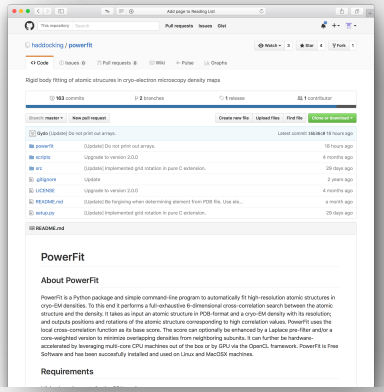
GitHub

INDIGO - DataCloud



docker

indigodatacloudapps/**disvis**



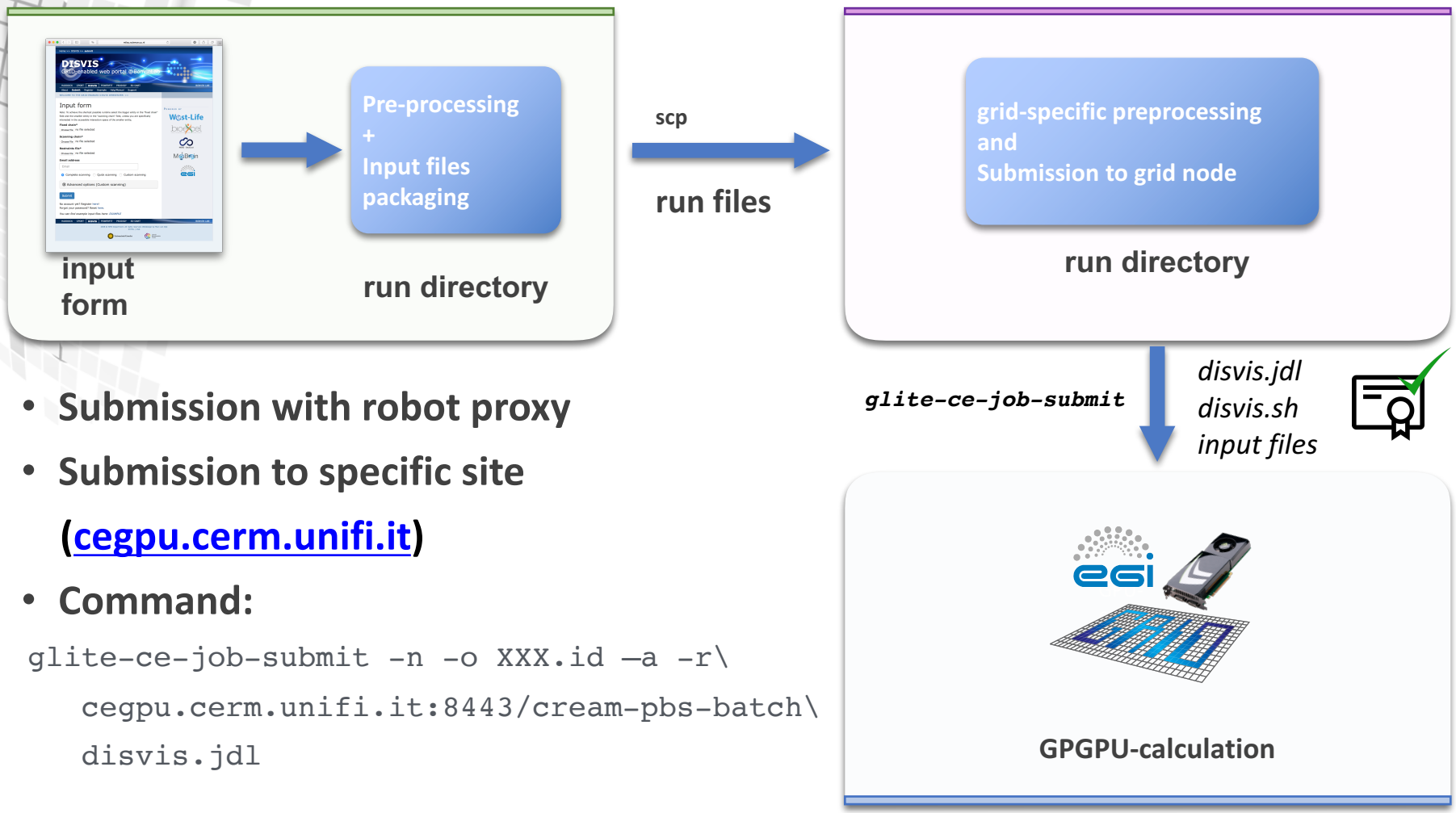
docker

indigodatacloudapps/**powerfit**

INDIGO - DataCloud

e-infrastructure

Grid submission



- **Submission with robot proxy**
- **Submission to specific site (cegpu.cerm.unifi.it)**
- **Command:**

```
glite-ce-job-submit -n -o XXX.id –a -r\
    cegpu.cerm.unifi.it:8443/cream-pbs-batch\
    disvis.jdl
```

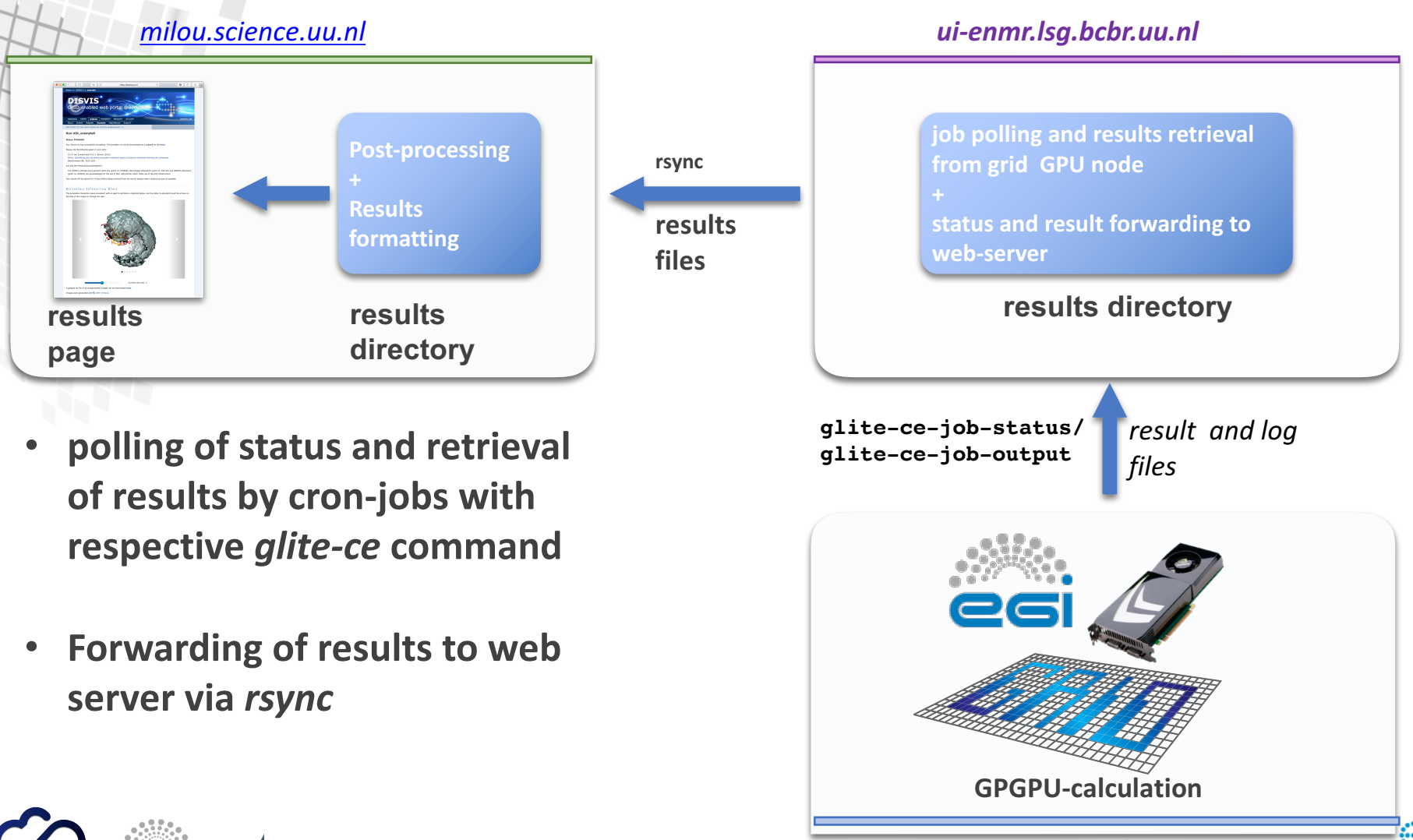
JDL-File

```
[
  executable = "disvis.sh";
  inputSandbox = { "aac_3mizzqnQ/disvis.sh",
                   "aac_3mizzqnQ/fixed_chain.pdb",
                   "aac_3mizzqnQ/scanning_chain.pdb",
                   "aac_3mizzqnQ/restraints.txt" };
  stdoutput = "disvis.out";
  outputsandboxbasedesturi = "gsiftp://localhost";
  stderror = "disvis.err";
  outputsandbox = { "disvis.out" , "disvis.err" , "aac_3mizzqnQ.tgz"};
  GPUNumber=1;
]
```

Executable

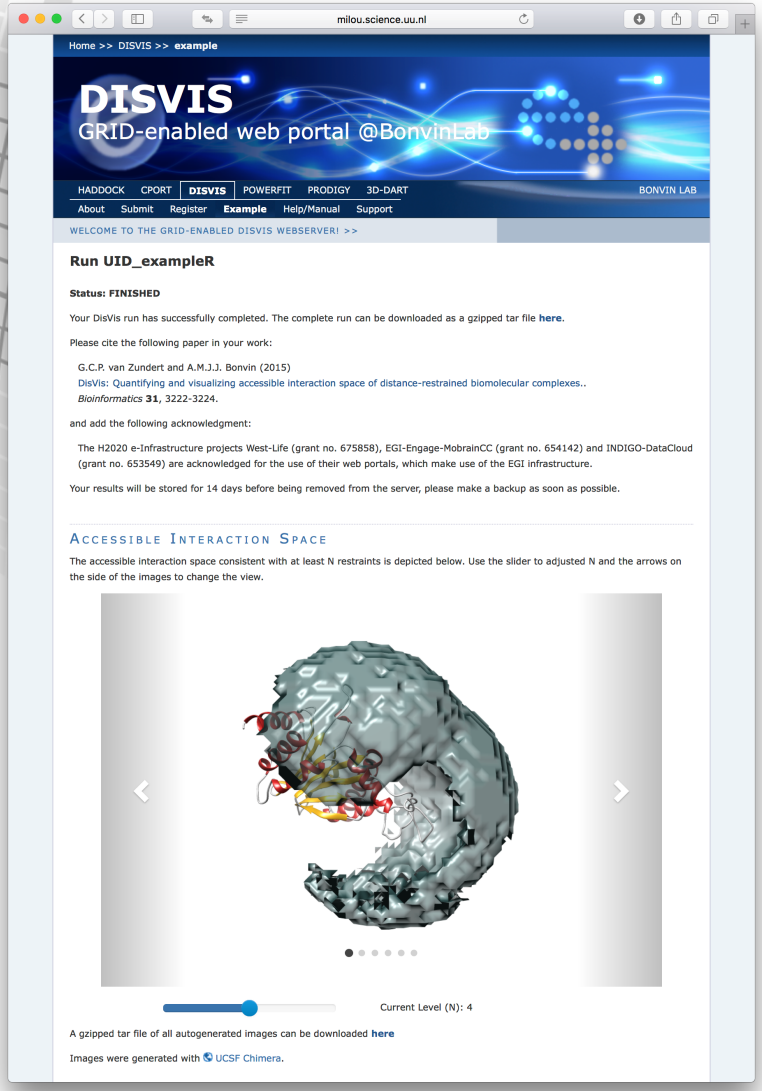
```
#!/bin/sh
driver=$(nvidia-smi | awk '/Driver Version/ {print $6}')
…
# Untar input files archive
tar xzf $jobid.tgz
…
# Get udocker from github
git clone https://github.com/indigo-dc/udocker
cd udocker
# Download docker image from indigodatacloud
./udocker.py pull indigodatacloudapps/disvis:nvdrv_$driver
rnd=$RANDOM
# Create temporary instance of docker image
./udocker.py create --name=disvis-$rnd indigodatacloudapps/disvis:nvdrv_$driver
#
mkdir $WDIR/results_$jobid
# Run disvis within the docker image
./udocker.py run -v /dev --volume=$WDIR:/home disvis-$rnd disvis /home/$pdb_A /home/$pdb_B \
                    /home/restraints.txt -g -a $a -vs $vs -ir $ir -cv $cv -iv $iv $is $oa \
                    –d /home/results_$jobid
…
# Archive results
tar cfz $jobid.tgz results_$jobid
```

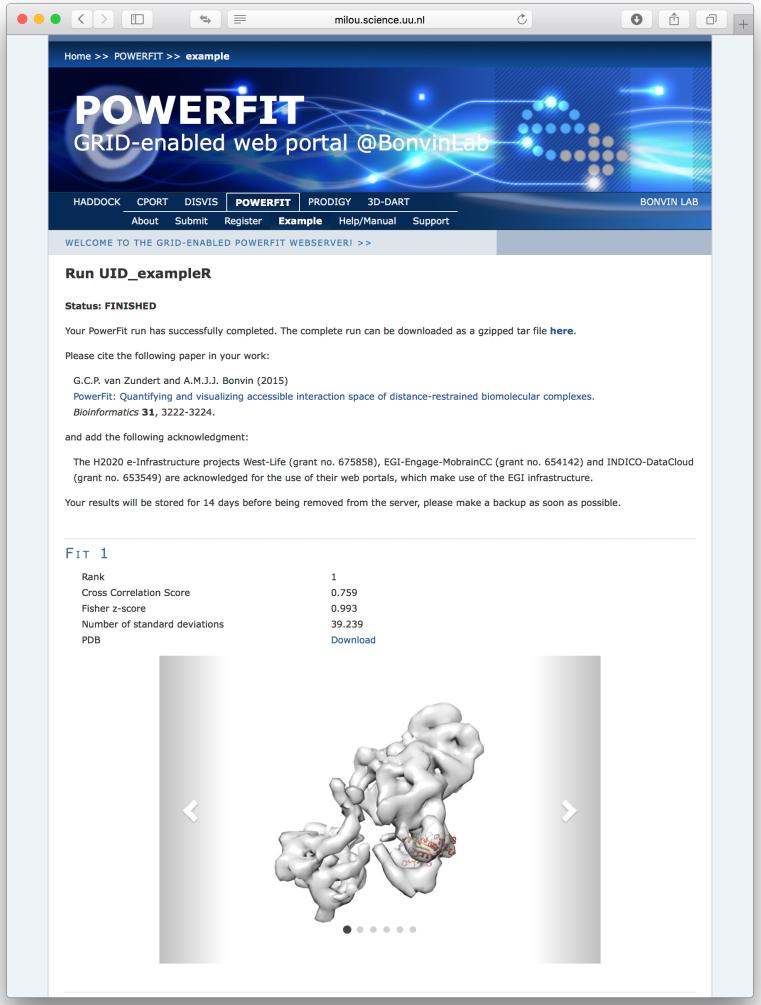
Results retrieval



- polling of status and retrieval of results by cron-jobs with respective *glite-ce* command
- Forwarding of results to web server via *rsync*

Rich result pages





Guided interpretation of results

z-Score

The table below features the z-Score for each restraint. The higher the score, the more likely the restraint is a false-positive. Z-scores above 1.0 are explicitly mentioned in the output of DisVis.

#	Restraint	Average violated fraction	Standard deviation	Z-score
8	**A1092(CB)-E152(CB)**	**1.00**	**0.01**	**2.05**
7	**A180(CB)-E122(CB)**	**0.80**	**0.33**	**1.29**
4	A15(CB)-E171(CB)	0.39	0.30	-0.29
5	A934(CB)-E201(CB)	0.38	0.29	-0.35
6	A938(CB)-E201(CB)	0.3		
2	A129(CB)-E161(CB)	0.2		
1	A1003(CB)-E166(CB)	0.2		
3	A129(CB)-E171(CB)	0.2		

Solutions

The table below lists the 15 best non-redundant solutions found by correlation score. The first column shows the rank, column 2 the correlation score, column 3 the Fisher z-score column 4 the zscore as factor of standard deviations (z/σ), and column 5 the sigma difference to the best fit. (see N. Volkmann 2009, and Van Zundert and Bonvin 2016).

Rank (N)	Cross Correlation Score	Fisher z-score	z-score/σ	Sigma difference $(z_1-z_N)/\sigma$
1	**0.560**	**0.633**	**31.0**	**0.00**
2	0.517	0.573	28.0	2.98
3	0.432	0.463	22.7	8.35
4	0.301	0.311	15.2	15.79
5	0.298	0.307	15.1	15.97
6	0.298	0.307	15.0	15.99
7	0.293	0.302	14.8	16.24
8	0.290	0.298	14.6	16.42
9	0.288	0.297	14.5	16.49
10	0.288	0.296	14.5	16.53

Some usage stats

Operational since Aug. 2016

DISVIS WEBSERVER STATS

Number of users: 70

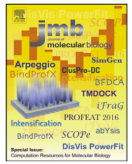
Total number of served requests as of August 1st 2016: 287 of which 113 on GPGPU-enabled EGI grid sites

POWERFIT WEBSERVER STATS

Number of users: 59

Total number of served requests as of August 1st 2016: 103 of which 53 on GPGPU-enabled grid sites

Published Dec. 2016

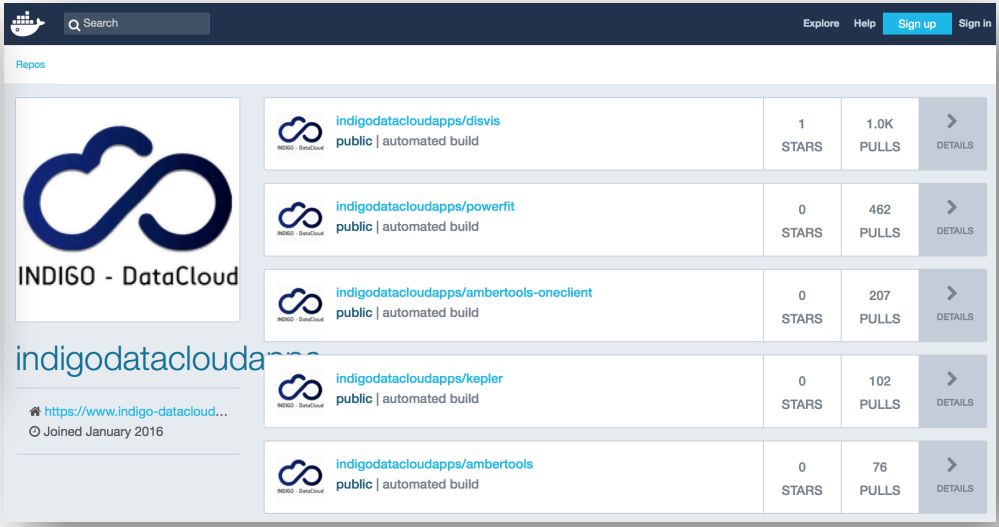


The DisVis and PowerFit Web Servers: Explorative and Integrative Modeling of Biomolecular Complexes

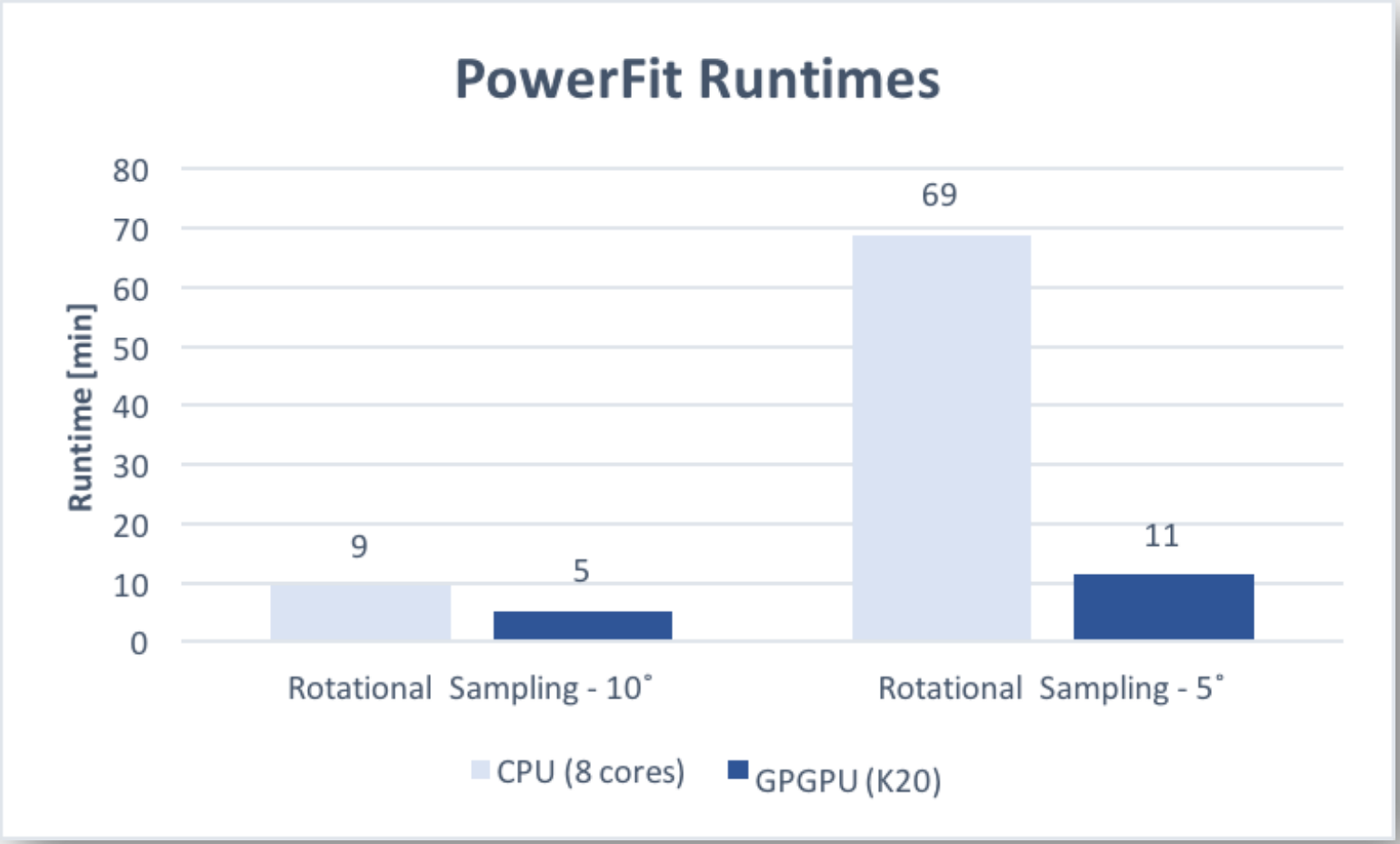
G.C.P. van Zundert[1], M. Trellet[1], J. Schaarschmidt[1], Z. Kurkcuoglu[1], M. David[2], M. Verlato[3], A. Rosato[4] and A.M.J.J. Bonvin[1]

*1 - **Bijvoet Center for Biomolecular Research,** Faculty of Science – Chemistry, Utrecht University, Padualaan 8, 3584CH Utrecht, The Netherlands*
*2 - **LIP - Laboratório de Instrumentação e Física Experimental de Partículas,** Avenida Elias Garcia 14, 1000 Lisbon, Portugal*
*3 - **Istituto Nazionale di Fisica Nucleare - Sezione di Padova,** Via Marzolo 8, 35131 Padova, Italy*
*4 - **Magnetic Resonance Center and Department of Chemistry "Ugo Schiff",** University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy*

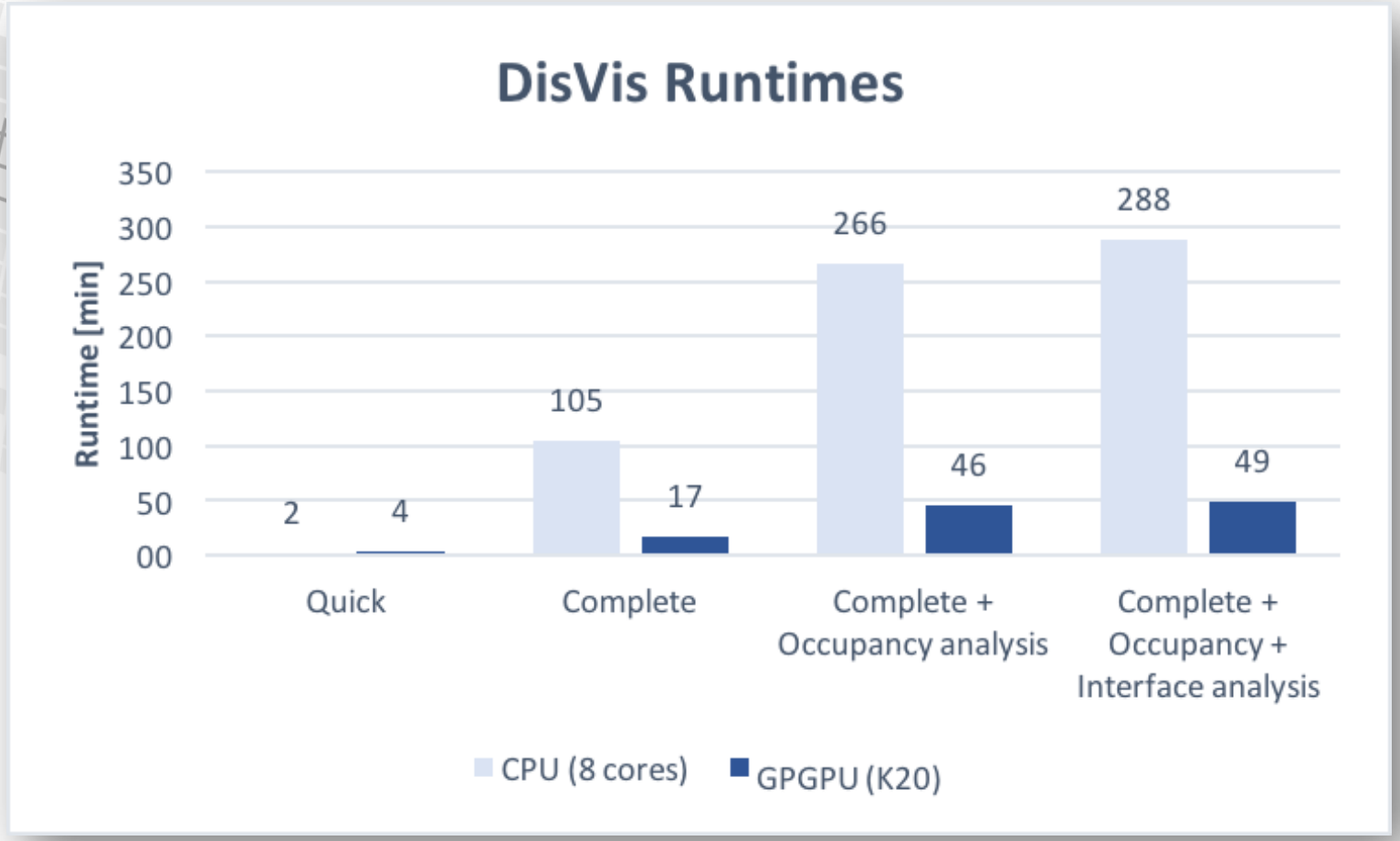
Top pulls in INDIGO docker hub



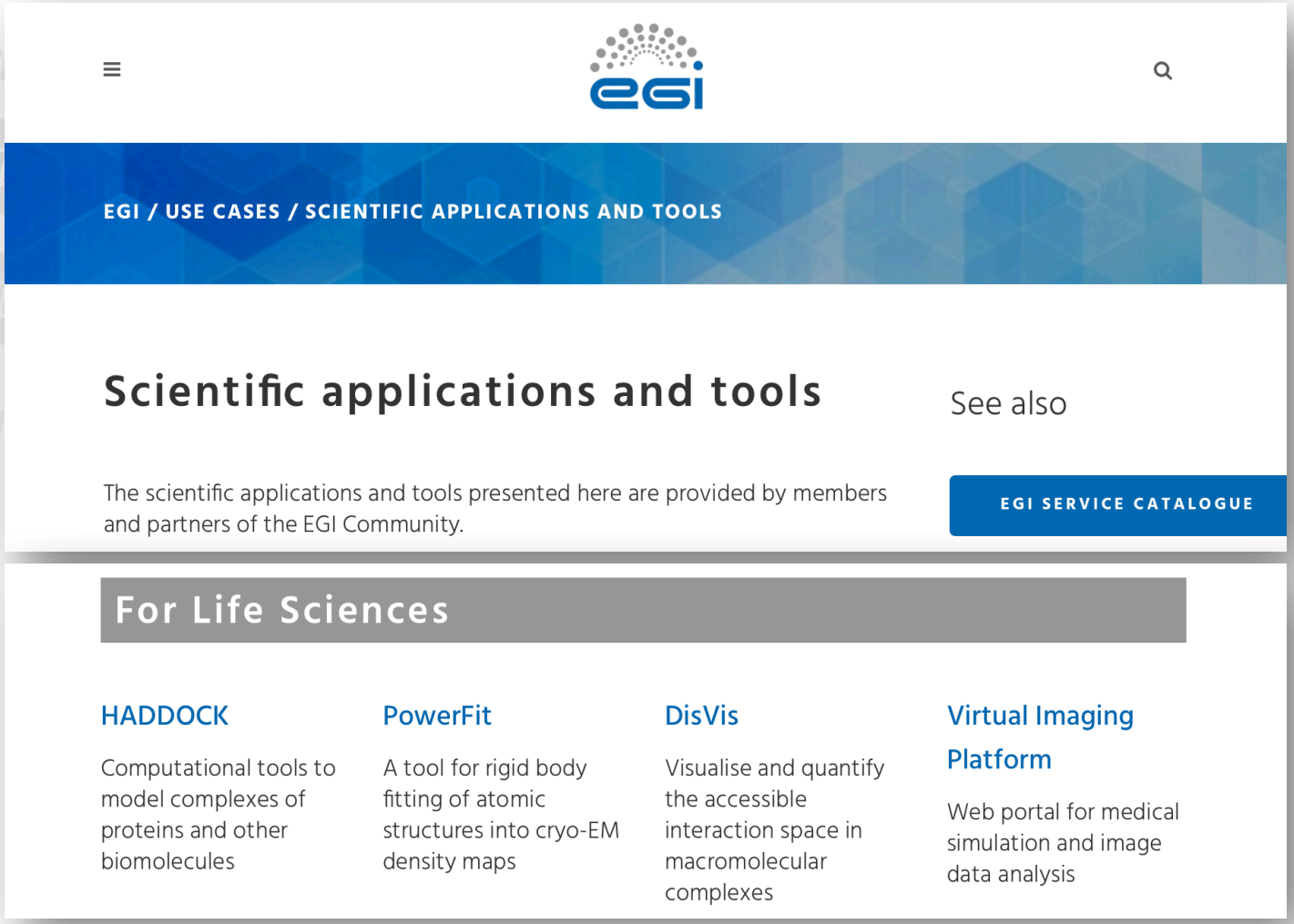
Some timing stats



Some timing stats



Listed in EGI applications



EGI / USE CASES / SCIENTIFIC APPLICATIONS AND TOOLS

Scientific applications and tools

The scientific applications and tools presented here are provided by members and partners of the EGI Community.

See also

EGI SERVICE CATALOGUE

For Life Sciences

HADDOCK

Computational tools to model complexes of proteins and other biomolecules

PowerFit

A tool for rigid body fitting of atomic structures into cryo-EM density maps

DisVis

Visualise and quantify the accessible interaction space in macromolecular complexes

Virtual Imaging Platform

Web portal for medical simulation and image data analysis

Conclusions

- **Two new, user-friendly web portals in operation**
- **Harvest grid GPGPU resources**
- **Complex software provisioning via Docker**
- **Build on INDIGO udocker solution**

- **Current issues**
  - Limited resources (only one site in production)
  - Automatic updates of nvidia drivers on sites must be coupled to automatic update of docker containers
  - Speed-up depends on graphic card (might thus require code optimization for new graphic cards).

West-Life
Structures for Life

Acknowledgements

€€

WeNMR
West-Life
EGI-Engage
INDIGO-
Datacloud
BioExcel CoE

Gydo van Zundert, Jörg Schaarschmidt, Mikael Trellet

Antonio Rosato (CERM), Marco Verlato (INFN), Mario David (LIP)

INDIGO - DataCloud

egi MoBrain

e-infrastructure

Thank you for your attention!