

West-Life: A VRE for Structural Biology

Alexandre Bonvin
Utrecht University



International Symposium on Grids and Clouds 2016

13-18 March 2016, Academia Sinica, Taipei, Taiwan



HORIZON 2020





View all your data in a single, coherent, place.



West-Life is your portal to processing data in the cloud.



Connect your processing together with a single portal.

West-life: your virtual research environment

West-Life is a H2020 Virtual Research Environment project that will provide the application level services specific to use cases in structural biology, covering all experimental techniques (e.g. Xray, cryo-EM, NMR, SAXS), enabling structural biologists to get the benefit of the generic services developed by EUDAT and the EGI.

The project

• 10 Partners:



○ STFC (UK) (lead partner, Martyn Winn Coordinator)



○ Dutch Cancer Institute (NKI) (NL)



○ EMBL (DE)



○ Masaryk University (MU) (CZ)



○ Consejo Superior De Investigaciones Cientificas (CSIC) (ES)



○ Consorzio Interuniversitario Risonanze Magnetiche Di Metallo Proteine (CIRMPP) (IT)



○ INSTRUCT (UK)

Universiteit Utrecht



○ Utrecht University (NL)

LUNA

○ Luna (FR) – (SME)



○ INFN (IT)

Budget: €4 000 0000

Duration: 36 months

Started: 1 Nov 2015

Proposal ID 675858

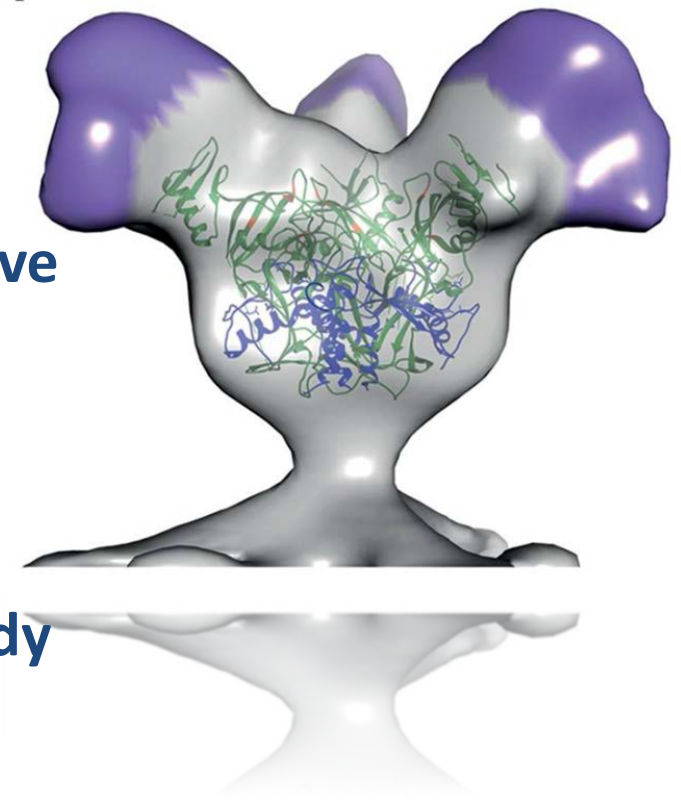


HORIZON 2020



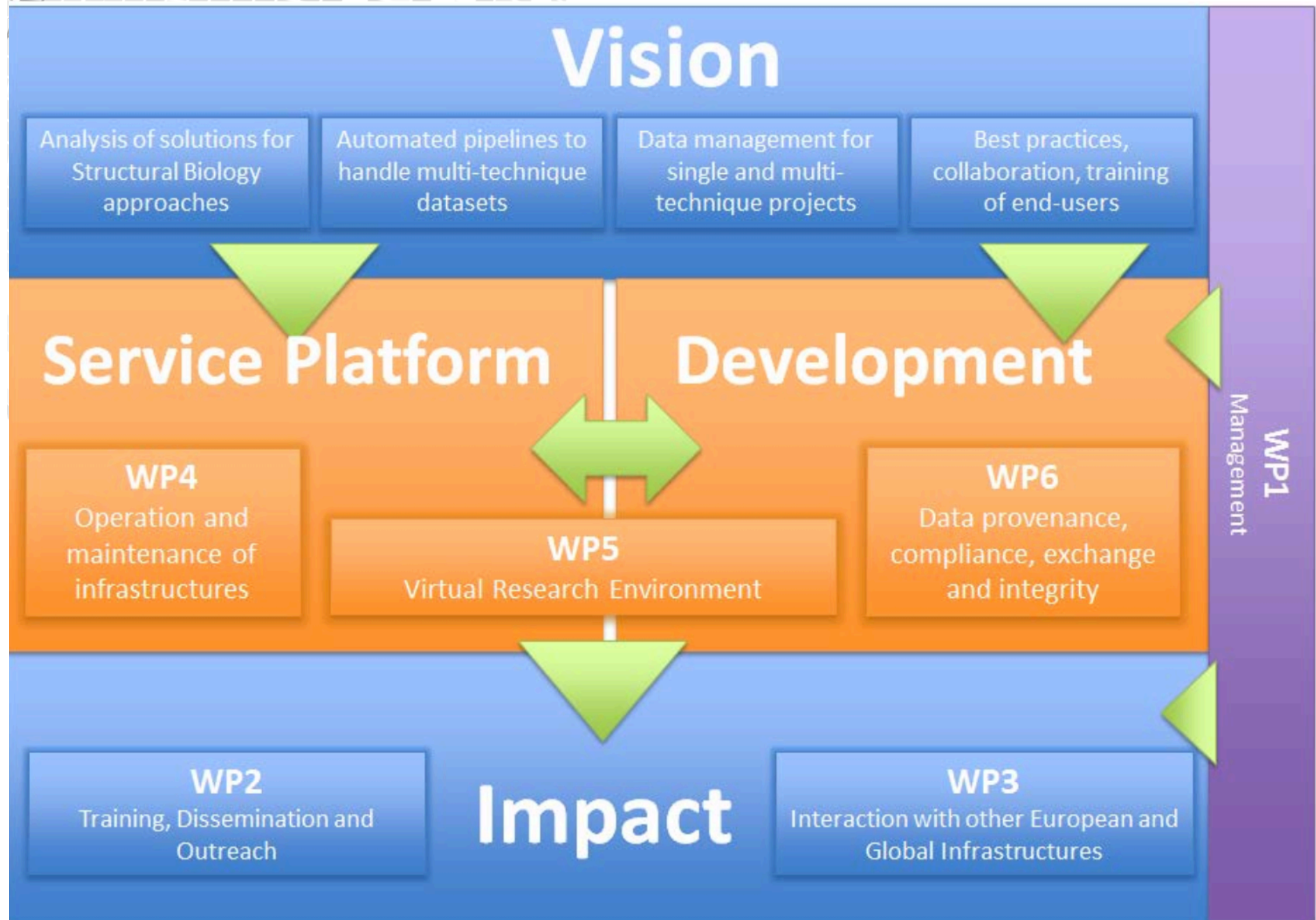
Main Concepts

- **Support for combined techniques:**
 - Multiple facilities visited for one project
 - Data management challenges incl. provenance
 - New algorithms needed for integrative approaches
 - Extends weNMR, uses iCAT, EGI and EUDAT resources
 - Will integrate and connect the already available services



Main objectives

1. Provide analysis solutions for the different Structural Biology approaches
2. Provide automated pipelines to handle multi-technique datasets in an integrative manner
3. Provide integrated data management for single and multi-technique projects, based on existing e-infrastructure
4. Foster best practices, collaboration and training of end users



West-Life: Life Sciences in the Cloud

The West-Life ecosystem

www.west-life.eu

West-Life builds upon the WeNMR achievements





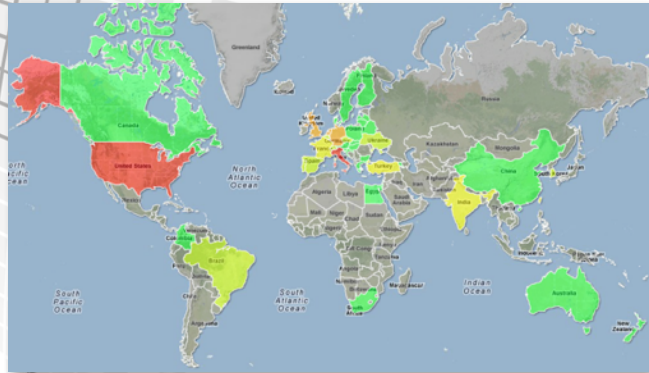
Virtual Research Community



HORIZON 2020



A worldwide e-Infrastructure for NMR and structural biology

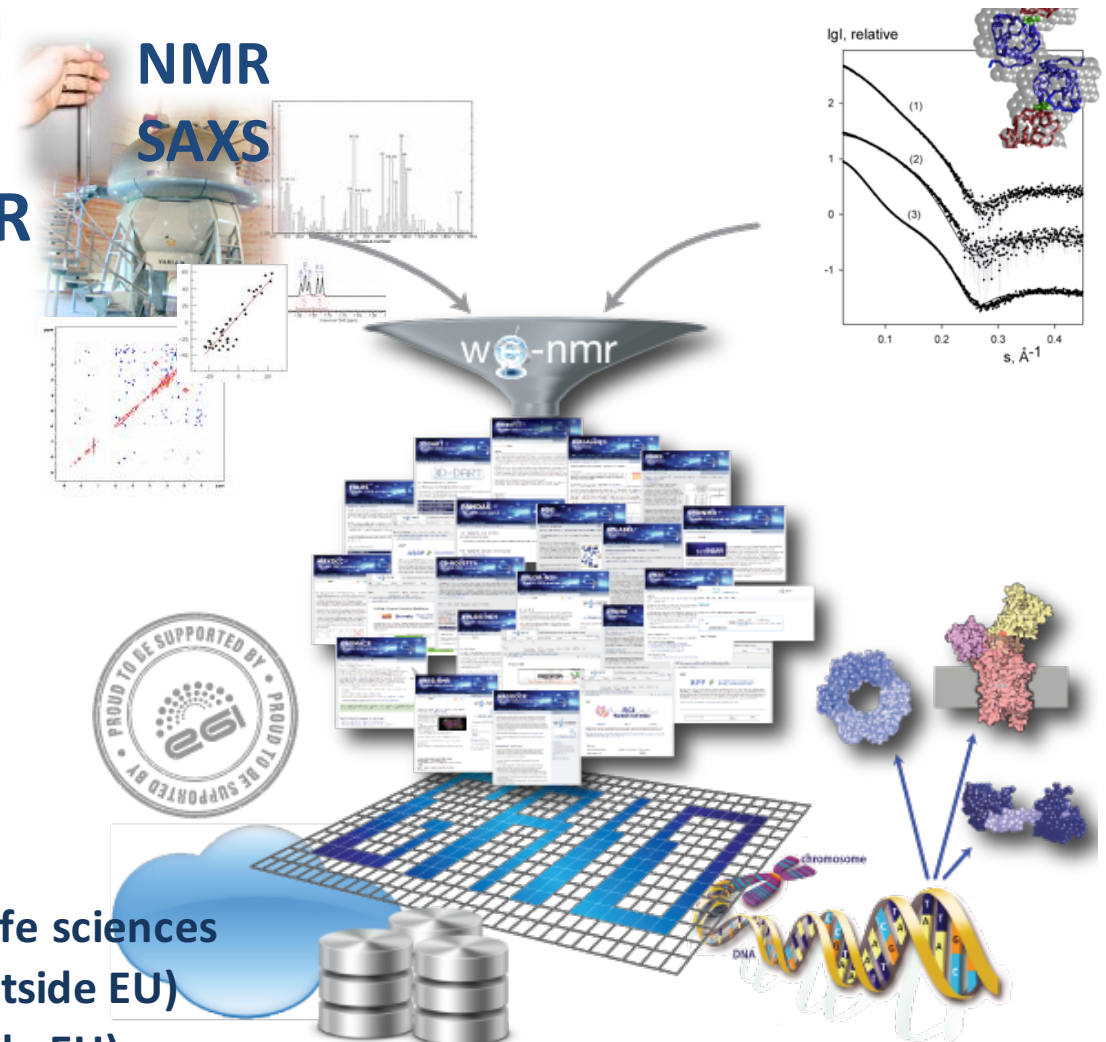


WeNMR VRC (December 2015)

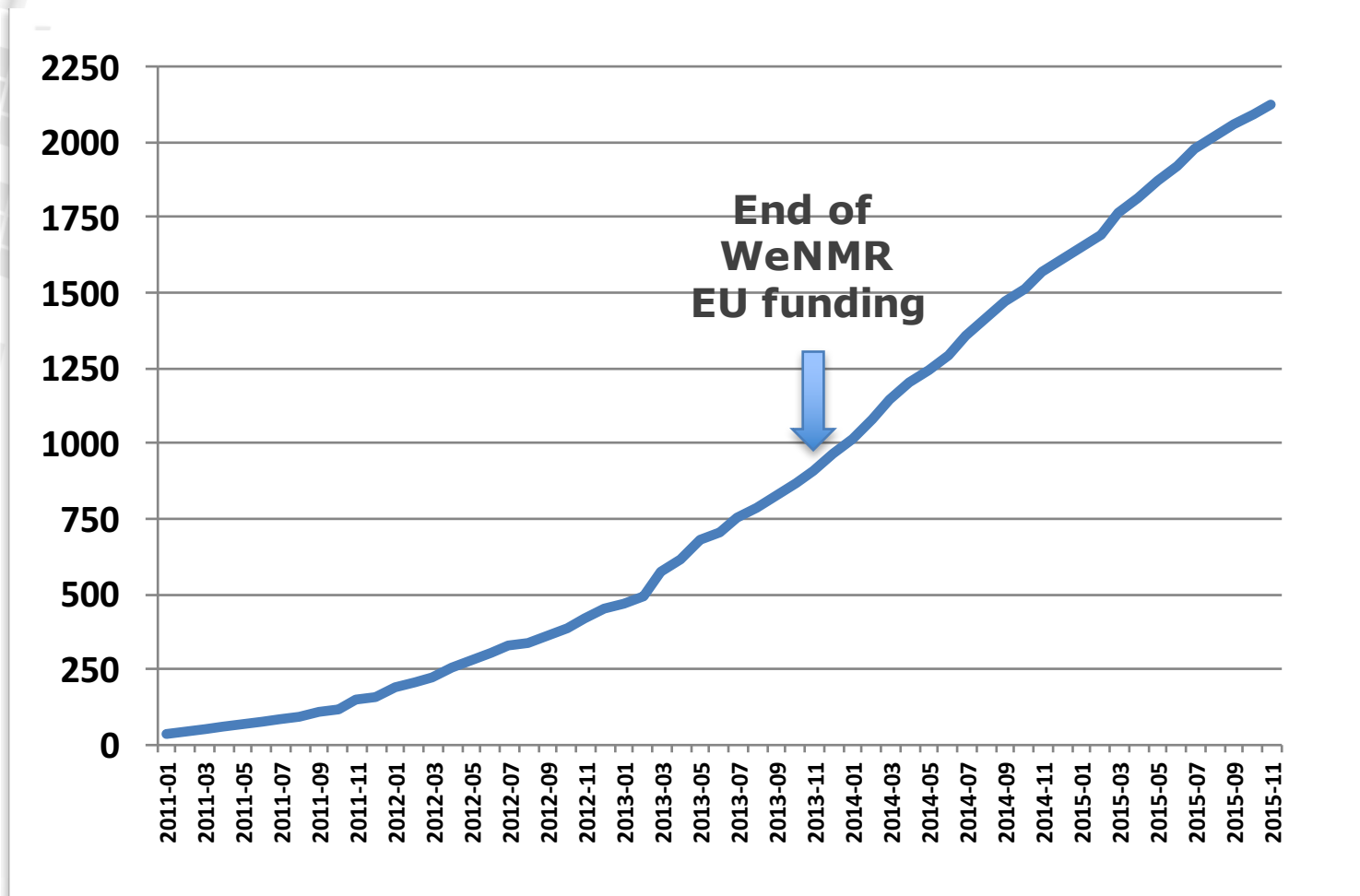
- One of the largest (#users) VO in life sciences
- > 720 VO registered users (36% outside EU)
- > 2250 VRC members (>60% outside EU)
- ~ 41 sites for >142 000 CPU cores via EGI infrastructure
- User-friendly access to Grid via web portals

www.wenmr.eu

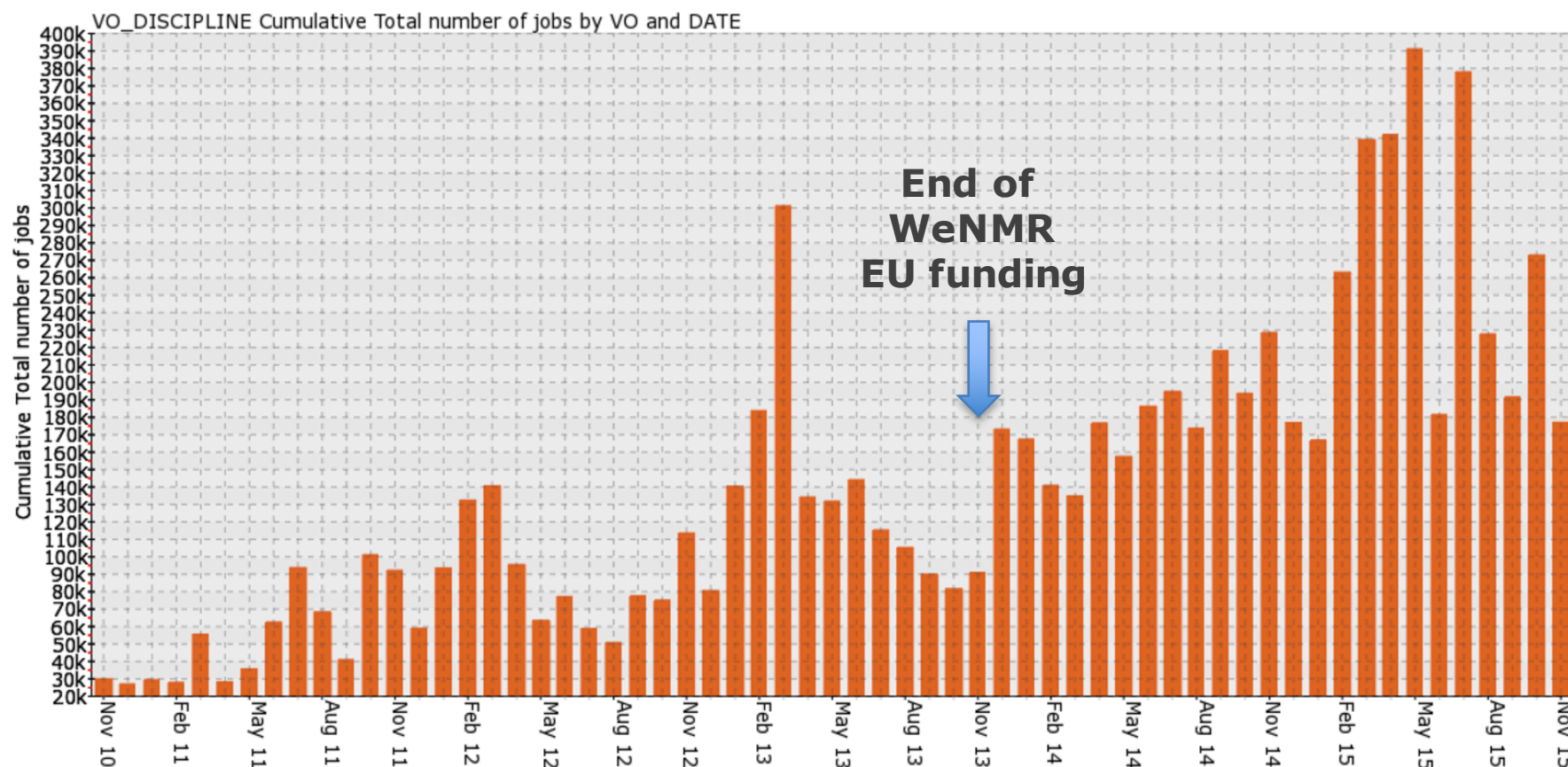
we-nmr



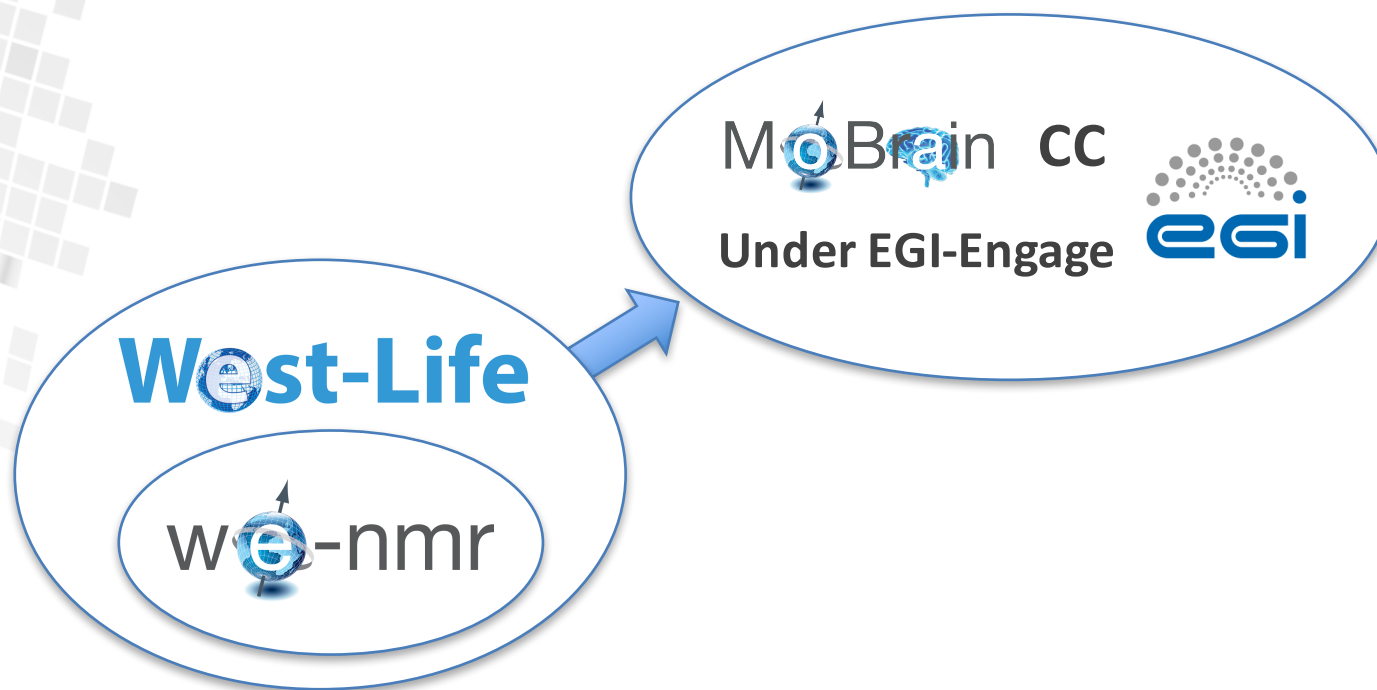
Sustained growth of the WeNMR VRC



Sustained # of jobs



West-Life is connected to EGI

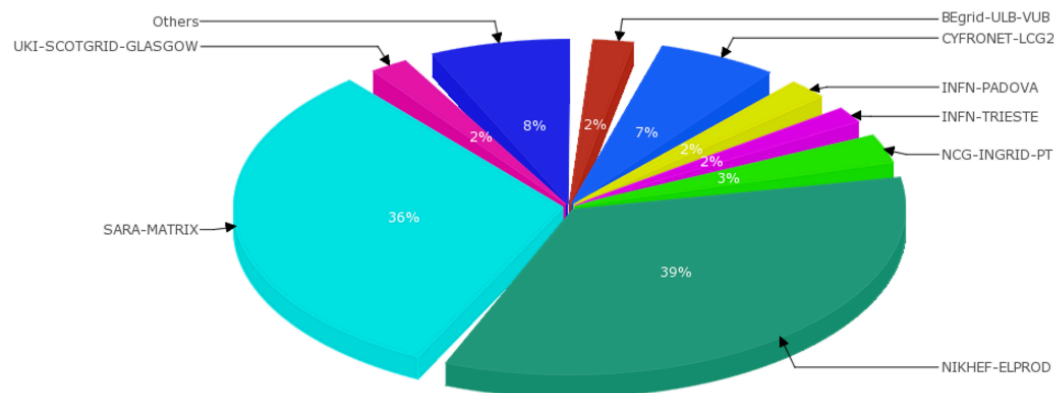


West-Life (and related projects) rely on EGI resources



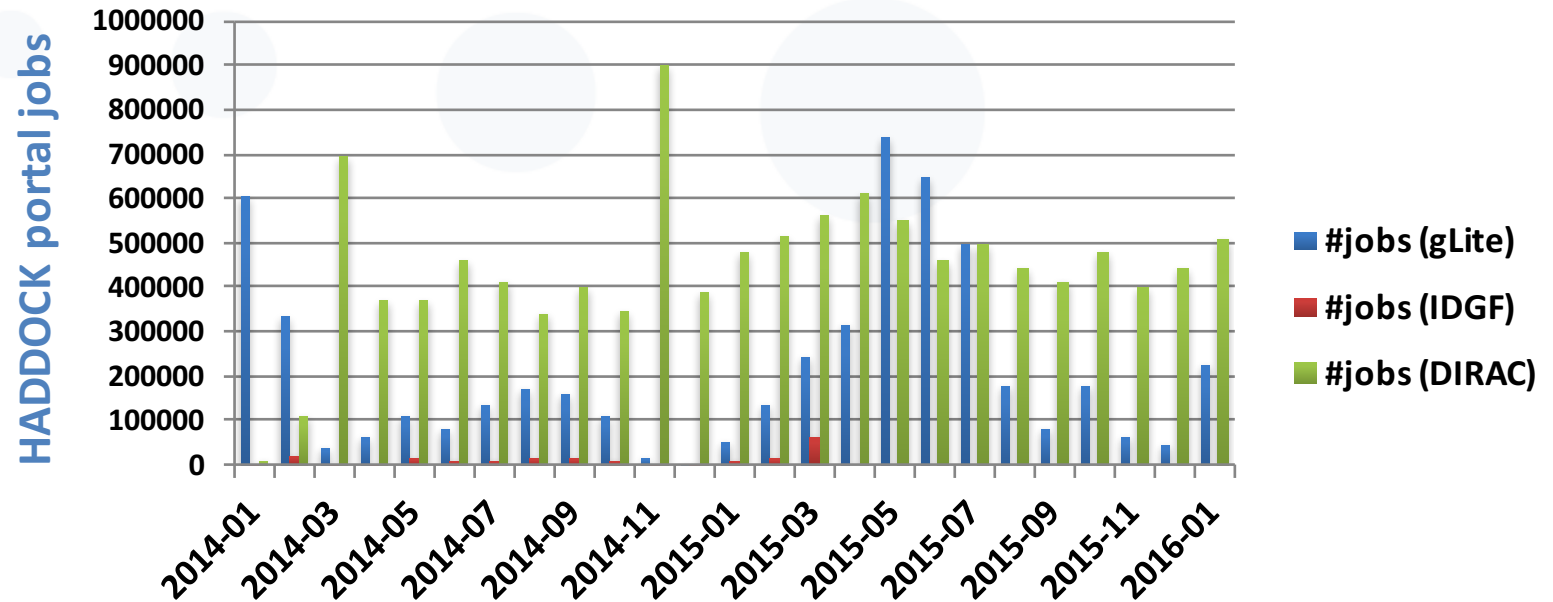
**World-wide: ~ 144'000 CPU
cores from 41 sites (EGI & OSG)
(Stats Jan. 2016)**

WeNMR data from 2014-2015



West-Life (and related projects) rely on EGI resources

- **Mainly grid, but also FedCloud** (e.g. the cryo-EM activities – see Jose Miguel de la Rosa's presentation later)
- **CVMFS** for software deployment
- Currently both **gLite** and **DIRAC4EGI** submission mechanisms



New SLA agreement

- Support for MoBrain/West-Life activities
- 75M CPU hours, 50 TB storage from 7 sites
 - INFN-PADOVA (Italy)
 - RAL-LCG2 (UK)
 - TW-NCHC (Taiwan)
 - SURFsara (The Netherlands)
 - NCG-INGRID-PT (Portugal)
 - NIKHEF (The Netherlands)
 - CESNET-MetaCloud (Czech Republic)



- Brings the micro (WeNMR) and macro (N4U) worlds together into one competence center under EGI Engage:



- With activities toward:
 - Integrating the communities
 - Making best use of cloud resources
 - Bringing data to the cloud (cryo-EM)
 - Exploiting GPGPU resources
- While maintaining the quality of our current services!

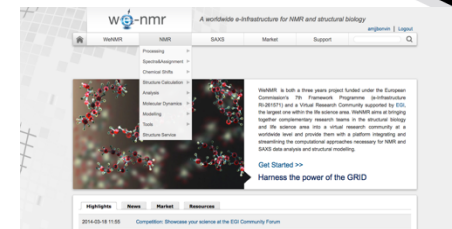
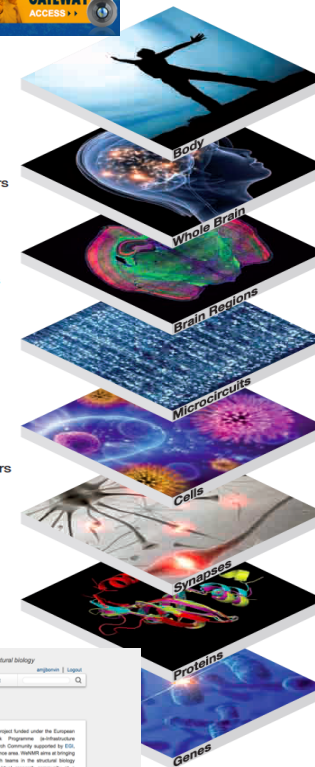


Meters
(10⁰)

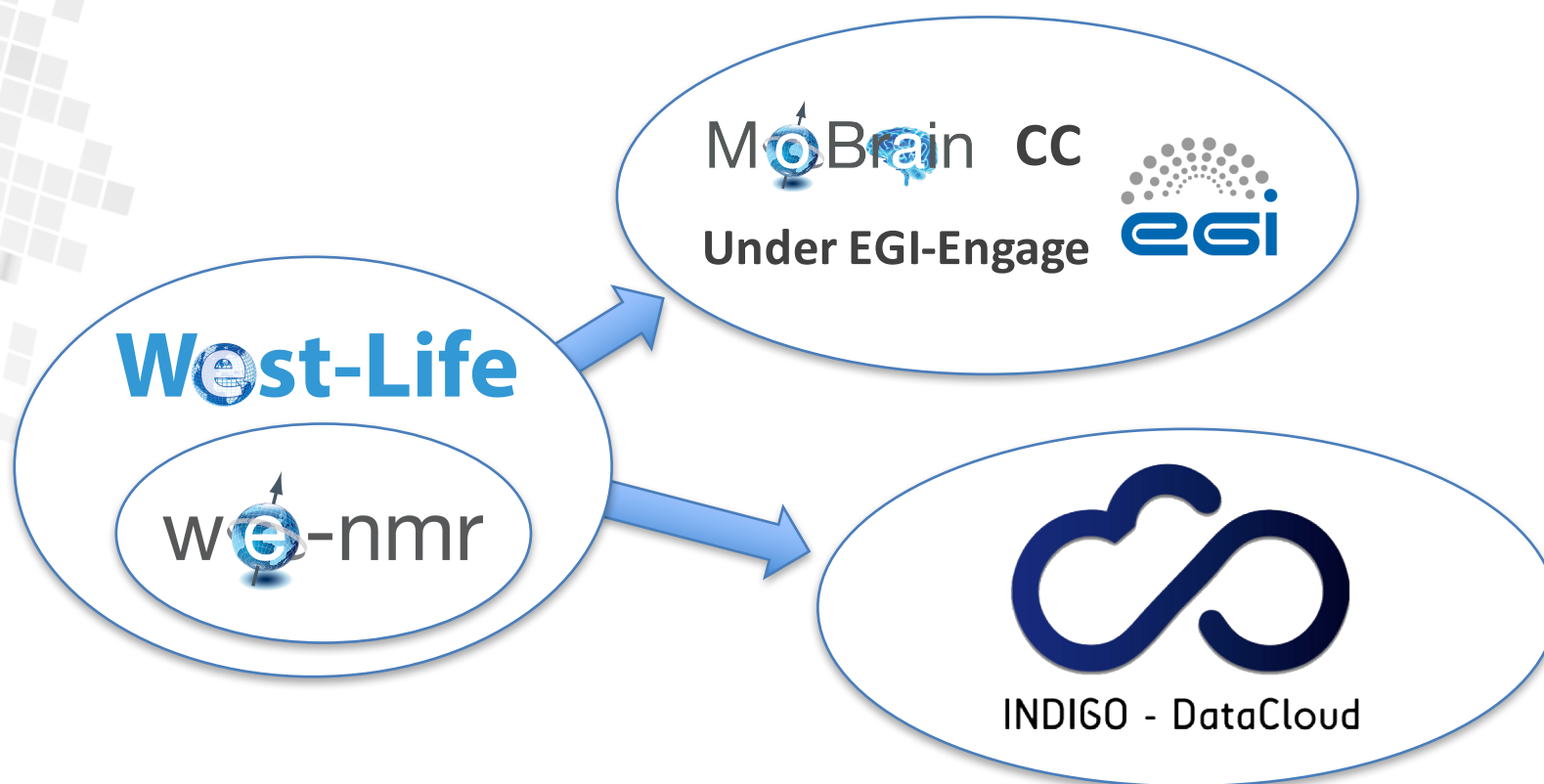
Centimeters
(10⁻²)

Millimeters
(10⁻³)

Micrometers
(10⁻⁶)



West-Life is connected to INDIGO-DataCloud





INDIGO - DataCloud

RIA-653549

Development of a software platform centered around two of the pillars of the EINFRA-1-2014 Horizon 2020 call:

- **Large scale virtualization of data/compute center resources.**
- **Development and adoption of a standards-based computing platform (with an open software stack).**



**West-Life partners
involved with use cases
and requirements**

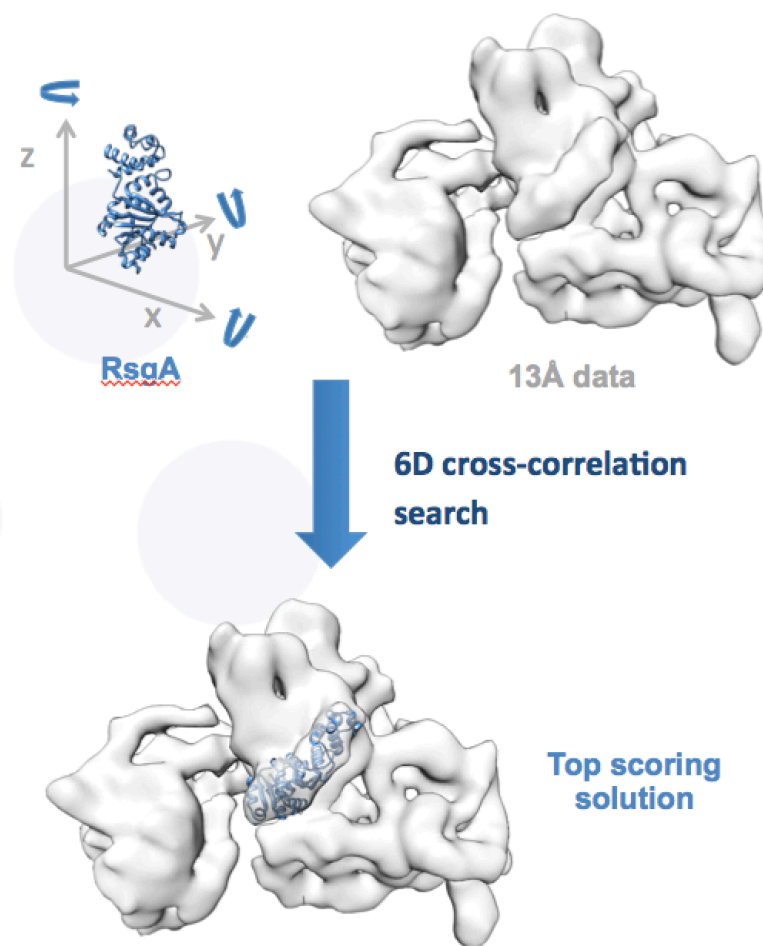


INDIGO-DataCloud is co-funded by the
Horizon 2020 Framework Programme

Exploring GPGPU resources: PowerFit

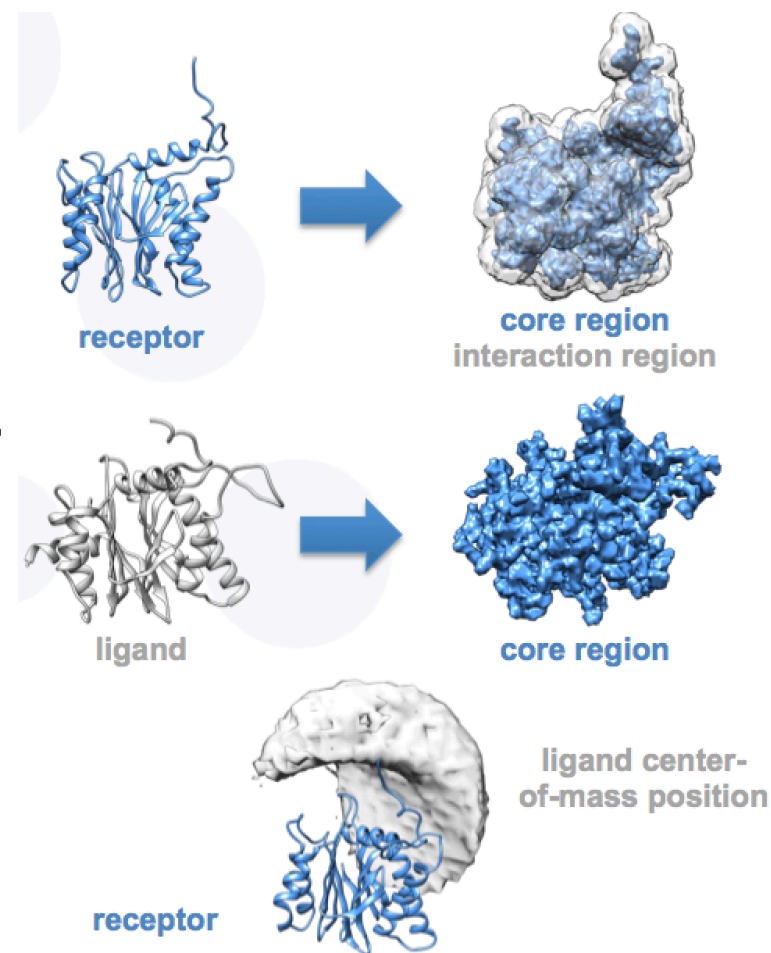
- Python package to automatically fit high-resolution biomolecular structures into cryo-EM densities
- Simple command-line program, able to run using single/multiple CPUs or GPU

van Zundert and Bonvin. *AIMS Biophysics* 2, 73-87 (2015)
www.github.com/haddock/powertfit



Exploring GPGPU resources: DisVis

- Python package to Python package to visualize and quantify the accessible interaction space of distance restrained binary biomolecular complexes.
- Simple command-line program, able to run using single/multiple CPUs or GPU



van Zundert and Bonvin. *Bioinformatics*. 31, 3222-3224 (2015)
www.github.com/haddock/diavis

Performance of CPU vs GPU

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

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

- CPU: AMD Opteron 6344 using FFTW3
- GPU: NVIDIA GeForce GTX 680 using clFFT

Requirements for PowerFit and DisVis

- **Basic:**
 - Python2.7
 - NumPy 1.8+
 - SciPy
 - GCC (or another C-compiler)
- **Optional for faster CPU version:**
 - FFTW3
 - pyFFTW
- **Optional for GPU version:**
 - OpenCL1.1+
 - pyopencl
 - cIFFT
 - gpyfft

Solution for grid and cloud computing:

Docker containers built with proper libraries and opencl support:

Base dockerfile with opencl:

<https://github.com/indigo-dc/docker-opencl>

Dockerfile for DisVis:

<https://github.com/indigo-dc/docker-disvis>

Accessing GPGPU resources via grid

- Test bed at CIRMMP in Florence
- New jdl requirements for GPUs:

```
executable = "disvis.sh";
inputSandbox = { ... };
...
GPUNumber=2;
```

- Exec script with docker command to access the GPU:

```
docker run --device=/dev/nvidia0:/dev/nvidia0 \
  --device=/dev/nvidia1:/dev/nvidia1 \
  --device=/dev/nvidiactl:/dev/nvidiactl \
  -v $WDIR:/home opencl_disvis /bin/sh \
  -c 'export LD_LIBRARY_PATH=/usr/local/lib64; \
  disvis /home/O14250.pdb /home/Q9UT97.pdb /home/restrain'
```

Baremetal vs docker vs cloud

ID	Type	GPU	#Cores CPU type	Mem (GB)
B-K20	Baremetal	Tesla K20	24 HT (12 real) Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz	32
B-K40	Baremetal	Tesla K40	48 HT (24 real) Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz	512
D-K20	Docker on K20	Tesla K20	24 Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz	32
K-K40	KVM on K40	Tesla K40	24 Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz	32

Courtesy of Mario David
INDIGO

Case	Machine	TimeGPU (sec)	TimeCPU 1 core	CPU1/GPU
B-K40	Baremetal	674	7928	11.8
K-K40	KVM	671	7996	11.9
B-K20	Baremetal	830	11839	14.3
D- K20	Docker	837	11926	14.3

No loss of performance

Virtualising portals using EC3

EC3: Elastic Cloud Computing Cluster

FEATURES

LEARN MORE

DEPLOY!

CONTACT

WHERE DO YOU WANT TO DEPLOY THE CLUSTER?

You will need to provide valid credentials for the Cloud provider. Not sure if this is safe? [Check the docs.](#)

Wanted to deploy a hybrid cluster? You can do it with the [CLI](#).



Amazon Web Services

Public Cloud provider



OpenNebula

On-premises Cloud provider



OpenStack

On-premises Cloud provider



EGI FedCloud

European Federated Cloud

[\(See a case study here\)](#)







Back to West-Life

VRE tasks

- **Development and integration of new service portals**
 - Interface X-ray and cryo-EM portals to the most suited e-Infrastructure solution
 - Select most suited submission mechanisms (e.g. DIRAC)
 - Incorporate newly identified portals into the VRE
 - User-friendly interfaces, with a VRE- integrated AAI
- **Custom VMs for pre-defined scenarios**

Newly identified tools to be implemented


Scipion Web Tools

Single particle analysis tools

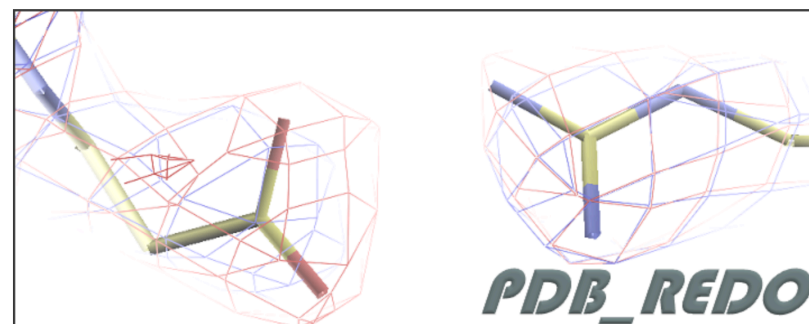
Create your initial volume [My first map](#)

Align your movies [My movie alignment](#)

Analyze your maps [My resolution map](#)


Collaborative Computational Project No. 4
 Software for Macromolecular X-Ray Crystallography

A databank of updated and optimised X-ray structure models



Enter a PDB code (4 characters):

Welcome to CCP4 online

[Login](#)

Other Options - [Register](#), [Forgotten Password](#), [Change Password](#)

Runnable programs

The following programs and pipelines are available:

Balbes

An automated Molecular Replacement (MR) pipeline - Balbes integrates into one system all the components necessary for solving a crystal structure by Molecular Replacement

MrBUMP

An automated Molecular Replacement (MR) pipeline - Given a target sequence and experimental structure factors, it will search for homologous structures, create a set of suitable search models from the template structures, do molecular replacement, and test the solutions with some rounds of restrained refinement.

Zanuda

Space group and crystallographic origin validation

jsPISA

Calculation and analysis of macromolecular surfaces and interfaces

AMPLE

Automated ab initio search model generation for molecular replacement.

SHELX

Automated SHELXC/D/E structure solution pipeline for fast routine experimental phasing. Accepts data in XDS, Scalepack, SHELX hkl or mtz formats and outputs phases and a poly-Ala trace. If a protein sequence is provided, BUCCANEER and REFMAC complete the structure. **BETA**

CRANK2

Automated structure solution pipeline for experimental phasing using maximum likelihood methods. **BETA**



ARP/wARP web service

7.6

[DISCLAIMER](#)
[ARP/wARP LICENCE](#)
[CCP4 LICENCE](#)
[CITATIONS](#)

ARP/wARP web service: step 1

Your email address:

Run ARPwARP starting from:

MTZ file: no file selected

OR use example data: (2.0Å Leishmanolysin) ☐

Proceed to step two



HORIZON 2020



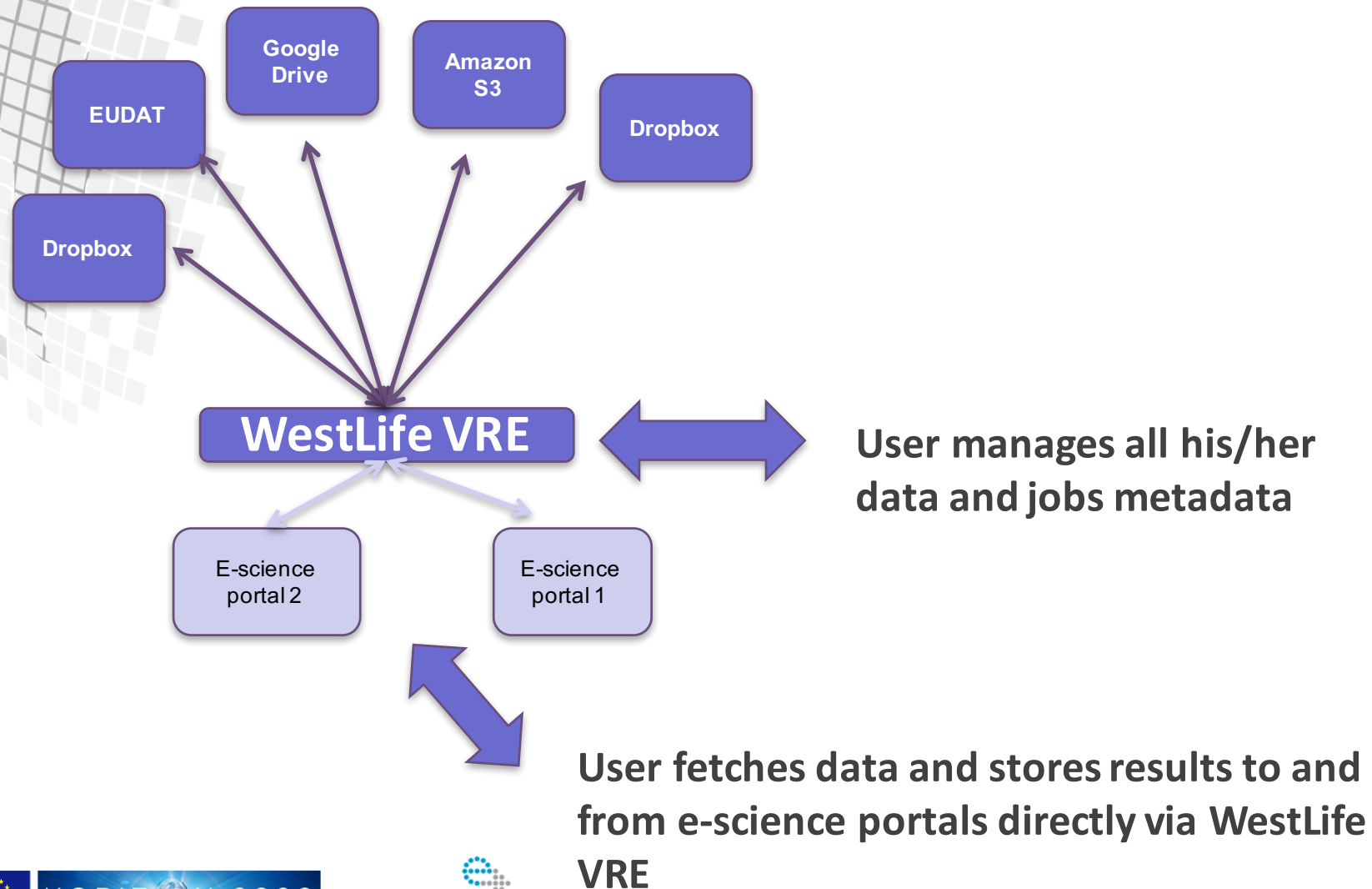
Baseline of services

Portal	Method	Service	grid/ cloud- enabled	Users	Users 2015	No. of user submissions 2015	No. of grid/cloud jobs 2015
Scipion	Cryo-EM	3D electron microscopy online processing workflows	-/+	-	-	100	-
GROMACS	Modelling	Molecular dynamics simulations	+/-	112	26	173	588
ViCi	Modelling	In silico ligand-based drug design	-/-	292	103	93	-
HADDOCK	NMR/modelling	Docking of biomolecular complexes	+/-	6699	1450	25k	7,5M
AMPS-NMR	NMR	Molecular dynamics simulations with AMBER	+/-	300	50	-	8k
CS-Rosetta3	NMR	Structure prediction with chemical shifts from NMR	+/-	51	16	67	189k
FANTEN	NMR	Determination of anisotropy tensors (NMR)	-/-	-	-	-	-
UNIO	NMR	Structure calculations including NOE assignment from NMR data	+/-	59	31	62	2285
XPLOR-NIH	NMR	Protein solution structure determination through structural restraints, simulated annealing calculations and energy minimization	+/-	100	10	-	80k
ARP/wARP	X-ray	Crystallographic Macromolecular Model Building	-/-	4088	760	3,2k	-
Auto-Rickshaw	X-ray	Automated crystal structure determination platform	-/-	2319	458	3,5k	-
CPP4 - Ample*	X-ray	Automated search model generation and molecular replacement (MR)	-/-	20	-	-	-
CPP4 - Balbes	X-ray	Automated MR pipeline	-/-	1155	728	3,3k	-
CPP4 - Crank2	X-ray	Structure solution pipeline for experimental phasing	-/-	10	-	-	-
CPP4 - MrBUMP	X-ray	Macromolecular structure solution by MR	-/-	580	473	1,3k	-
CPP4 - Shelx*	X-ray	SHELXC/D/E structure solution	-/-	19	-	-	-
CPP4 - Zanuda	X-ray	Space group and crystallographic origin validation	-/-	264	189	371	-
PDB REDO	X-ray	Optimization of crystallographic structure models	-/-	700	500	3k	-
CCD	X-ray/Molecular Biology	Design of constructs for protein crystallography	-/-	-	-	~3k	-

Toward a new VRE portal

- The WestLife VRE should provide a web portal where structural biologists can:
 - Have an aggregated view of all their data across repositories (EUDAT, Dropbox etc)
 - Upload data directly into e-science portals
 - Download results directly from e-science portals into their repositories (no need to download/upload to and from your laptop)
 - Store their jobs metadata

Toward a new VRE portal



Toward a new VRE portal



West-Life Conclusions

- **A lively and growing community in life sciences**
 - building on the achievements of WeNMR
 - integrating key software developers in X-ray and cryo-EM
 - making use of various e-Science and e-Infrastructure solution in collaboration with relevant e-Infra projects
- **to best serve the INSTRUCT and worldwide structural biology communities**



West-Life Life Sciences in the Cloud

Teams of scientists and engineers around the world use West-Life to push the boundaries in life sciences.

**Thank you for your
attention!**