WISDOM, a grid enabled virtual screening initiative

Yannick Legré
LPC Clermont-Ferrand, France (CNRS/IN2P3)

On behalf of the WISDOM Collaboration

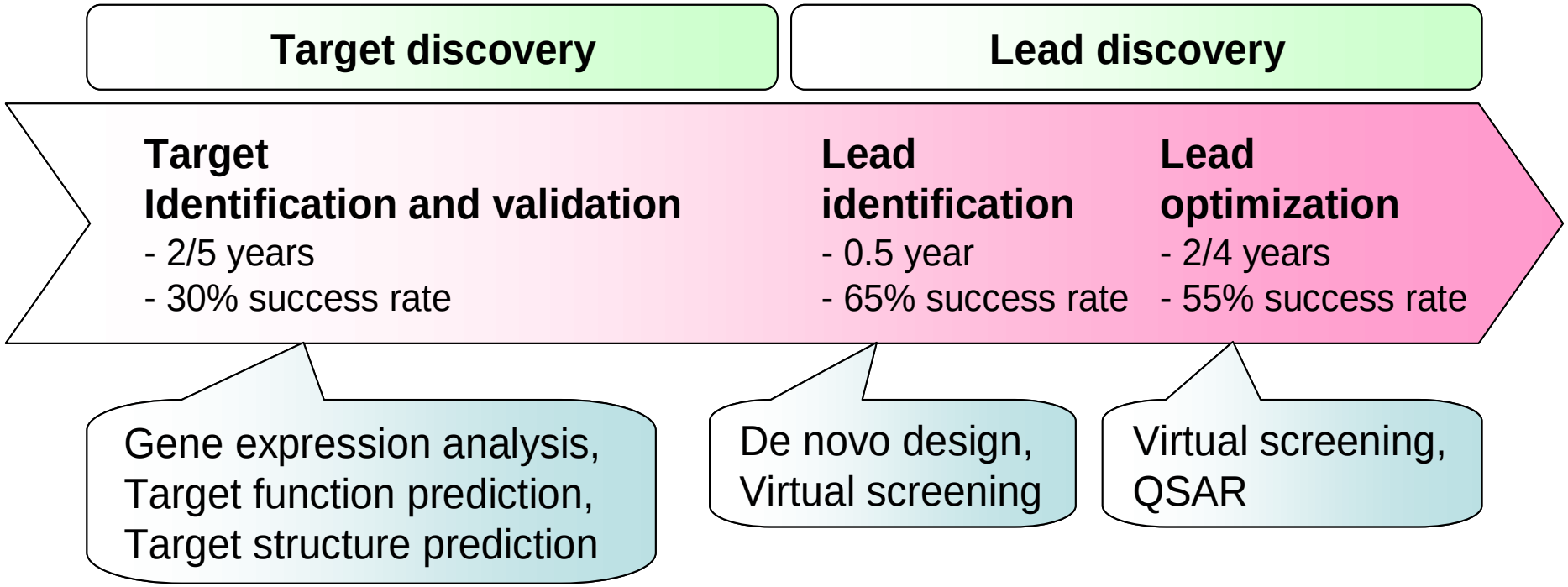
Visit us on EGEE booth

www.eu-egee.org

Information Society and Media

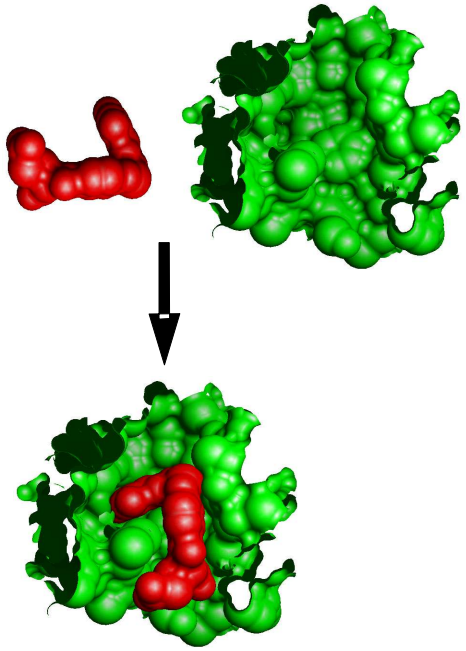
Searching for new drugs

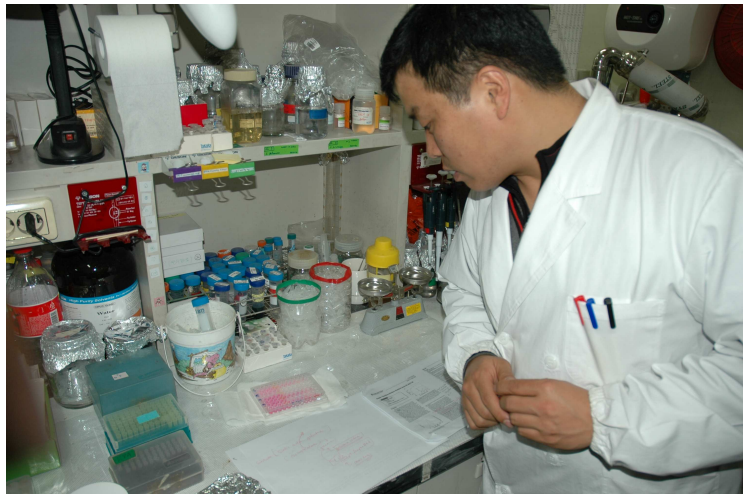
- **Drug development is a long (10-12 years) and expensive (~800 MDollars) process**
- **In silico drug discovery opens new perspectives to speed it up and reduce its cost**



Simplified principle of screening

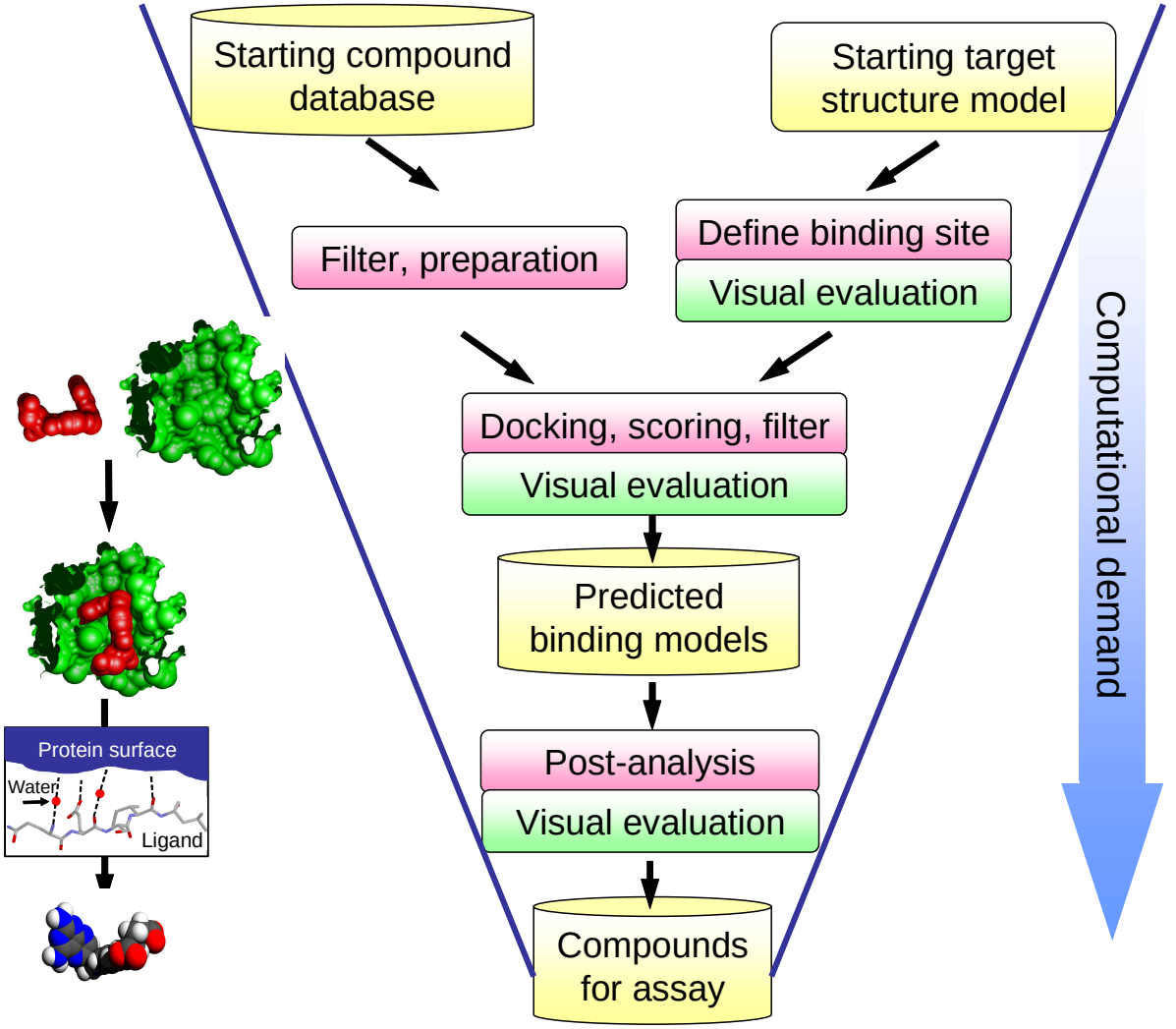
- **Biologists isolate a protein, the target, which plays an important role in the life cycle**
  - of the malaria parasite
  - of the H5N1 virus
  - of a cancer cell
  - …
- **The objective is to find other molecules which will block the action of that protein: the hits**
  - Docking of the molecule on the protein active site
- ***in silico* docking vs *in vitro docking***
  - *In silico:* calculation of the binding energy between molecules
  - *In vitro:* measurement of the chemical reaction coefficient





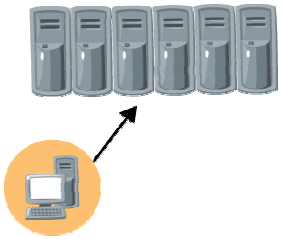
A first step towards in silico drug discovery: virtual screening

- **In silico virtual screening**
  - Starting from millions of compounds, select a handful of compounds for in vitro testing
  - Very computationally intensive but potentially much cheaper than in vitro testing
- **Where to find CPUs to make it time effective ?**

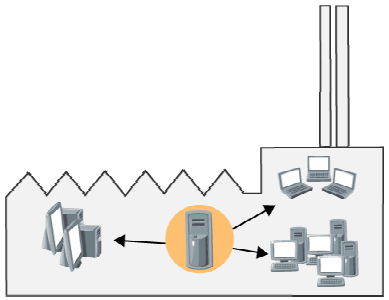


Distributed Computing in a nutshell

Cluster



Enterprise Grids

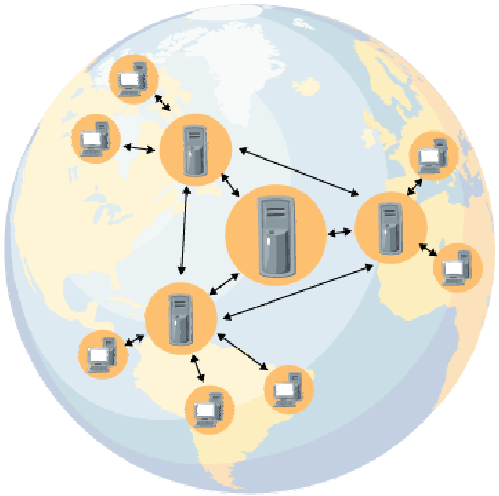


Example:
United Devices

Volunteer Computing

Example:
World Community Grid
Africa@home

‘The Grid’ (in this talk!)



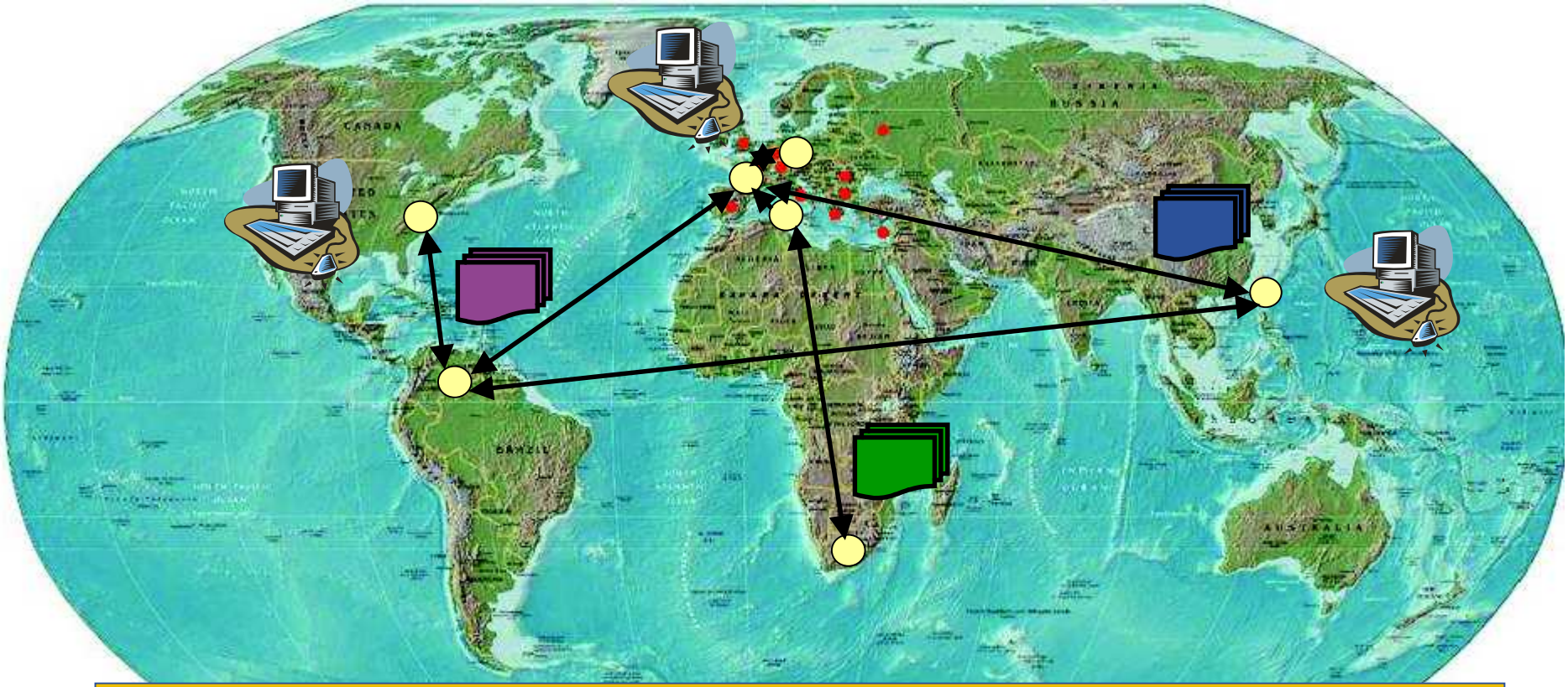
Example:
EGEE

What is the Grid?

- The World Wide Web provides seamless access to information that is stored in many millions of different geographical locations
- In contrast, the Grid is a new computing infrastructure which provides seamless access to computing power, data and other resources distributed over the globe
- The name Grid is chosen by analogy with the electric power grid: plug-in to computing power without worrying where it comes from, like a toaster

The grid added value for international collaboration

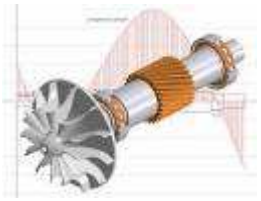
- **Grids offer unprecedented opportunities for sharing information and resources world-wide**

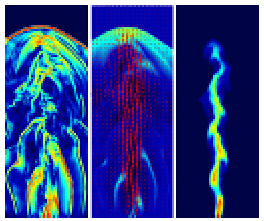


Grids are unique tools for :
-Collecting and sharing information (Epidemiology, Genomics)
-Networking experts
-Mobilizing resources routinely or in emergency (drug discovery)

The EGEE-II project

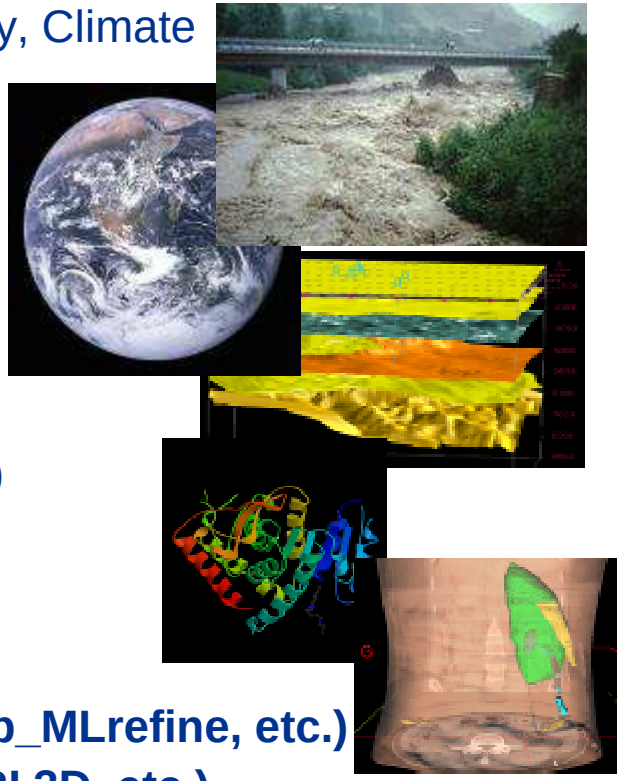
- **EGEE**
  - 1 April 2004 – 31 March 2006
  - 71 partners in 27 countries, federated in regional Grids
- **EGEE-II**
  - 1 April 2006 – 31 March 2008
  - 91 partners in 32 countries
  - 13 Federations
- **Objectives**
  - Large-scale, production-quality infrastructure for e-Science
  - Attracting new resources and users from industry as well as science
  - Maintain and further improve “gLite” Grid middleware





Applications on EGEE

- **More than 25 applications from 9 domains**
  - Astrophysics
    - MAGIC, Planck
  - Computational Chemistry
  - Earth Sciences
    - Earth Observation, Solid Earth Physics, Hydrology, Climate
  - Financial Simulation
    - E-GRID
  - Fusion
  - Geophysics
    - EGEODE
  - High Energy Physics
    - 4 LHC experiments (ALICE, ATLAS, CMS, LHCb)
    - BaBar, CDF, DØ, ZEUS
  - Multimedia
  - **Life Sciences**
    - **Bioinformatics (Drug Discovery, GPS@, Xmipp_MLrefine, etc.)**
    - **Medical imaging (GATE, CDSS, gPTM3D, SiMRI 3D, etc.)**



WISDOM

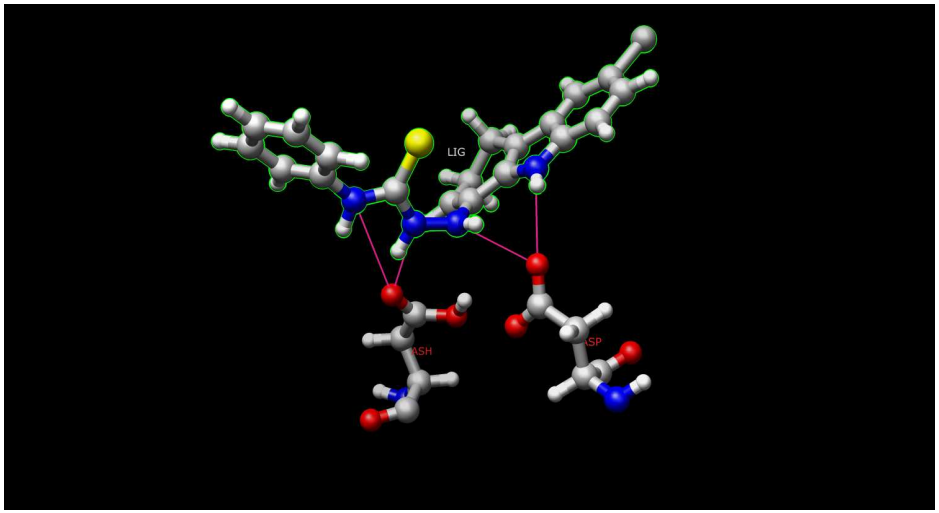
- **WISDOM stands for World-wide In Silico Docking On Malaria**

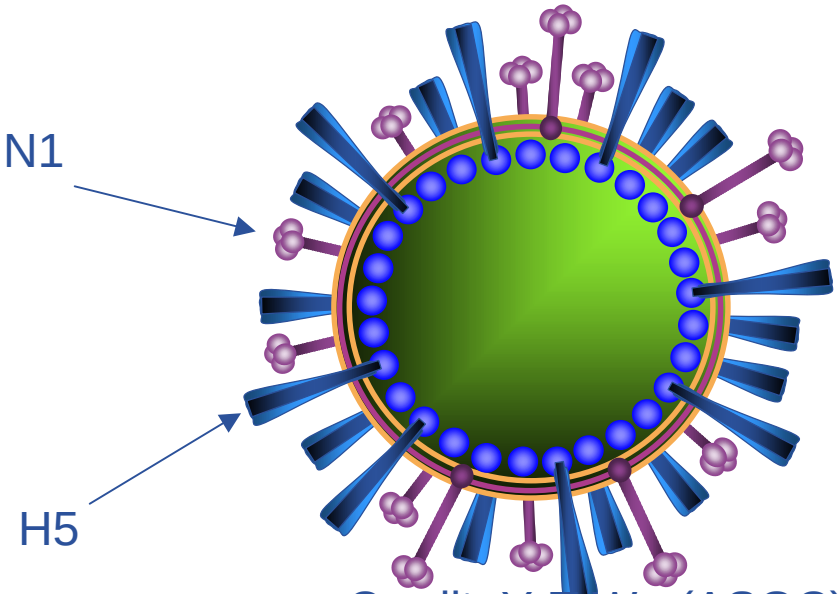
- **Goal: find new drugs for neglected and emerging diseases**
  - Neglected diseases lack R&D
  - Emerging diseases require very rapid response time

- **Method: grid-enabled virtual docking**
  - Cheaper than in vitro tests
  - Faster than in vitro tests

Biological objectives

- **Malaria: Find active molecules**
  - on a known mutated protein (DHFR)
  - on new targets:
    - Plasmepsins
    - GST
    - Tubulin
- **Avian Flu**
  - Study the impact of point mutations of the N1 enzyme
    - Tamiflu active on N1
  - Find new molecules active on N1





Credit: Y-T Wu (ASGC)

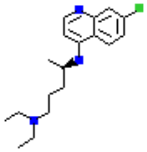
Grid-enabled virtual docking

Millions of potential drugs to test against interesting proteins!

~~High Throughput Screening ~ 10$/compound several hours~~

Too costly for neglected disease!

Compounds:

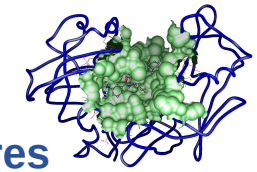


ZINC: 4.3M

Chembridge: 500 000

Molecular docking (FlexX, Autodock)
~1 to 15 minutes

Targets:

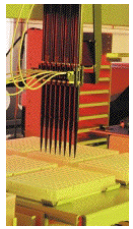


PDB: 3D structures

Data challenge on EGEE
~ 2 to 30 days on ~5000 computers

Cheap and fast!

Selection of the best hits



Hits screening using assays performed on living cells

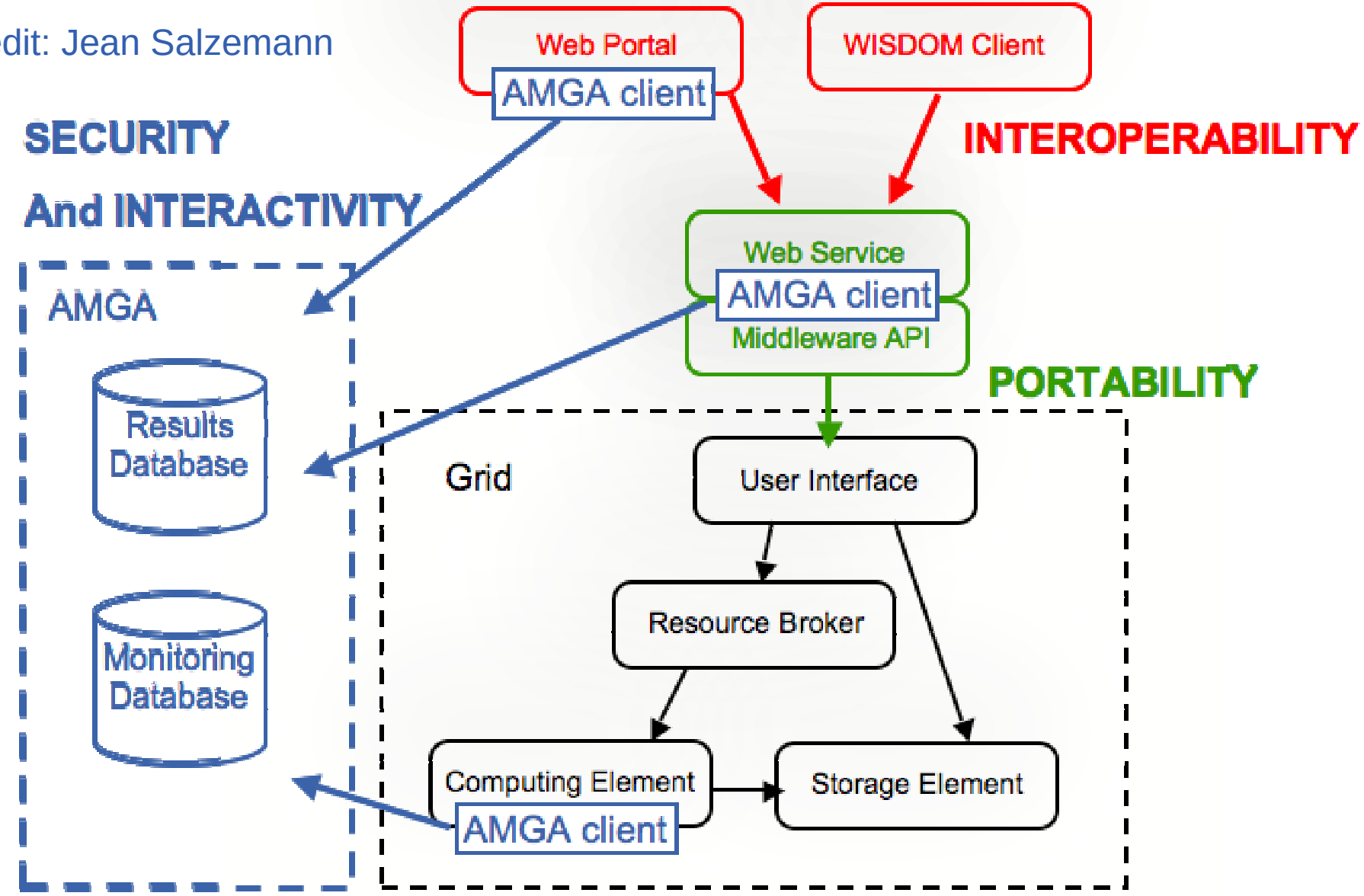
Leads

Clinical testing

Drug

Production Environment

Credit: Jean Salzemann



Statistics of deployment

- **First Data Challenge: July 1st - August 15th 2005**
  - Target: malaria
  - 80 CPU years
  - 1 TB of data produced
  - 1700 CPUs used in parallel
  - 1st large scale docking deployment world-wide on a e-infrastructure

- **Second Data Challenge: April 15th - June 30th 2006**
  - Target: avian flu
  - 100 CPU years
  - 800 GB of data produced
  - 1700 CPUs used in parallel
  - Collaboration initiated on March 1st: deployment preparation achieved in 45 days

- **Third Data Challenge: October 1st - 15th December 2006**
  - Target: malaria
  - 400 CPU years
  - 1,6 TB of data produced
  - Up to 5000 CPUs used in parallel
  - Very high docking throughput: > 100.000 compounds per hour

Post docking analysis on Avian Flu

- **Selection of the most promising molecules in a 2-step process**
  - 1st step: rejection of 85% based on docking score.
  - 2nd step: re-ranking of the remaining 15% and selection of the best 5%

- **6 known active inhibitors included in the analyse to validate the process**
  - 5 out of 6 kept in the 2250 selected compounds

Global effectiveness:

$(Hits_{sampled}/N_{sampled})/(Hits_{total}/N_{total})$

Pearlman & Charifson, JMC, 2001

Pre-sceening (AUTODOCK) over collection and sample first 15%

$EF^1 = (5/6)/15\% = 5.5$

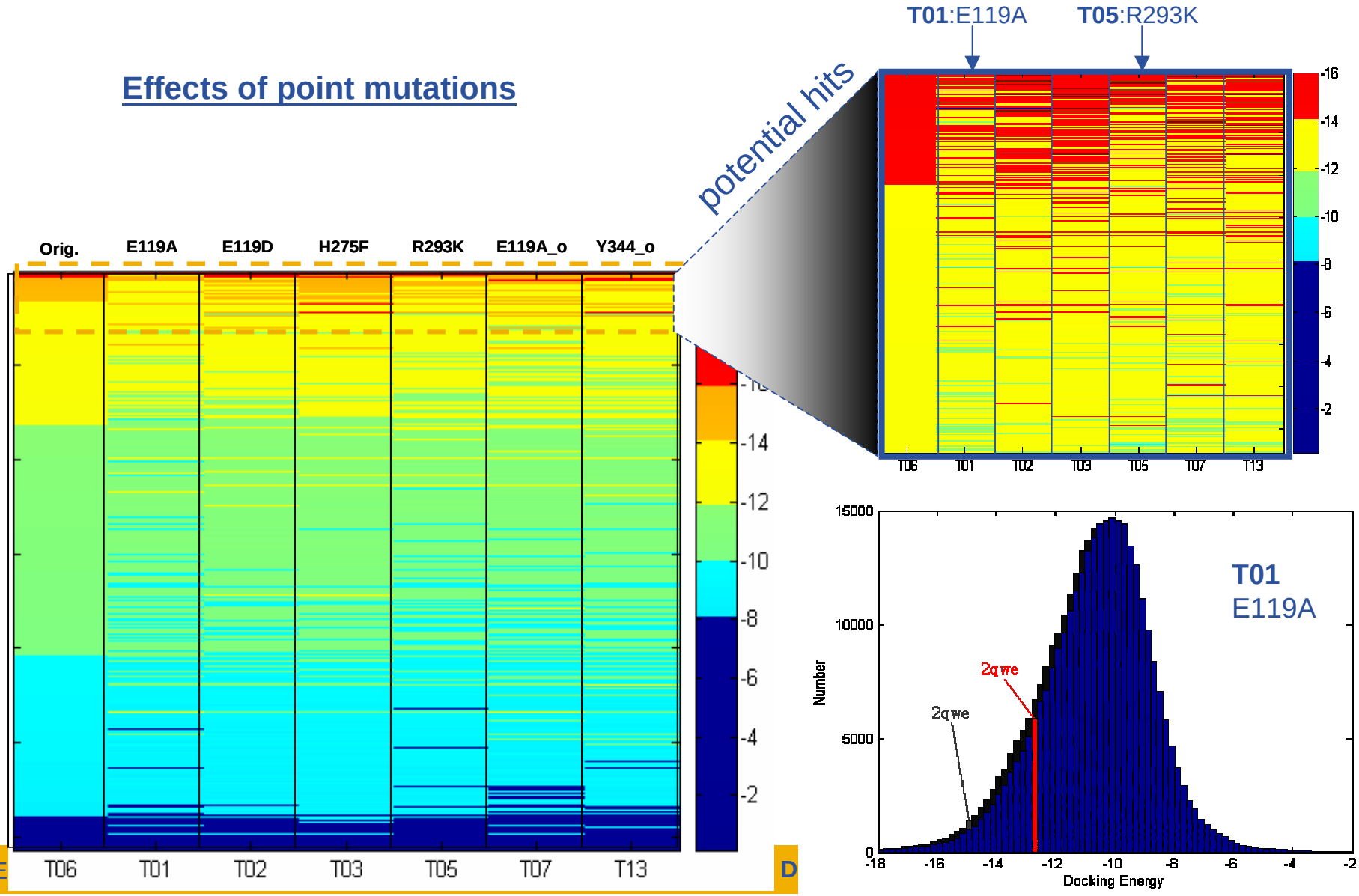
Re-ranking (SDDB) first 15% and sample first 5%

$EF^2 = (5/6)/(5\%*15\%) = 111$

Credit: Y-T Wu (ASGC)

Significant impact of mutations on the activity of molecules

Effects of point mutations

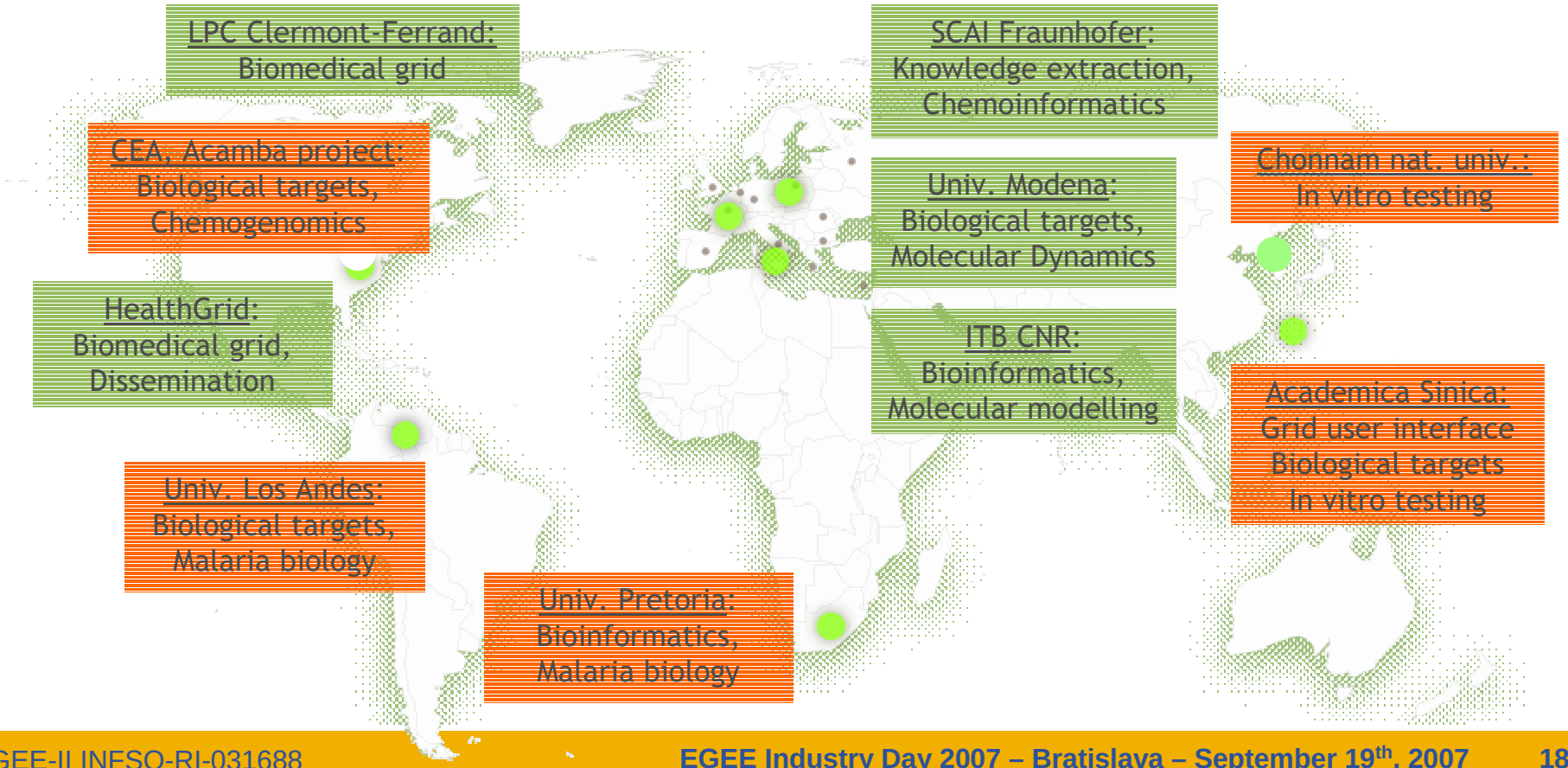


Status of in vitro tests

- **Avian Flu**
  - Initial number of compounds: 300.000
  - 123 compounds bought and tested out of the 2250 selected
    - 7 out of 123, approximately 6%, are active
  - Usual average success rate for in vitro tests: 0,1%
  - Factor 60 increase to be confirmed on more compounds
  - Tests under way at Chonnam National University (ROK)
- **Malaria**
  - Initial number of compounds: 500.000 (WISDOM-I)
  - Selection of 30 molecules in 2 steps
    - 1000 molecules selected on docking score
    - Selection of 30 molecules through molecular dynamics
  - Tests under way at Chonnam National University (ROK)
  - First results are very encouraging

Grid added value for in silico drug discovery

- The grid provides the centuries of CPU cycles required for virtual screening
- The grid provides the reliable and secure data management services to store and replicate the biochemical inputs and outputs
- The grid offers a collaborative environment for the sharing of data in the research community on Avian Flu and Malaria



Industrial perspective

- **A secure and reliable production environment was developed for Wisdom**
  - Up to 100.000 docked compounds per hour (WISDOM-II)
  - Distributed Secured Data Management
- **This environment has been used for other life sciences applications (e.g. PDB refinement within the EMBRACE project - CMBI)**
- **We are ready to address industrial requirements!**

Visit us on EGEE booth

Credits

Academia Sinica
BioSolveIT
CNR-ITB
CNRS
CEA
Chonnam National University
HealthGrid
IN2P3
LPC
SCAI Fraunhofer
Università di Modena e Reggio Emilia
Université Blaise Pascal
University of Pretoria
University of Los Andes

Auvergrid
Accamba
BioInfoGRID
EELA
EGEE
EMBRACE
EUChinaGRID
EUMedGRID
SHARE
TWGrid

Conseil Regional d'Auvergne
European Union

With the sponsorship of Università di Modena e Reggio Emilia