THE UNIVERSITY of TENNESSEE KNOXVILLE

Center for Molecular Biophysics

OAK RIDGE National Laboratory

Drug Discovery at the Petascale & Big Data

Jerome Baudry

Jeremy Smith

Sally Ellingson

UTK Biochemistry & Cellular and Molecular Biology

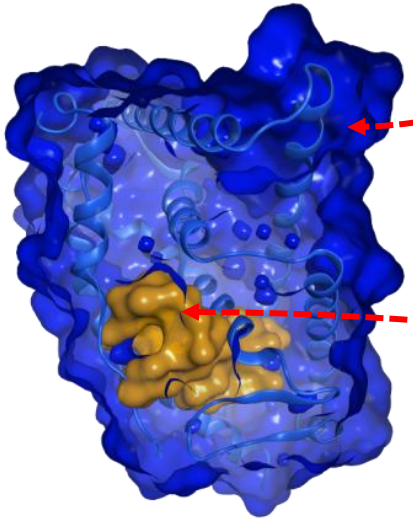
UT/ORNL Center for Molecular Biophysics

Genome Science and Technology

From Big Compute to Big Data

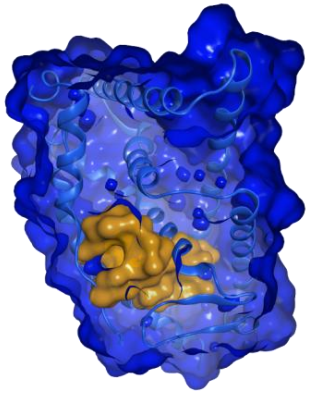
How do drugs do their 'magic'?

Drug discovery/development paradigm (~2/3 of drugs) =
a small organic molecule binding a protein receptor

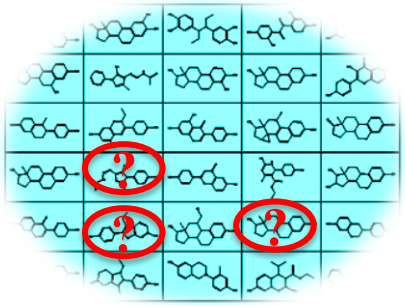


Drug discovery =
i) linking a disease with a protein and ii) finding small molecule effectors

Computational drug discovery & (virtual) screening



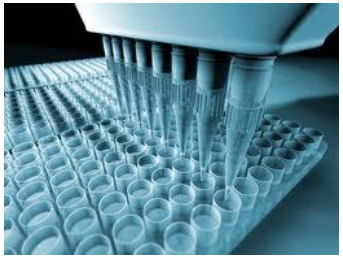
+



protein model

chemicals databases

Identify molecules that are likely to bind a given protein (*in vitro* binding assay)



screening 1 drug candidate: $10 to $20

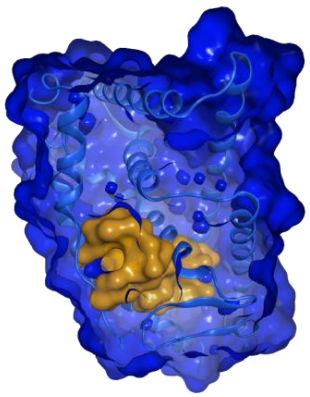
Chemical databases out there : up to 80M molecules

Virtual screening :

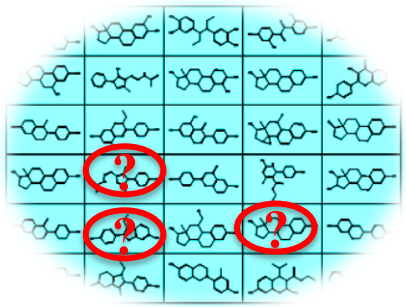
Virtual *in vitro* binding assays

Virtual docking of chemicals in proteins:

identify molecules that may bind in a protein

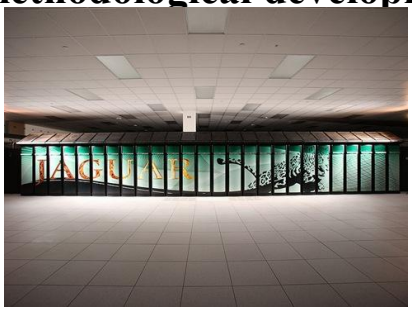


+



Important technology in pharmaceutical industry: drug discovery (hits) and drug design (lead optimization)

methodological developments

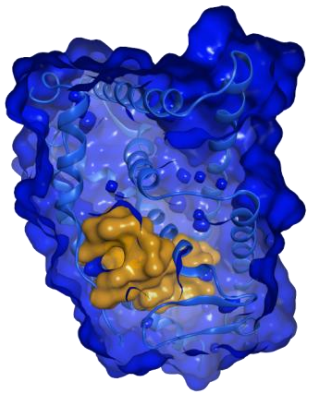


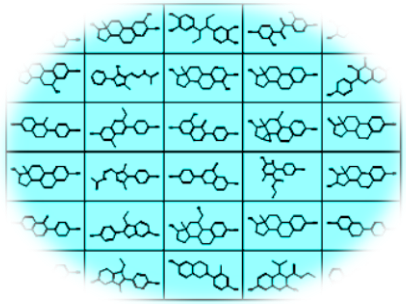
<u>An interdisciplinary approach:</u>
Computational Sciences
Structural Biology, Biophysics, Biochemistry
Chemistry/Medicinal Chemistry

Leverage the power of the world's top supercomputers

Virtual docking of chemicals in proteins:

identify molecules that may bind in a protein





Identify molecules that are likely to bind a given protein : reducing the size of the haystack (NOT trying to find the needle directly)

= estimating the binding free energy of the drug candidates

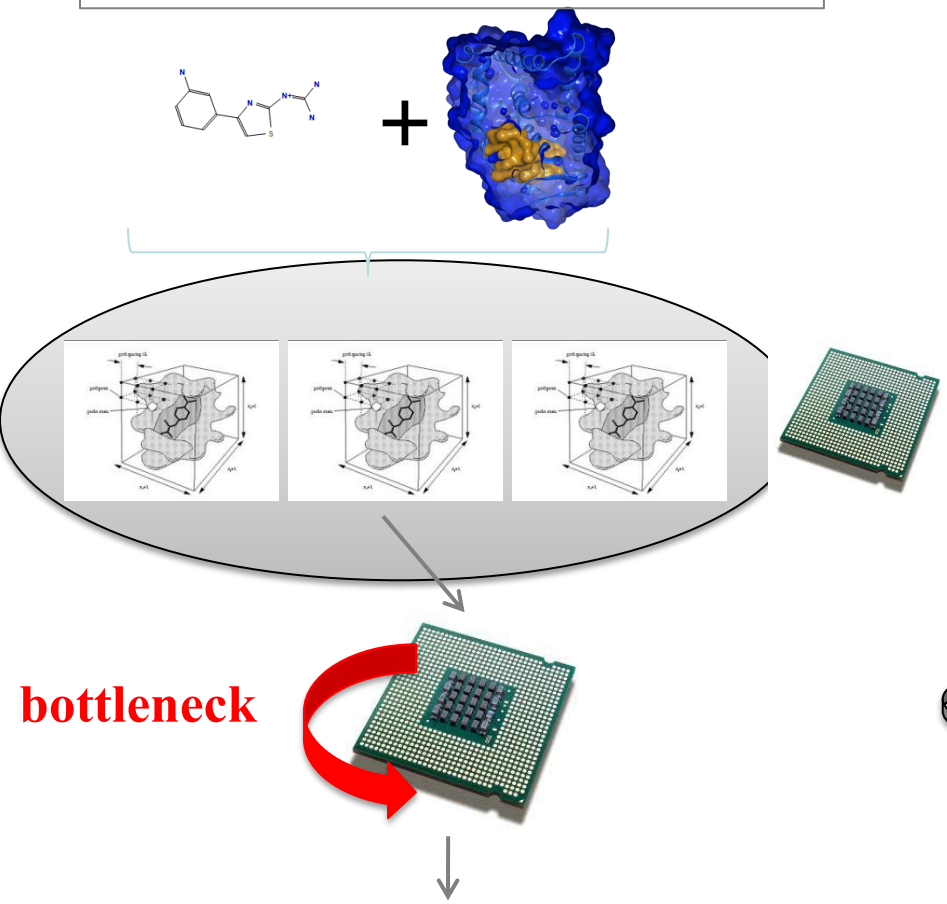
1) **Translating the structures in atomic parameters (atomic coordinates, physico-chemical properties)**
2) **Inserting the drug candidates' data, in many conformations, inside the protein**
3) **Calculating the interaction potential energy (a la MD, but more parametrized)**
4) **Translating this into a binding free energy**
5) **Rank the compounds & output energies & structures**

We use the Autodock engine (A. Olson, Scripps, open source, well validated)

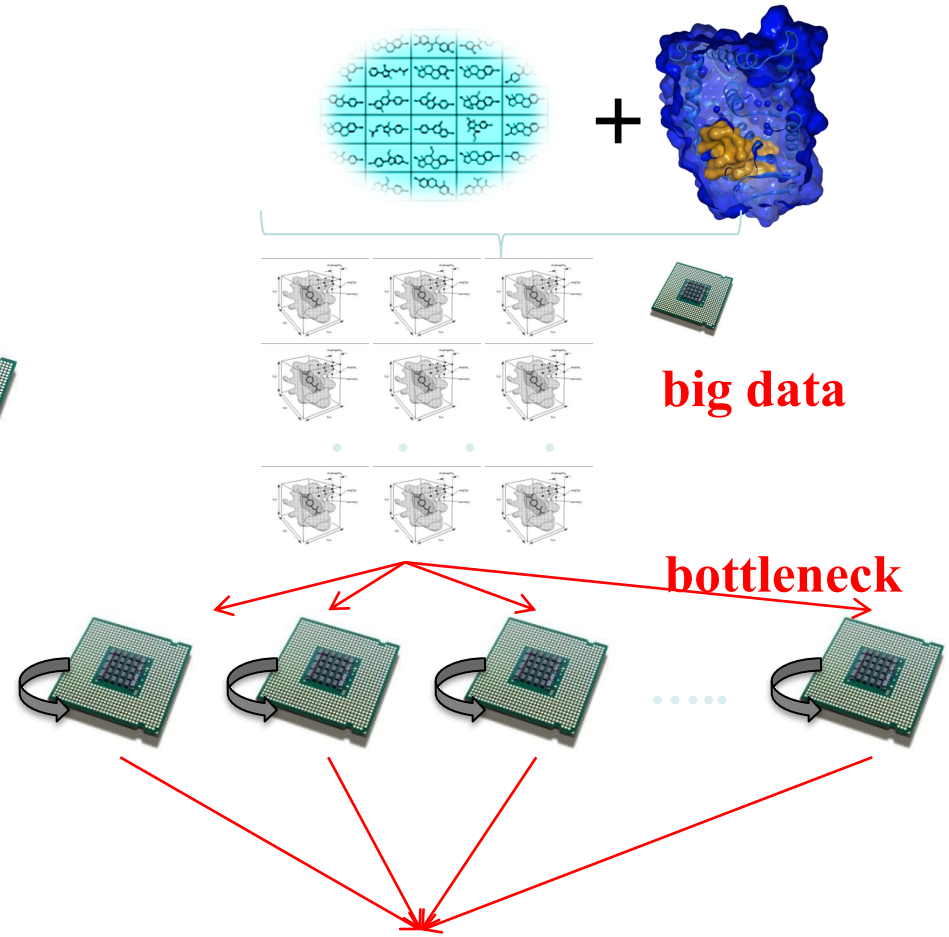
A first encounter with "Big Data"

(on the small side)

Low # of compounds to dock



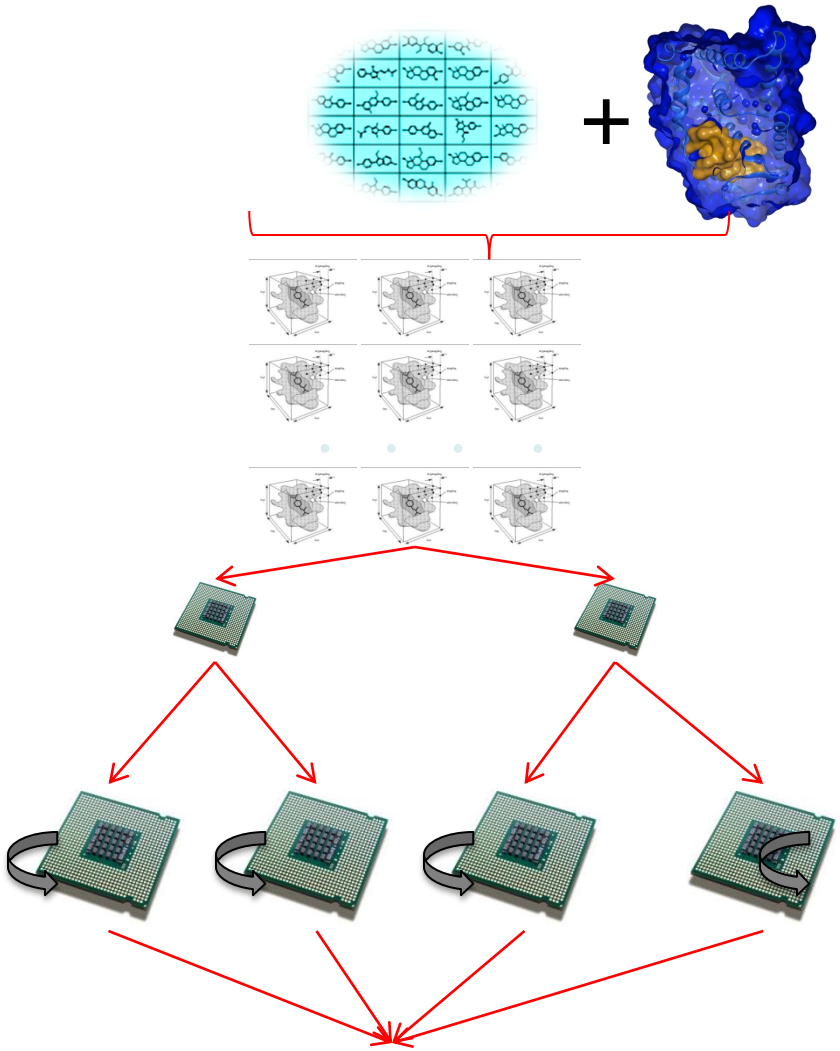
High # of compounds to dock



A first Encounter with "Big Data"

(Collignon, Schulz, Smith, Baudry. *J. Comp. Chem*. 2011,)

High # of compounds to dock



Optimize pre-processing and post processing workflow on Lens

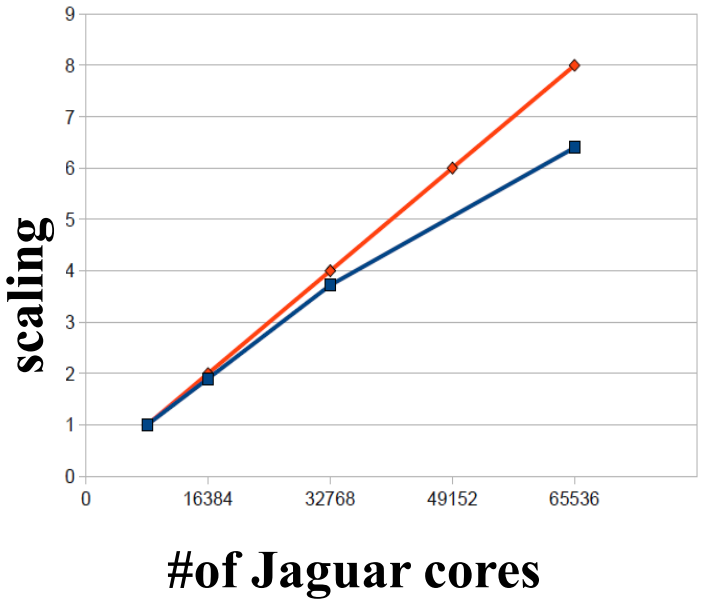
Optimization of primary data format (binary grids)

MPI handle job distribution between master & workers

Parallelization of Autodock on Jaguar

(Collignon, Schulz, Smith, Baudry. *J. Comp. Chem*. 2011,)

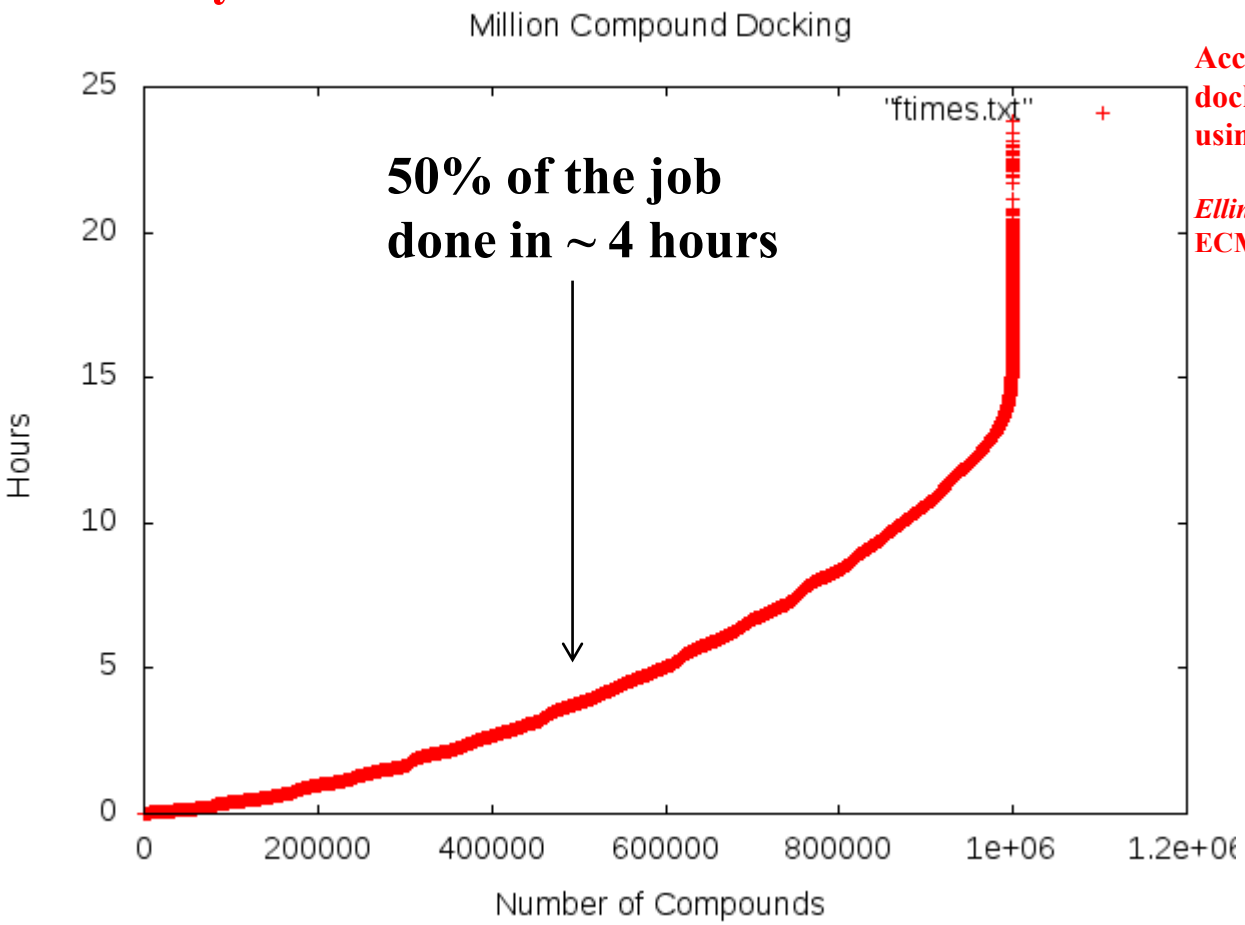
Our Autodock4 MPI parallelization



before data optimization:	**1 protein**	**10^4 ligands**	**1 day**
Presently:	**1 protein**	**10^6 ligands**	**1 day**
	10 proteins	**10^5 ligands**	**1 day**
Exascale (2018):	**100 proteins**	**10^7 ligands**	**1 day**

Parallelization of Autodock on Jaguar

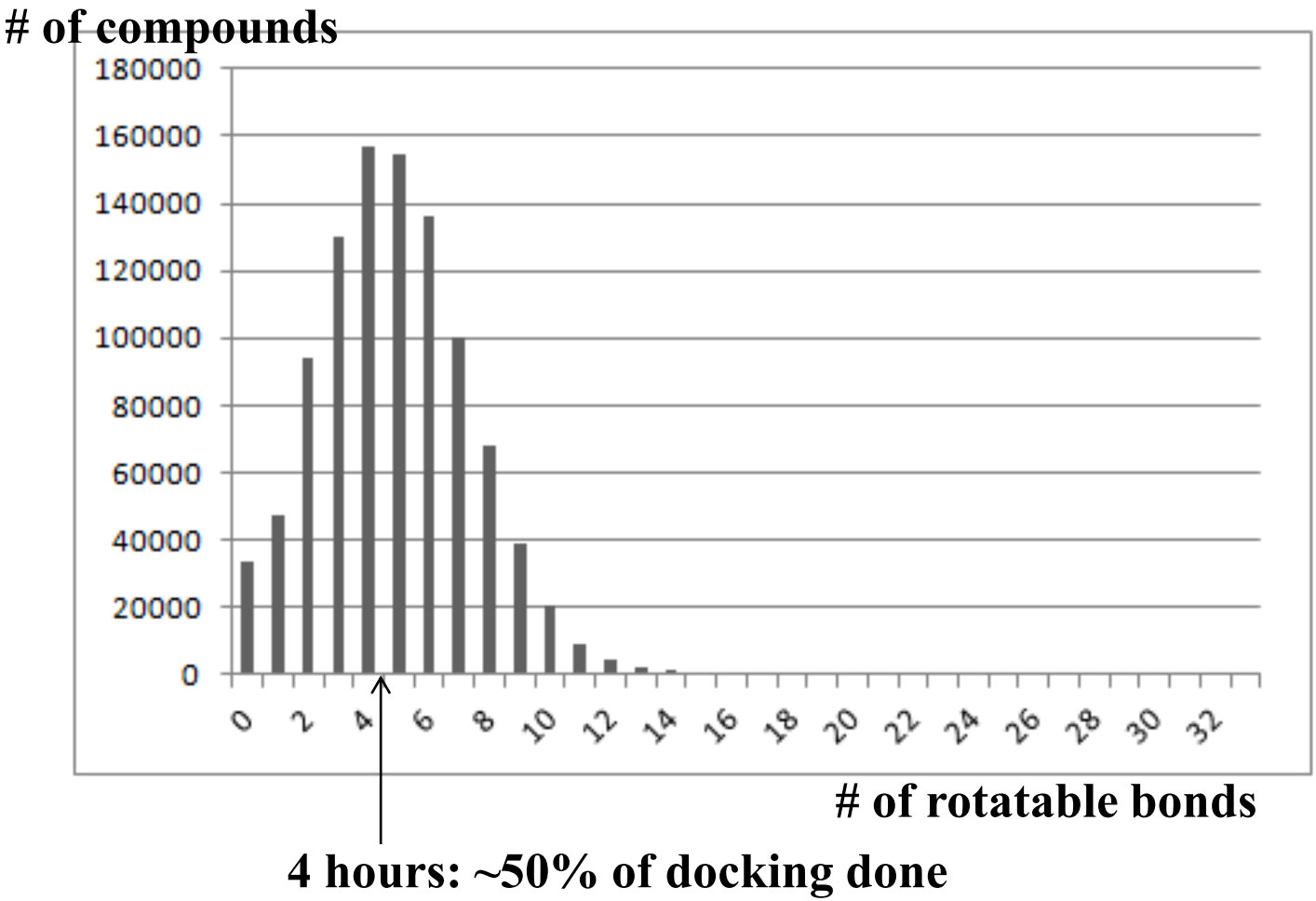
NIH-funded Application Georgetown University / ORNL :
1M compounds screening (most of them non 'drug-like'), flexible docking, HDAC4 target:
~ 1 day on 65K JAGUAR cores



Accelerating virtual high-throughput ligand docking: screening one million compounds using a petascale supercomputer.

Ellingson, Dakshanamurthy, Brown, Smith, & Baudry.
ECMLS12

Application: GU's 1M compounds, flexible docking, HDAC4 target:
~ 1 day on Jaguar's 65K processors (most of then non "drug-like")



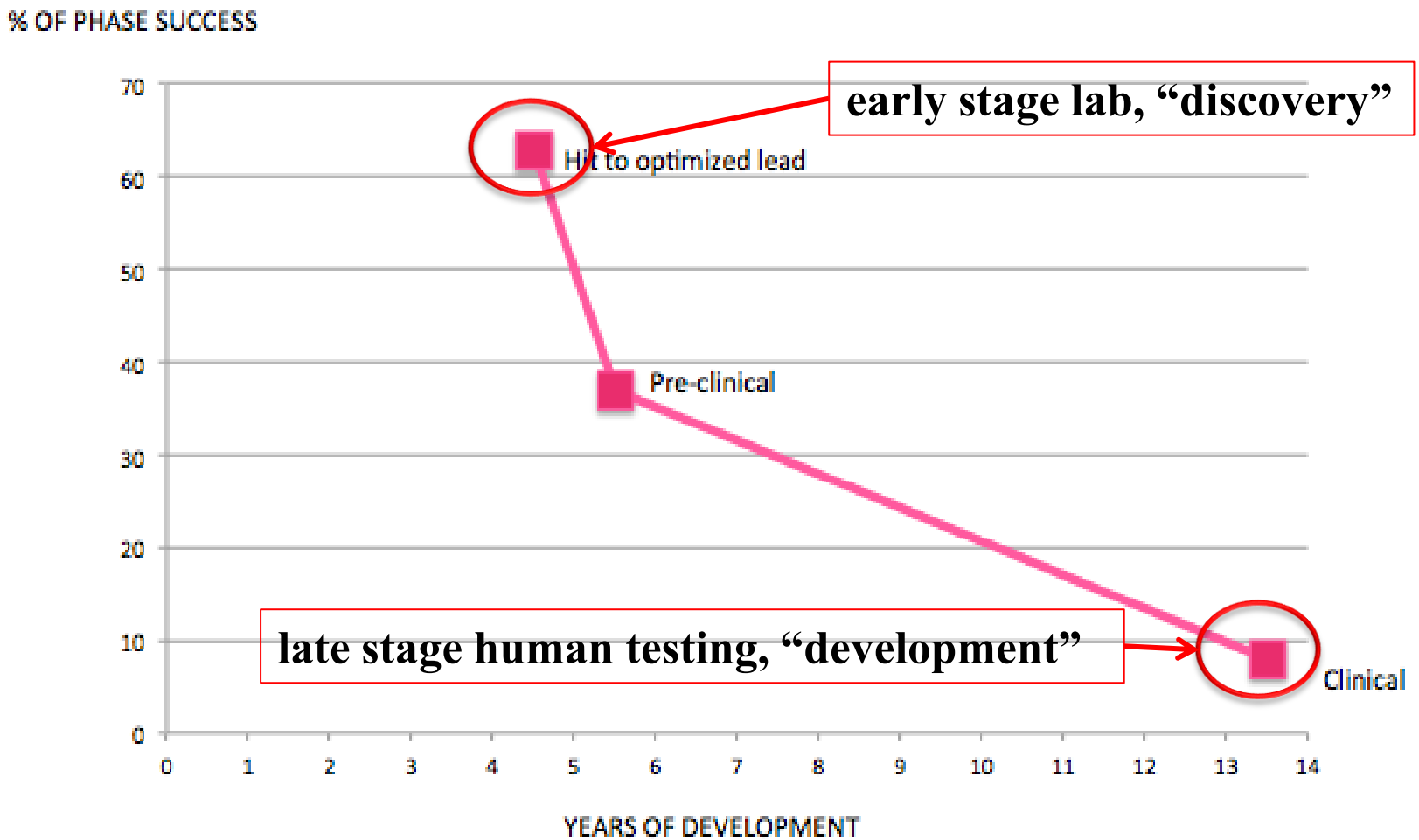
S. Ellingson, S. Dakshanamurthy, M. Brown, J.C. Smith, J. Baudry, ECMLS12

Toward Big Science & Big Data

We have an efficient technology to discover new drug candidates

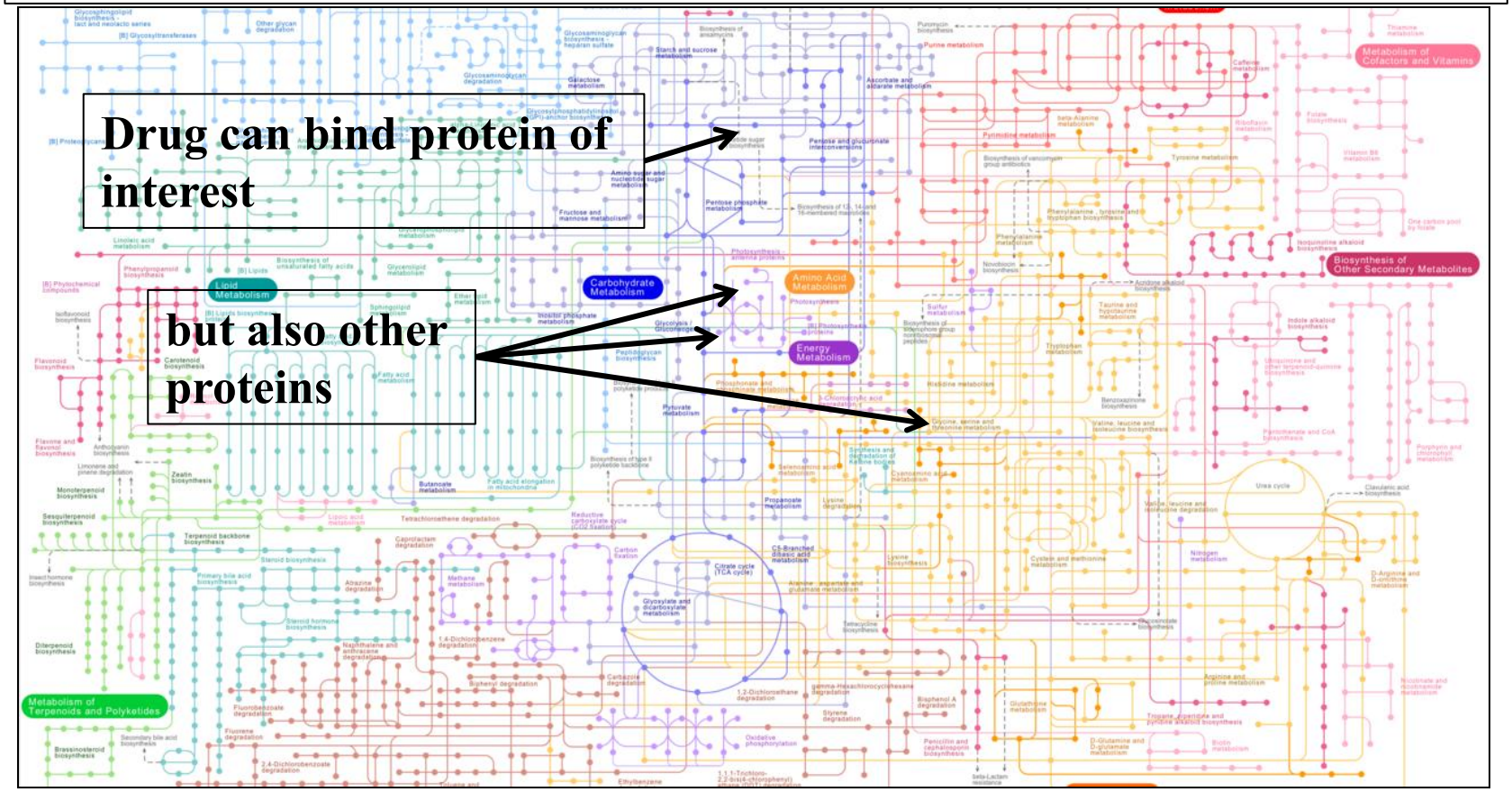
But discovering new drug candidates is only half of the problem (and a small fraction of CPU time & data misery)

Why do drugs fail? (after >10 years and >$1B)



Why do drugs fail? (after >10 years and >$1B)

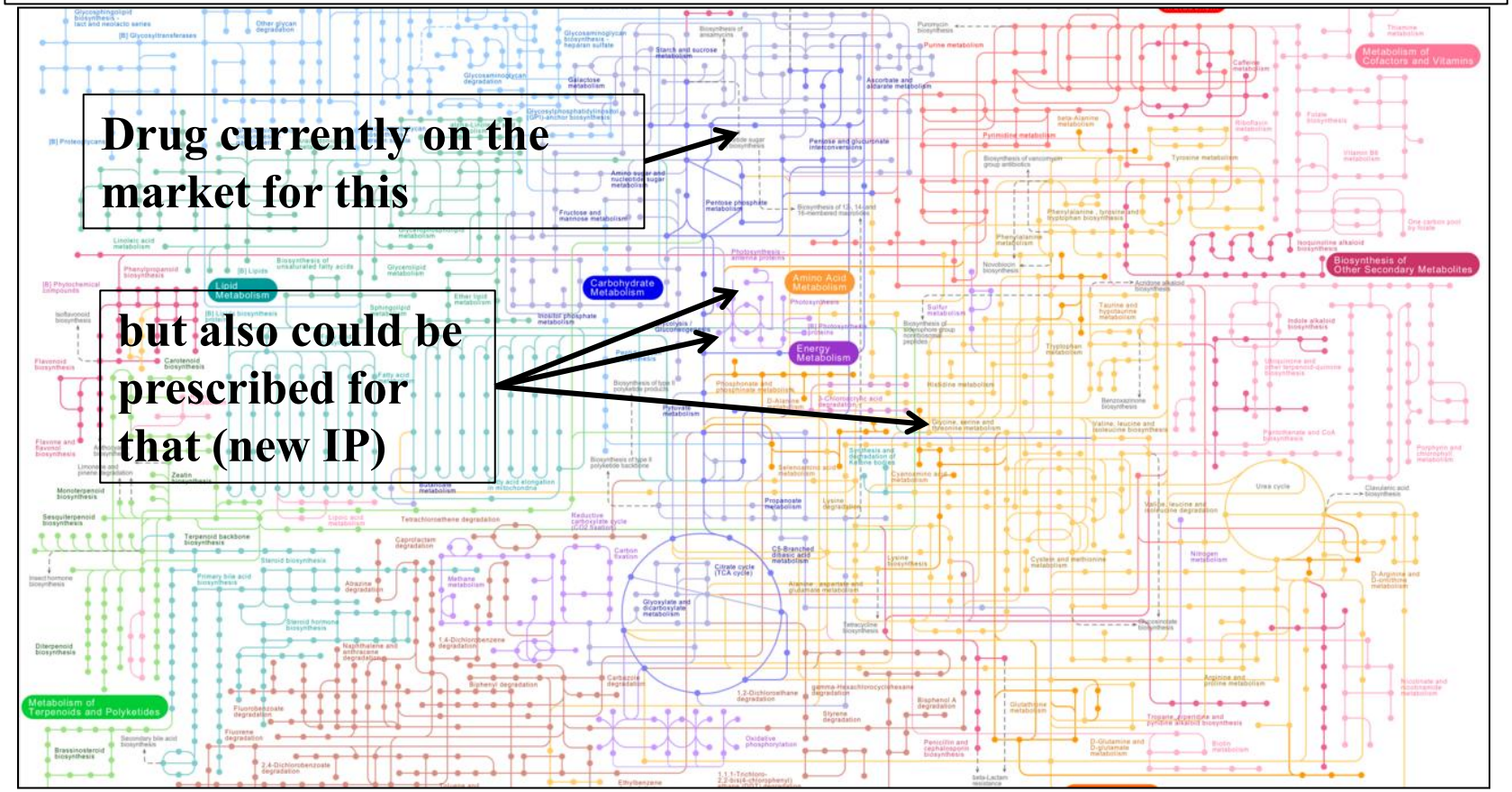
Because they bind to many proteins: not specific enough & TOXICITY



We use HPC to identify ligands that will bind to several proteins: From virtual *in vitro* to virtual cell-based assays

By the way, binding to multiple proteins can also be a very good thing

I : repurposing

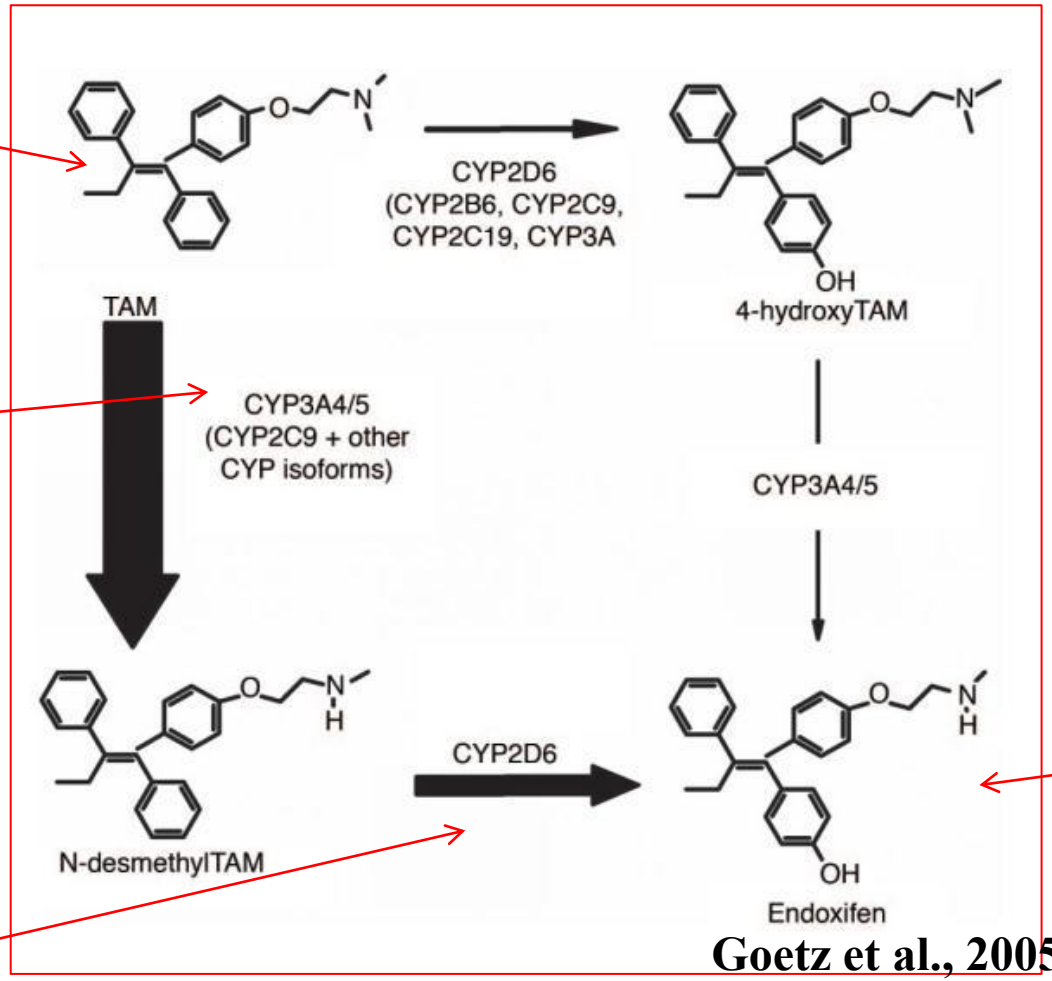


By the way, binding to multiple proteins can also be a very good thing

II : metabolites may be the interesting species, e.g., tamoxifen

patients get this

Protein #1

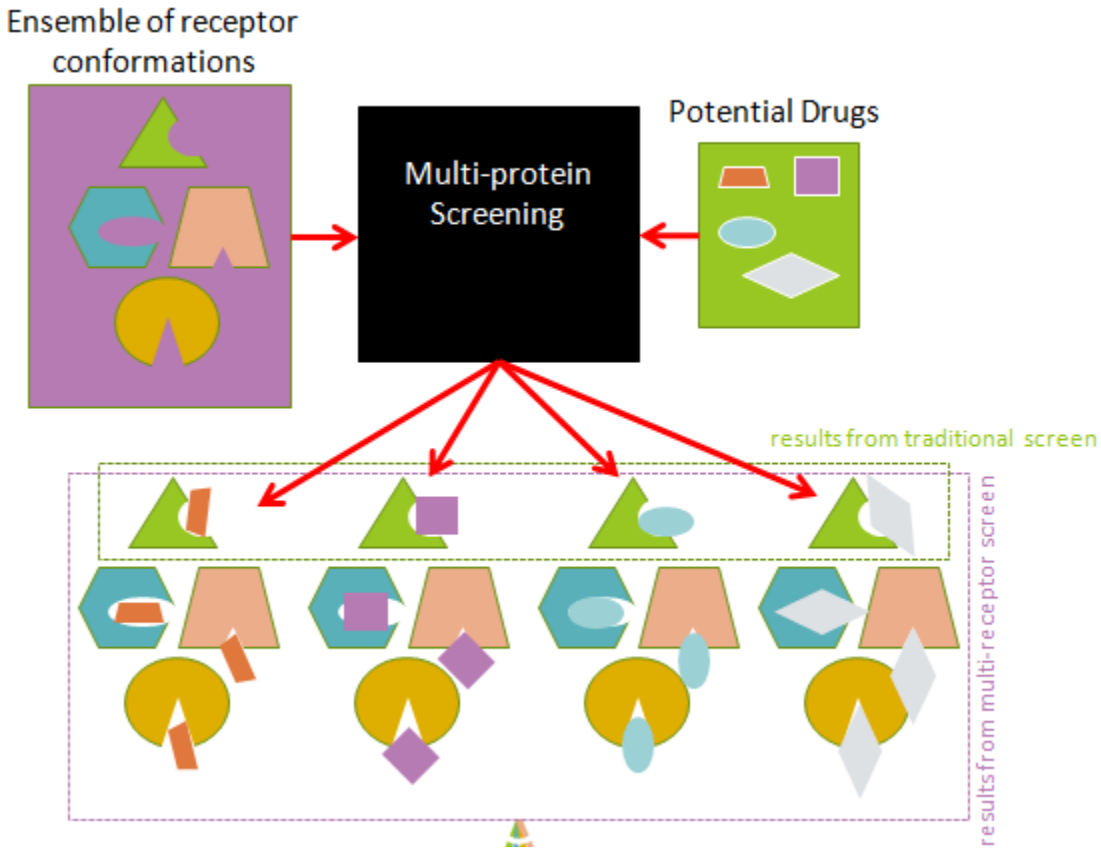


active molecule

Protein #2

Goetz et al., 2005

Multiple protein docking: we are getting more interesting & more complicated



+ MD of each target: sample of structural diversity of receptors

crystal structure snapshot 1 snapshot 2 snapshot … snapshot n

Big Data Challenges

The existing technology has not been too concerned with data so far

But going for multiple protein docking gets insane quickly:

- Applications (1M compounds) ~0.1 TB of compressed, unstructured data: already a pb to find a way to communicate quickly to our collaborators (secure way, file size restrictions, time of shuffling files around)

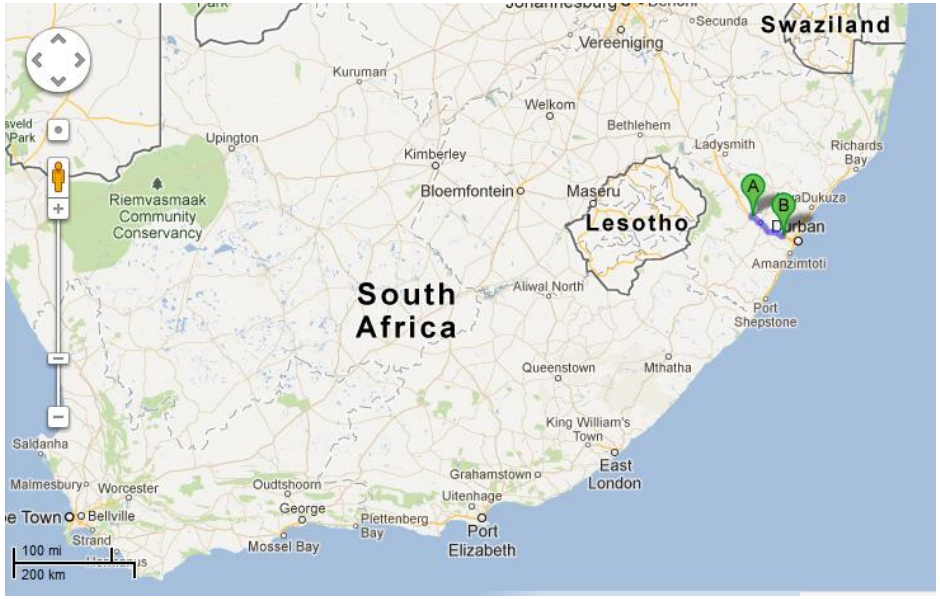
- 80 M chemicals screened in 1protein: 3.5 TB of compressed, unstructured data.
- 70K protein structures: ~0.25 EB
- Using Molecular Dynamics to generate structural snapshots of proteins: > 2EB

We need to go on a data diet, we need to make sure we only produce and store what we need to mine down the road.

But what is it that we need to mine? Different 'customers' have different requests (energy, structure in xyz coordinates, structures in chemical symbolism…)

“Big Data”: transmit the data

- **IPoAC?** (IP over Avian Carriers, RFC1149)



2009, Howick to Hillcrest in SA: ~60 km

4GB of data

Winston the homing pigeon & microSD

vs.

Telkom ADSL line

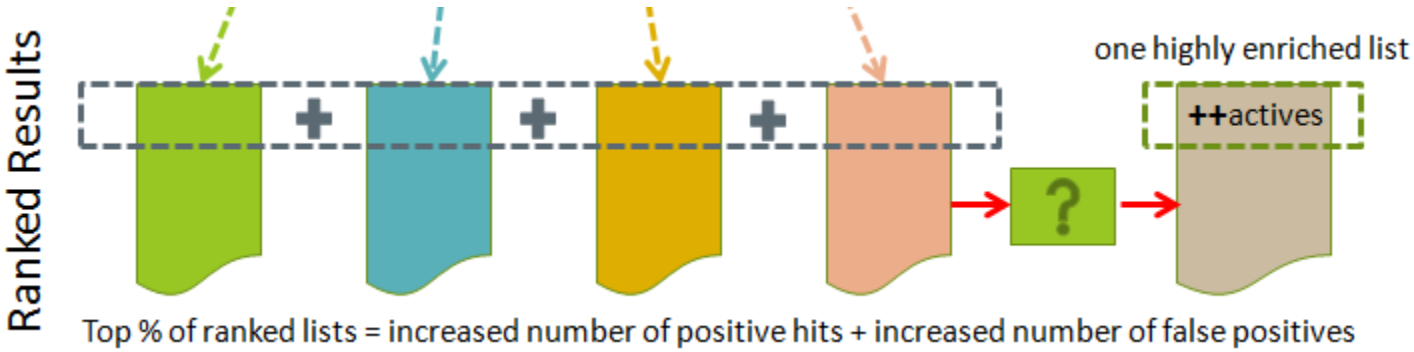
Results: Winston: 2hours, 6 minutes and 57 seconds (from upload to download)

Telkom ADSL: transfer 4% complete by that time

We would need about ~70M pigeons so we should look for an alternative: adapted infrastructure with bigger, better, faster & more secure pipes

“Big Data”

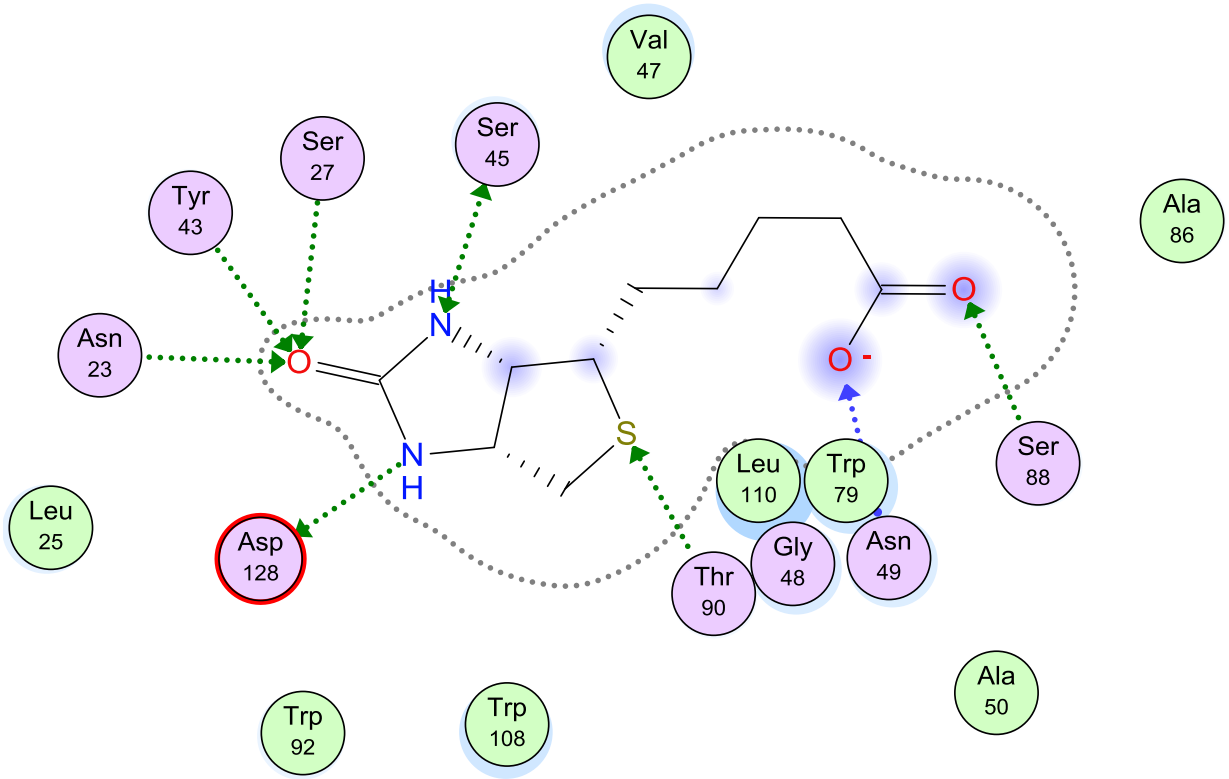
<u>Data analysis & multiple protein docking</u> : we are breaking new ground



How to create this “one highly enriched list”?
We know the data is there, but how to access it?

“Big Data”

<u>Visualization</u> is very important to interact with medicinal chemists



Additional data to create and handle, albeit most likely for a fraction of the screening results (but top 0.5% of 80M compounds = 400K figures like this one)

Conclusions

Virtual assays: moving away form simulating a test tube toward simulating a patient:

- We now have the hardware
- We now have the software

- We will be facing important qualitative & quantitative data issue: data infrastructures a pillar of exascale computing

Thanks

- Jeremy Smith, Sally Ellingson, Jeff Nichols
- Barbara Collignon, Roland Schulz,
- UT/ORNL Center for Molecular Biophysics

- UTK, DOE, NIH, NSF