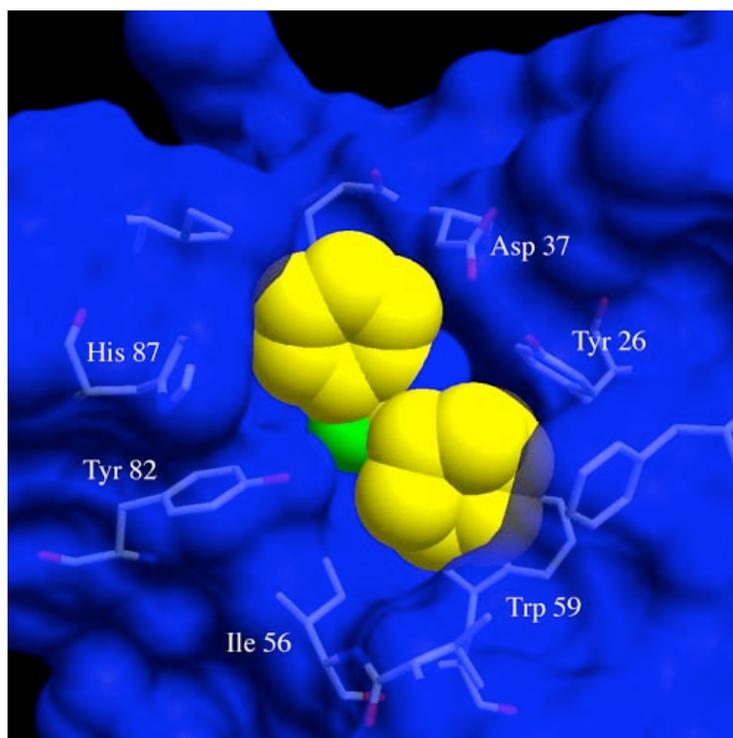# An In Silico Drug Screening Pipeline on BlueGene/L

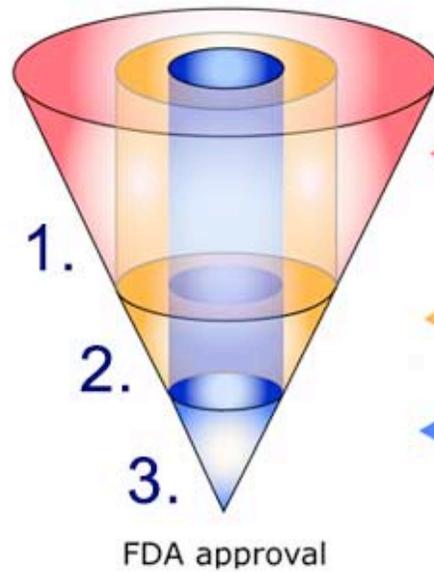Fangfang Xia

The University of Chicago
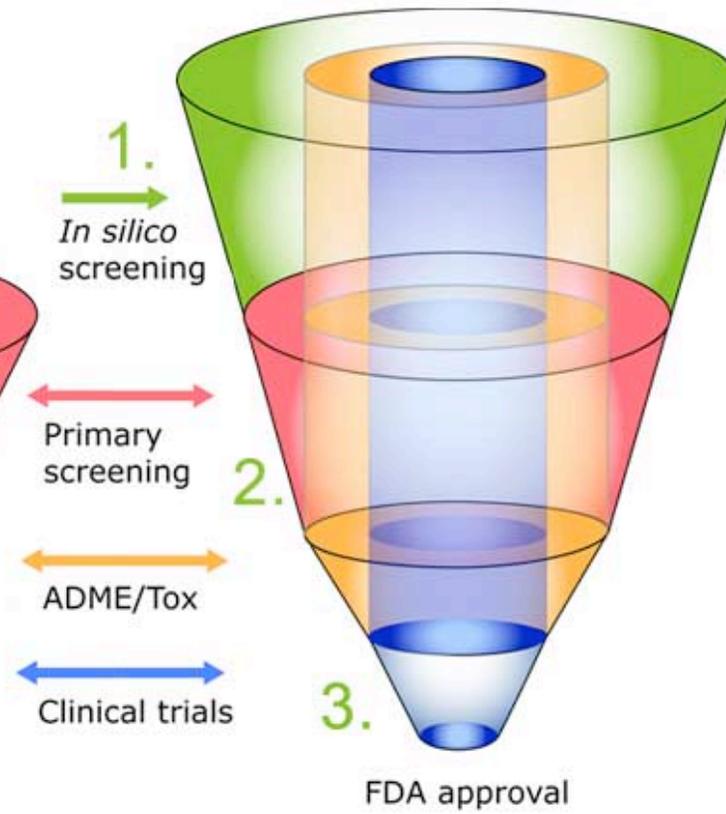
In silico Screening

Classic Drug Discovery Funnel

Optimised Drug Discovery Funnel

1. In silico screening

2. Primary screening

ADME/Tox

3. Clinical trials

1.
2.
3.

FDA approval

FDA approval

Participants

- Rick Stevens (co-PI), Ross Overbeek (co-PI)
- Matt Cohoon, Mike Kubal, Kaitlyn Hwang, Jenifer Zinner, Gohar Margaryan, Mindy Shi, Fangfang Xia

Gene-to-Screen Workflow

1. IDENTIFY GENES/PROTEIN TARGETS FROM THE LITERATURE
   - KNOWN ANTIBIOTIC TARGETS AND KNOWN ANTIBIOTIC RESISTANCE FACTORS
   - ESSENTIAL GENES AND VIRULENCE ASSOCIATED FACTORS
2. DETERMINE FUNCTIONAL ROLE OF EACH TARGET
   - COMPARATIVE ANALYSIS (CLUSTER ANALYSIS, SUBSYSTEM RECONSTRUCTION)
3. SEARCH FOR KNOWN INHIBITORS OF EACH TARGET
   - LITERATURE MINING AND COMPUTATIONAL SCREENING (SEE STEP 8)
4. DETERMINE DEGREE OF CONSERVATION ACROSS SPECIES
   - PHYLOGENY AND SEQUENCE ALIGNMENT
   - CHARACTERIZATION OF THE ACTIVE SITE
   - EARLY SCREEN IN HUMAN AND MODEL SYSTEMS
5. DETERMINE STRUCTURE OF EACH TARGET (PDB, COMPUTATION)
   - DATABASE SEARCH/SIMILARITY AND STRUCTURAL MODELING
6. DETERMINE ACTIVE SITE OF EACH TARGET
   - COMPUTATIONAL ANALYSIS OF EACH STRUCTURE
7. DETERMINE DRUGABILITY OF EACH TARGET
   - SIZE OF POCKET, NUMBER OF POCKETS
8. SCREENING OF COMPOUNDS FOR BINDING AFFINITY ETC.
   - COMPUTATIONAL AND HIGH-THROUGHPUT EXPERIMENTS
9. TOXICITY SCREENING IN HUMAN AND MODEL SYSTEMS
   - COMPUTATIONAL AND HIGH-THROUGHPUT EXPERIMENTS

DOCK5 pipeline
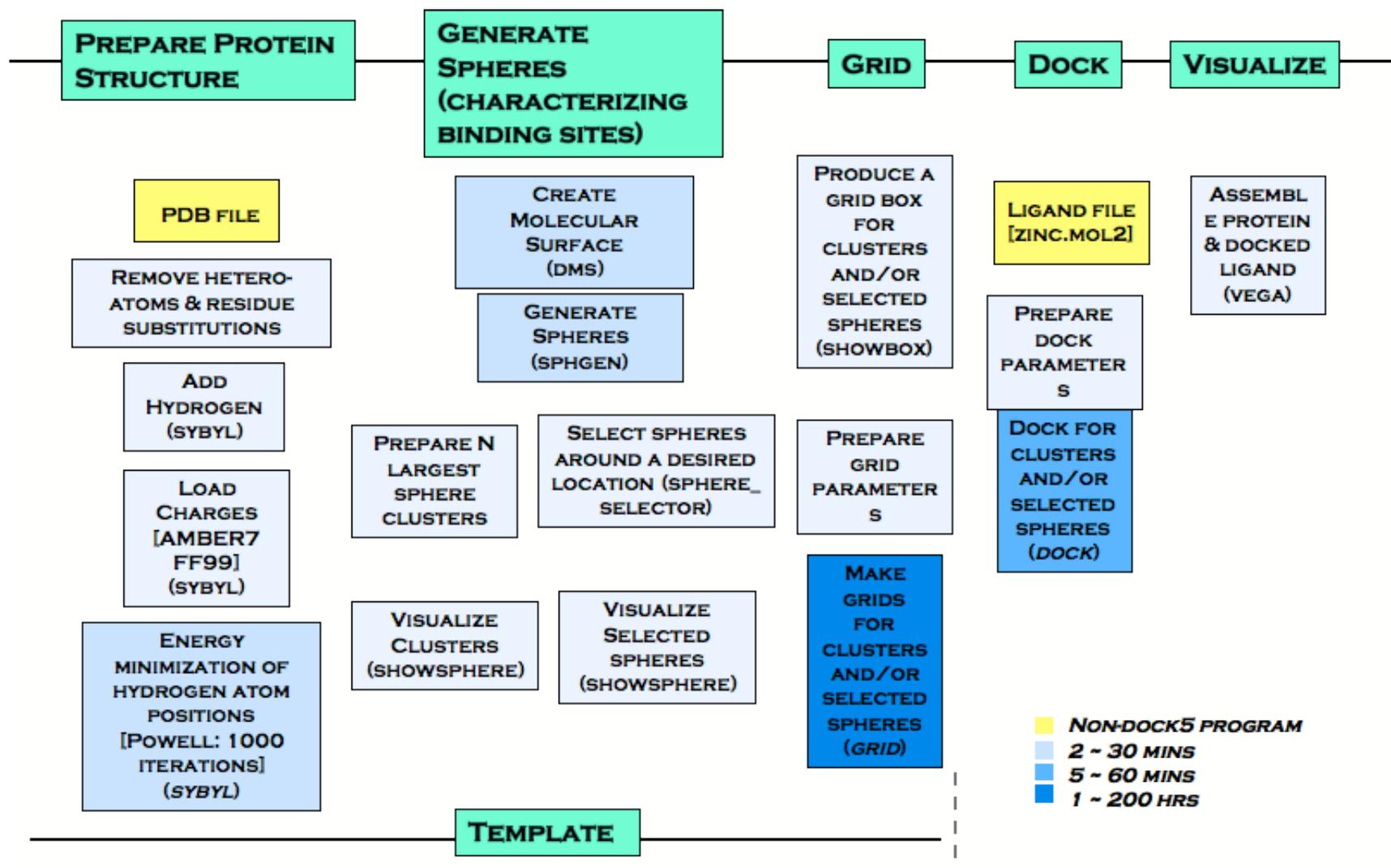
Parallel Programming

- Ported DOCK5 (UCSF) and AutoDock3 (Scripps) to BG/L

- Trivially parallellizable: Master/Slave w/ MPI

- I/O patterns:

  - Master: parses one ligand database file and passes individual ligand information to slaves

  - Slave: does actual "docking" and generates a small output file with a docking score and optimal molecular configurations

Visualization and Analysis

- Top ligands
- Compare docking score and configuration w/ those of natural ligands
- Look into very interesting results w/ protein visualization tools such as Protein Explorer

Visualization and Analysis

CoaD

Status and Scalability

- ## DOCK5
  - Scales poorly beyond 256 processes
  - Currently run with 32/64 processes
  - Memory footprint often exceeds 512M
  - Templates have to be prepared on big-endian machines
- ## AutoDock3
  - Scales well to 1024 processes
  - Runs in VN mode on BG/L

Roadmap

- Current throughput
  - DOCK5: 10,000 successful screens per BlueGene/L rack (1024 nodes) per day
  - AutoDock3: 50,000 screens per rack per day
- Searching for new antibiotics
  - 300 essential-gene-products x 3.3 million compounds $\Rightarrow$ 990 million drug docking computations (each one involves about 20 different computations) $\Rightarrow$ over 10 billion jobs
- Adding a neural network component to filter unpromising ligands based on high level features