PubChem Mining - From Small Molecule to Structures and Bioactivity

Luke Huan
Associate Professor
Electrical Engineering and Computer Science
University of Kansas
http://people.eecs.ku.edu/~jhuan/




Group Members

Ph.D. Students:

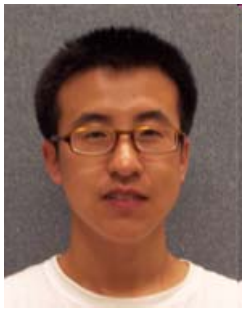
Aaron Smalter, Bria Quanz, Hongliang Fei, Leo Zhang, Jia Yi

Department of EECS, University of Kansas

Master Student:

Xiaohong Wang





Collaborators

- Dr. Jeff Aubé, KU School of Pharmacy
- Dr. Deepak Bandyopadhyay, GSK
- Dr. Gerald H. Lushington, KU Molecular Graphics and Modeling Laboratory
- Dr. Leming Shi, FDA
- Dr. Alex Tropsha, UNC School of pharmacy

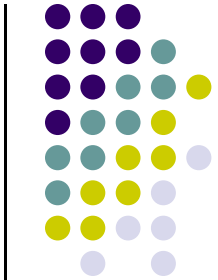




Acknowledgments

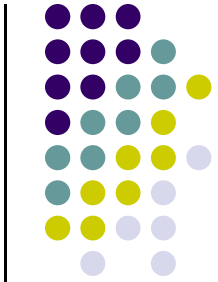
- The work is partially supported by
  - National Science Foundation, “CAREER: Mining Genome-wide Chemical-Structure Activity Relationships in Emergent Chemical Genomics Databases”, (IIS 0845951)
  - National Human Genome Research Institute “KU Special Chemistry Center” (U54 HG005031)
  - National Center for Research Resources, “KU Bioinformatics Computing Facility Core Renovation and Improvement” (RR031125)
  - University of Kansas Faculty General Research Fund

Why Talking about Data Mining in Drug Discovery



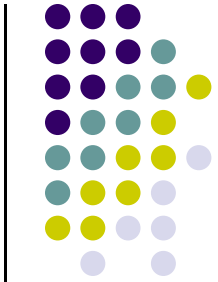
- Drug discovery is highly data driven
  - Chemical structure
  - Protein sequence, structure, and expression
  - Genome and gene
  - Biological network
  - Pharmacokenetics and pharmacodynamics
- Data are increasingly becoming public available
- Having ample data, demanding more knowledge!
- We see many different data types
  - Vector, semi-structured, time-series, spatial-temporal, images, video, hypertext, literature
- Data analysis and data management challenges are from all aspects
  - Large volume, high dimensional, high noise, large amount of missing values, non iid data, structured input and output, unlabeled data
  - Multi-instance (label, class, task)
- Spans the full data analysis cycles
  - Data collection, data cleasing, data semantics, data integration, data representation
  - Model inference, model selection, modal average, model interpretation

Outline



- Drug Discovery Pipeline
- Overview of PubChem
- Chemical Structure Based Prediction Problems with Kernel Methods
- Advanced Topics of Data Analysis in Drug Discovery

Part I: Drug Discovery Pipeline



- Overview of Drug Discovery and Development
- Pre-discovery of Drugs
  - Target identification/validation, assay development, hit identification, lead identification, early safety tests, lead optimization, preclinical testing
- Drug Discovery
  - Investigational new drug (IND), clinical trials phase I, II, and III, new drug application (NDA), manufacturing, post-market analysis
- Concluding Remarks

Selected Landmarks in Drug Discovery

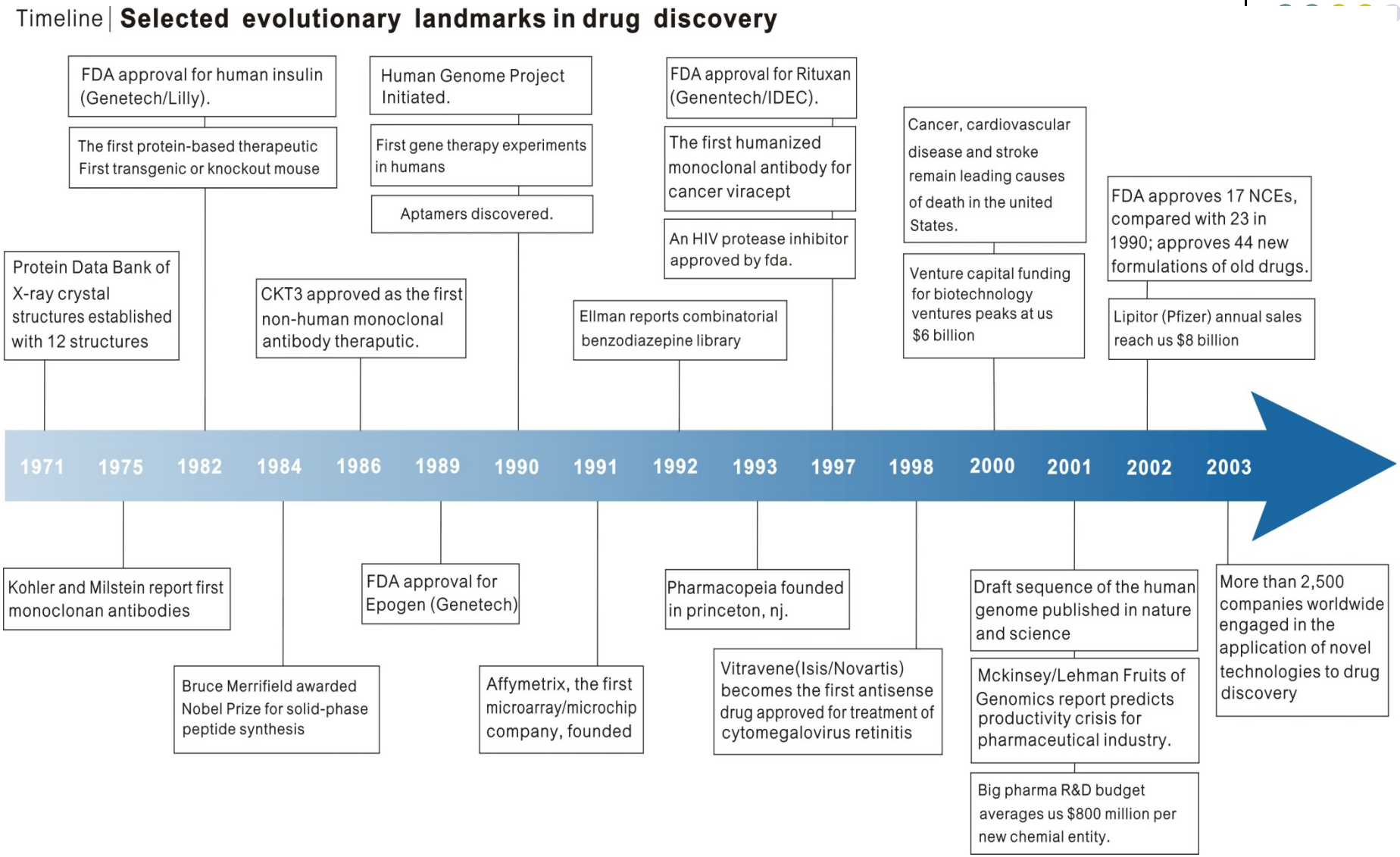


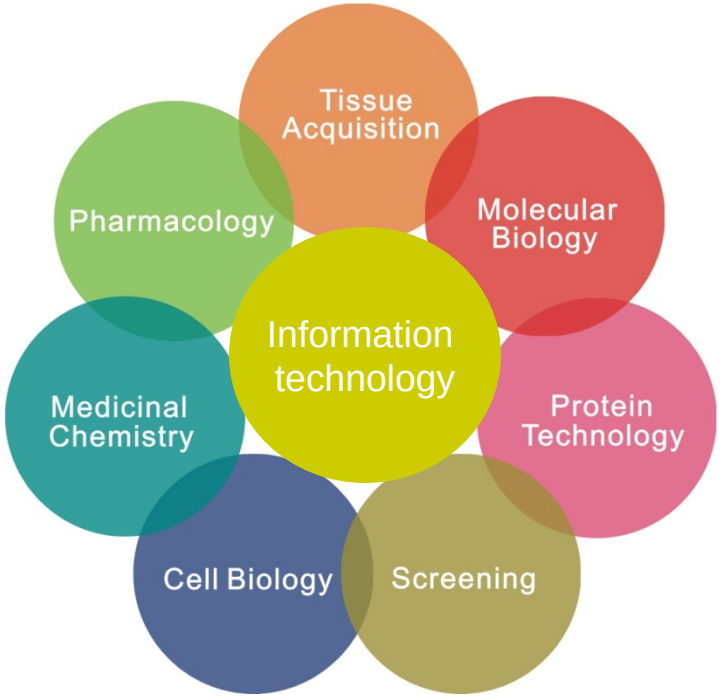
Figure adopted from: L.J. Gershell *et al.* A brief history of novel drug discovery technologies, *Nat. Rev. Drug Discov.* 2, 321-327 (2003)

Overview: **Drug Discovery and Development**

- **Highly Interdisciplinary**: Recent advances in genomics, proteomics and computational power present new ways to understand human diseases at the molecular level.

- **High Attrition Rate**: For every **5,000-10,000** compounds that enter the research and development (R&D) pipeline, ultimately only one receives approval.

- **Complex**: Success requires immense resources — the best scientific minds, highly sophisticated technology, complex project management, and sometimes, luck.



Timescale: **Drug Discovery and Development**

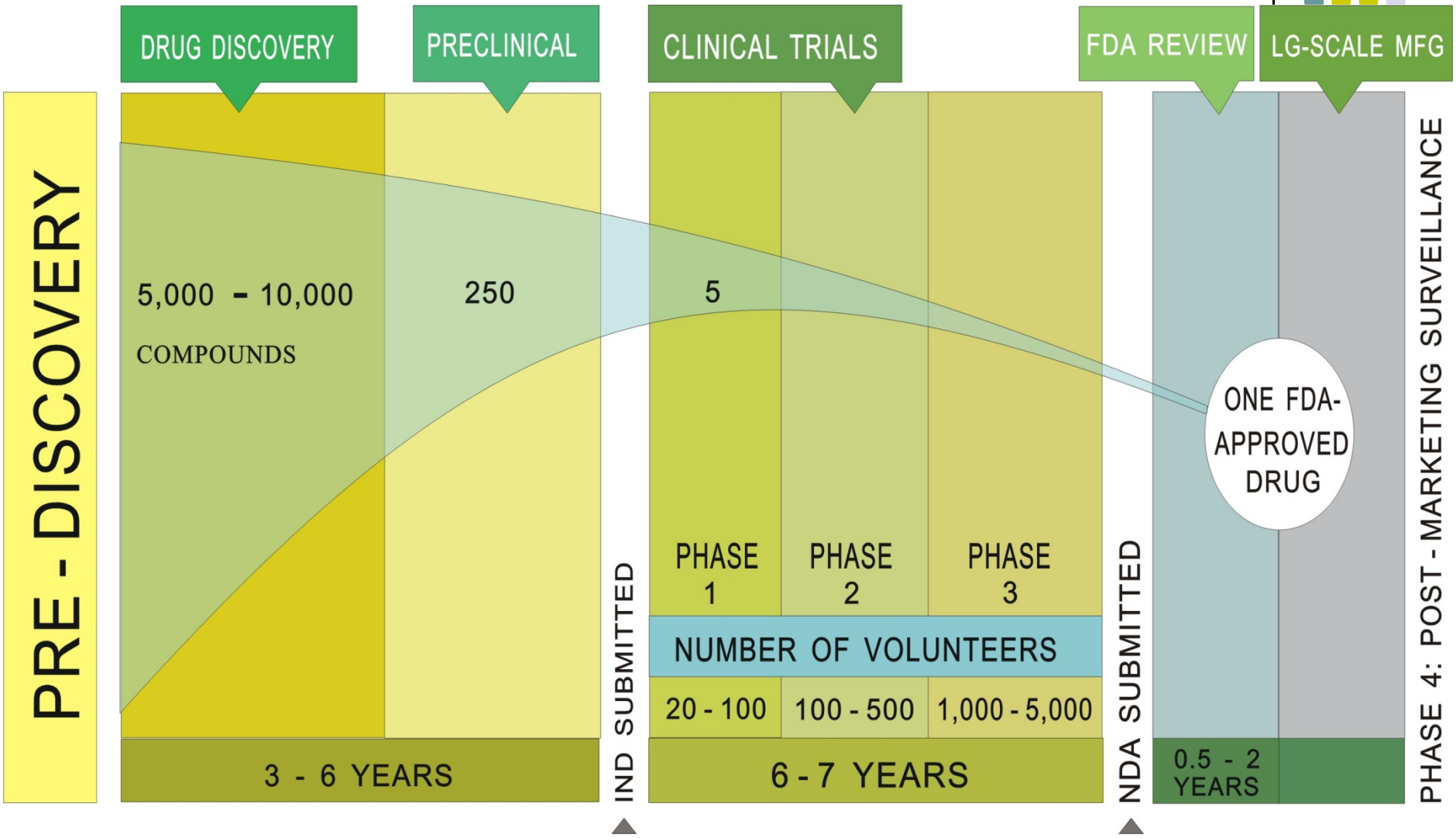


Figure adopted from the brochure of INNOVATION.ORG "*Drug Discovery and Development: Understanding the R&D Process*".

Process: **Drug Discovery and Development**

- This whole process takes an average of **10-15 years.**

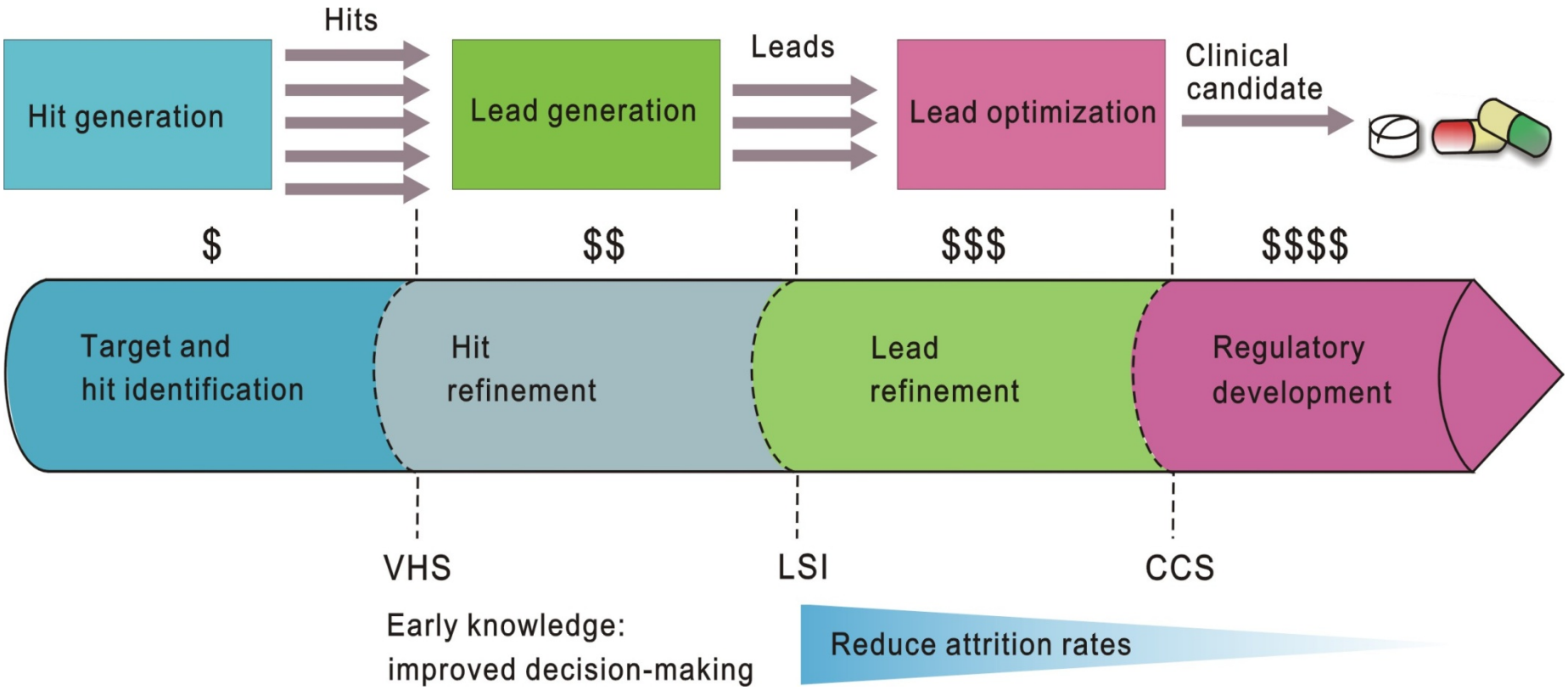
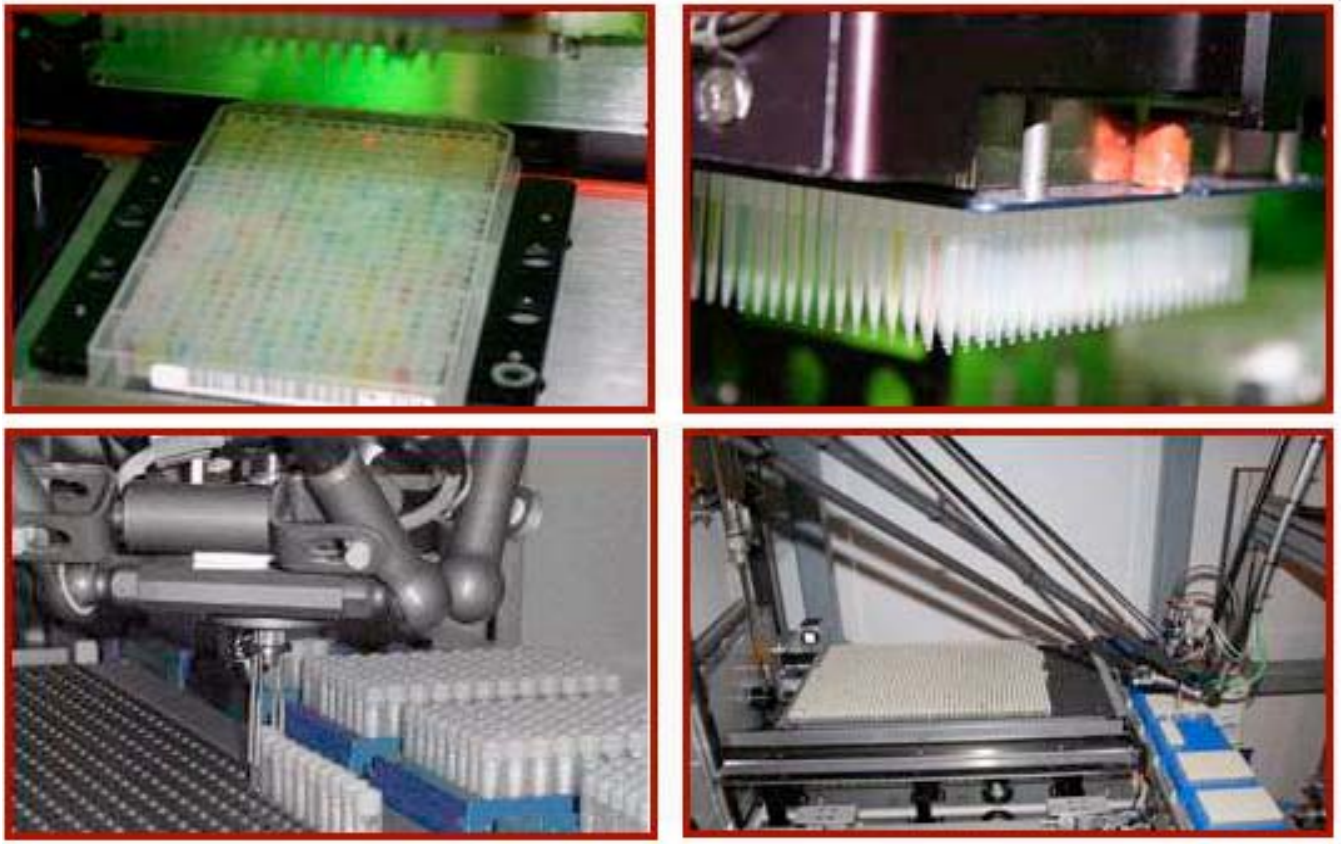


Figure adopted from: M.A. Cooper, Optical biosensors in drug discovery, *Nat. Rev. Drug Discov.* 1, 515-528(2002)

Drug Discovery: **Assay Development**

- **High-throughput Screening** is a widely used approach to identify leads.
  - Advances in robotics and computational power allow researchers to test hundreds of thousands of compounds against the target to identify any that might be promising.



Cited from the Internet: http://www.osip.com/scires_coretech

Drug Discovery: **Lead Identification**

- Newly invented pharmacologically active moieties may have poor drug-likeness and may require chemical modification to become drug-like enough to be tested biologically or clinically.

- A **lead compound** is a starting point for chemical modifications in order to improve potency, selectivity, or pharmacokinetic parameters.

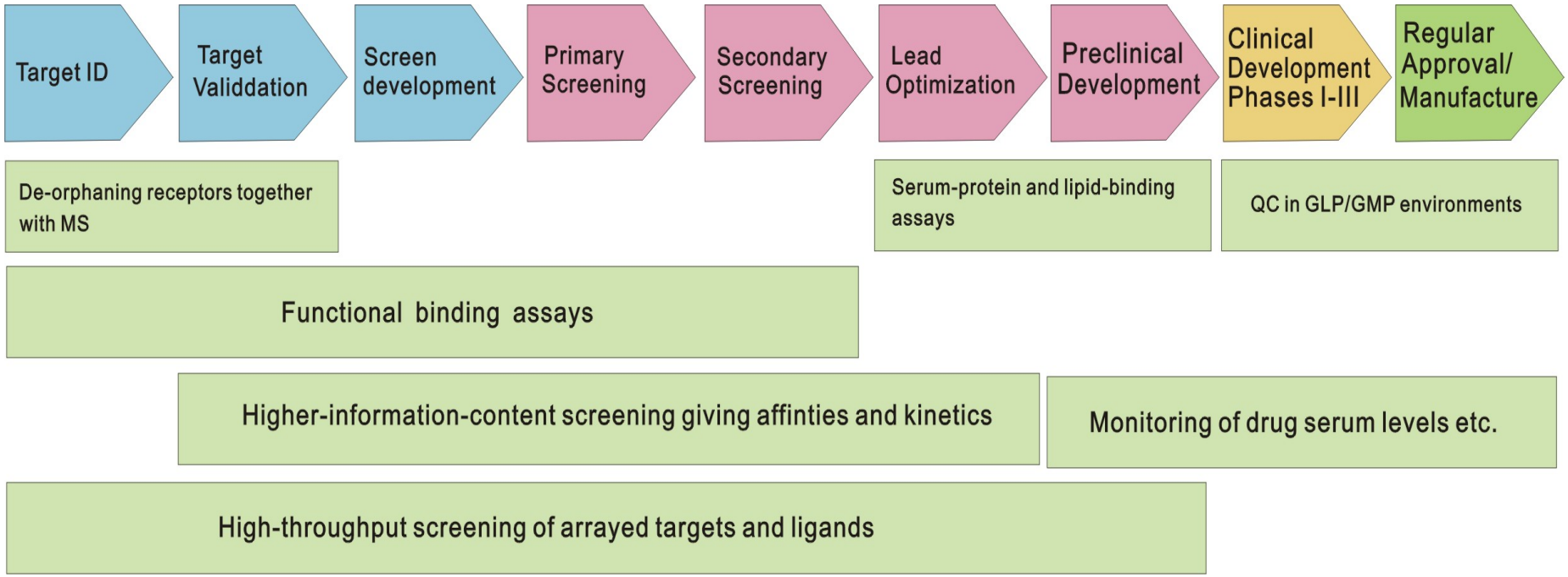
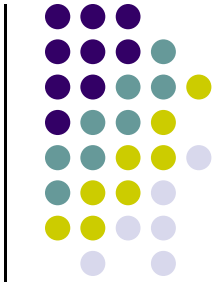


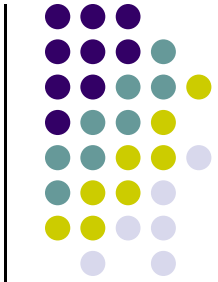
Figure adopted from: K.H. Bleicher *et al.* Hit and Lead Generation: Beyond High-throughput Screening, *Nat. Rev. Drug Discov.* 2, 369-378 (2003)

Drug Discovery: Early Safety Test



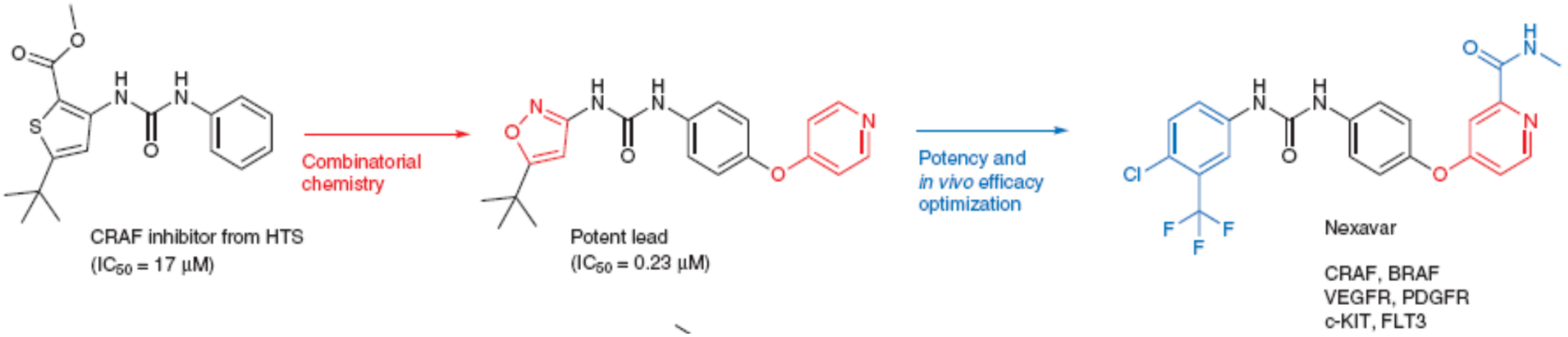
- Lead compounds go through a series of tests (ADME/Tox) to provide an early assessment of the safety of the lead compound.
- Successful drugs must be:
  - **A**bsorbed into the bloodstream;
  - **D**istributed to the proper site of action in the body;
  - **M**etabolized efficiently and effectively;
  - **E**xcreted from the body successfully;
  - demonstrated to be not **Tox**ic.
- These studies help researchers prioritize lead compounds early in the discovery process. ADME/Tox studies are performed in living cells, in animals and via computational models.

Drug Discovery: Lead Optimization

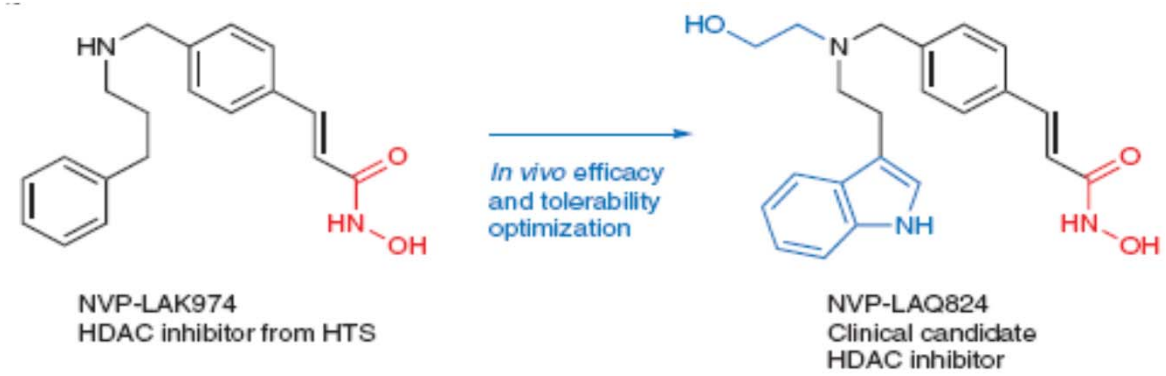


- Lead compounds that survive the initial screening are then “optimized,” or altered to make them more effective and safer.
- By changing the structure of a compound, its properties can be changed, e.g. making it less likely to interact with other chemical pathways and thus reducing the potential for side effects.
- Even at this early stage, researchers begin to think about how the drug will be made, considering formulation and large-scale manufacturing.
- The resulting compound is the candidate drug.

Case Study of Lead Optimization



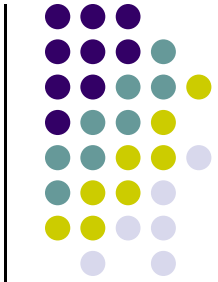
Combinatorial variation of the two substituents on the central urea generated a potent lead (red). Lead optimization focused on improving potency and *in vivo* activity (blue).



The hydroxamate zinc-binding functionality typical of many HDAC inhibitors (red). Lead optimization to the clinical candidate NVPLAQ824 concentrated on improvements to *in vivo* activity and tolerability (blue).

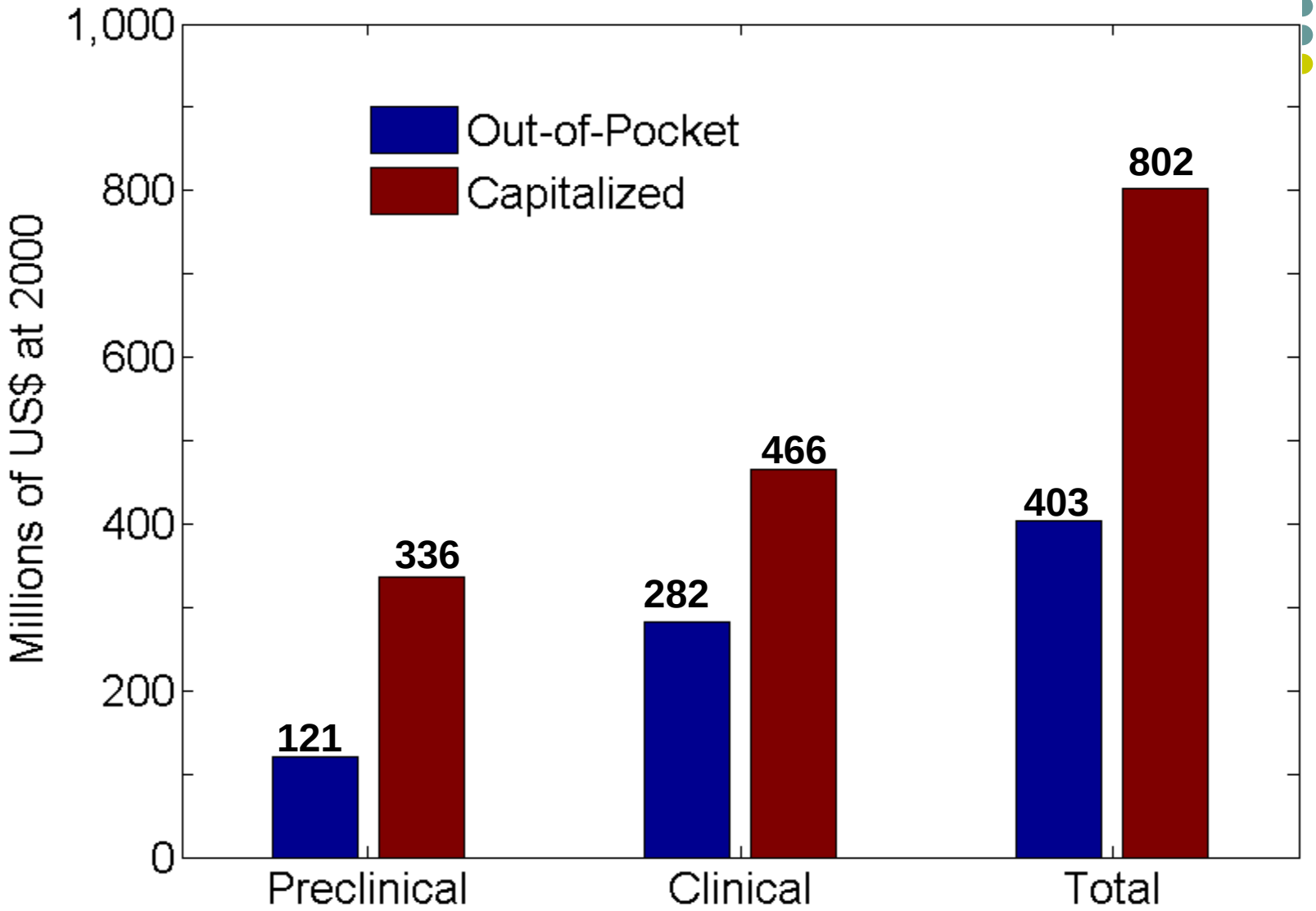
I. Collins & P. Workman, New approaches to molecular cancer therapeutics, Nat. Chem. Biol. 2, 689-700 (2006)

Drug Discovery: Preclinical Testing



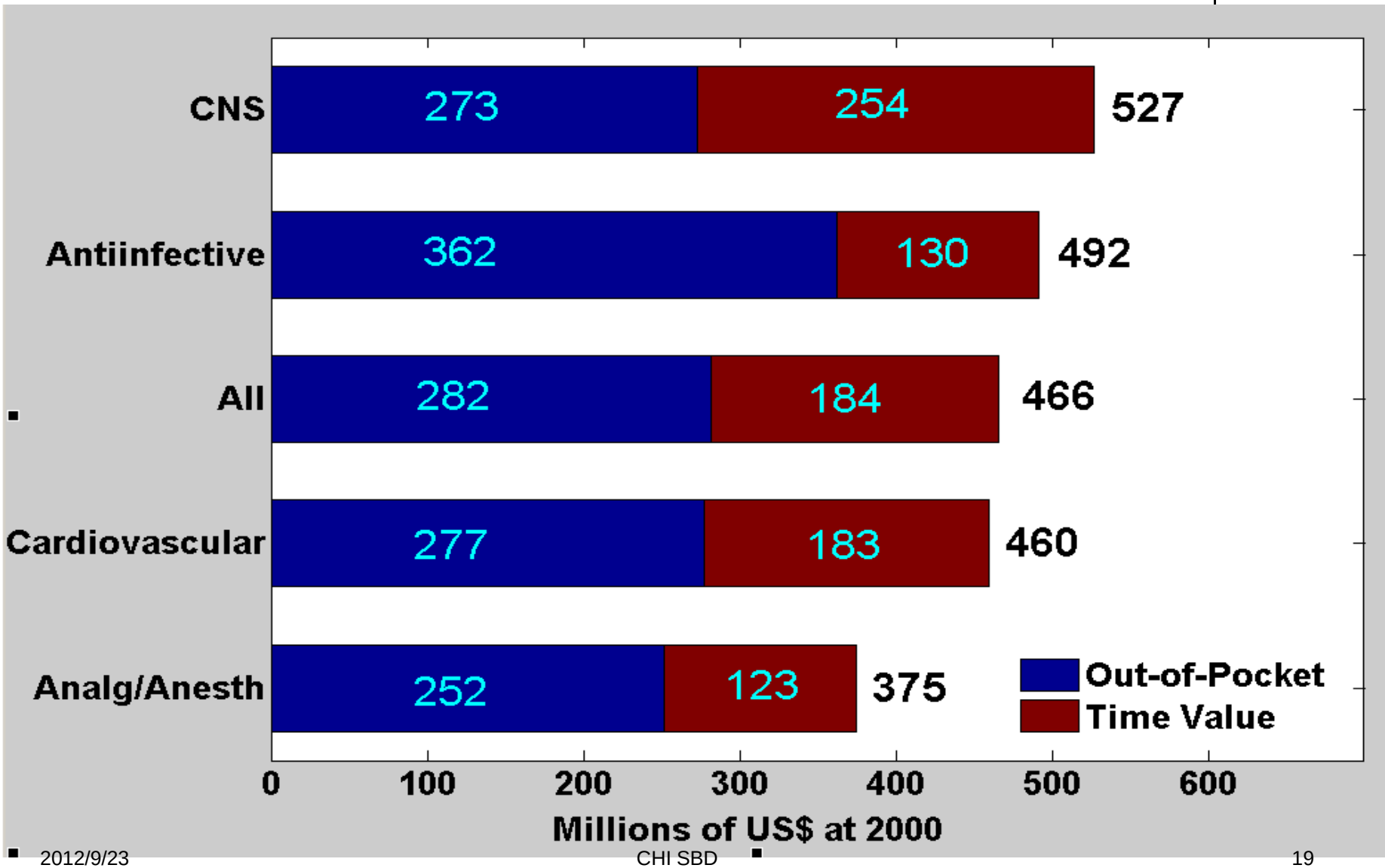
- With one or more optimized compounds, lab and animal testing is used to determine if the drug is safe enough in humans:
  - The FDA requires extremely thorough testing before the candidate drug can be studied in humans;
  - *in vitro* and *in vivo* tests (in living cell cultures and animal models) are carried out to understand how the drug works and what its safety profile looks like.
  - First scale up: how to make large enough quantities of the drug for clinical trials.
  - From 5,000 to 10,000 compounds, one to five molecules, called “candidate drugs,” will be studied in clinical trials.

Pre-approval R&D Cost



Data source: DiMasi et al., J Health Economics 2003;22(2):151-185

Clinical Cost by Therapeutic Category



Data source: DiMasi et al., J Health Economics 2003;22(2):151-185

Drug Development: Phase I Clinical Trial

- **Initial testing in a small group of healthy volunteers for safety**
  - These studies are usually conducted with about 20 to 100 healthy volunteers.
  - The main goal of a Phase 1 trial is to discover if the drug is safe in humans.
  - Researchers look at the pharmacokinetics of a drug: How is it absorbed? How is it metabolized and eliminated from the body? Does it cause side effects? Does it produce desired effects?
  - These closely monitored trials are designed to help researchers determine what the safe dosing range is and if it should move on to further development.

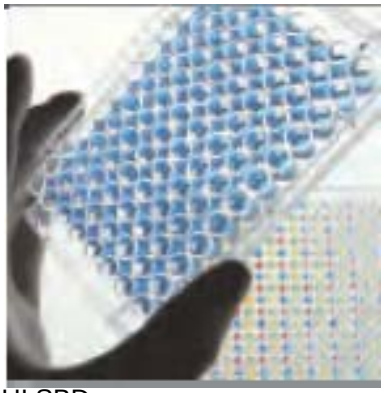


Figure adopted from the brochure of INNOVATION.ORG "*Drug Discovery and Development: Understanding the R&D Process*".

Drug Development: Phase II Clinical Trial

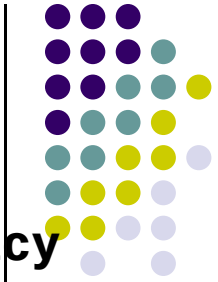
- **Phase 2a and 2b Trials:** Sometimes combined with a Phase I trial

  - Phase 2a trial is aimed not only at understanding the safety of a potential drug, but also getting an early read on efficacy and dosage in a small group of patients.

  - The resulting Phase 2b trial would be designed to build on these results in a larger group of patients for the sake of designing a rigorous and focused Phase III trial.

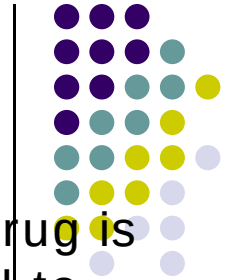
Figure adopted from the brochure of INNOVATION.ORG "*Drug Discovery and Development: Understanding the R&D Process*".

Drug Development: Phase III Clinical Trial



- **Test in a large group of patients to show safety and efficacy**
  - Study the drug candidate in a larger number of patients: about 1,000-5,000;
  - Generate statistically significant data about safety, efficacy and the overall benefit-risk relationship of the drug;
  - Key in determining whether the drug is safe and effective;
  - Provides the basis for labeling instructions to help ensure proper use of the drug (e.g., information on potential interactions with other medicines).

Drug Development: New Drug Application (NDA)



- If the results of all three phases of clinical trials show that the drug is both safe and effective, a NDA with the FDA requesting approval to market the drug.
  - It can be as long as 100,000 pages or more;
  - The NDA includes all of the information from the previous years of work, as well as the proposals for manufacturing and labeling of the new medicine;
  - The FDA can either approve or deny the NDA. It may issue an “approvable” letter requesting more information or studies before approval can be given;
  - Review of an NDA may include an evaluation by an advisory committee. Committees vote on whether the FDA should approve an application, and under what conditions.

What if Phase III Fails?

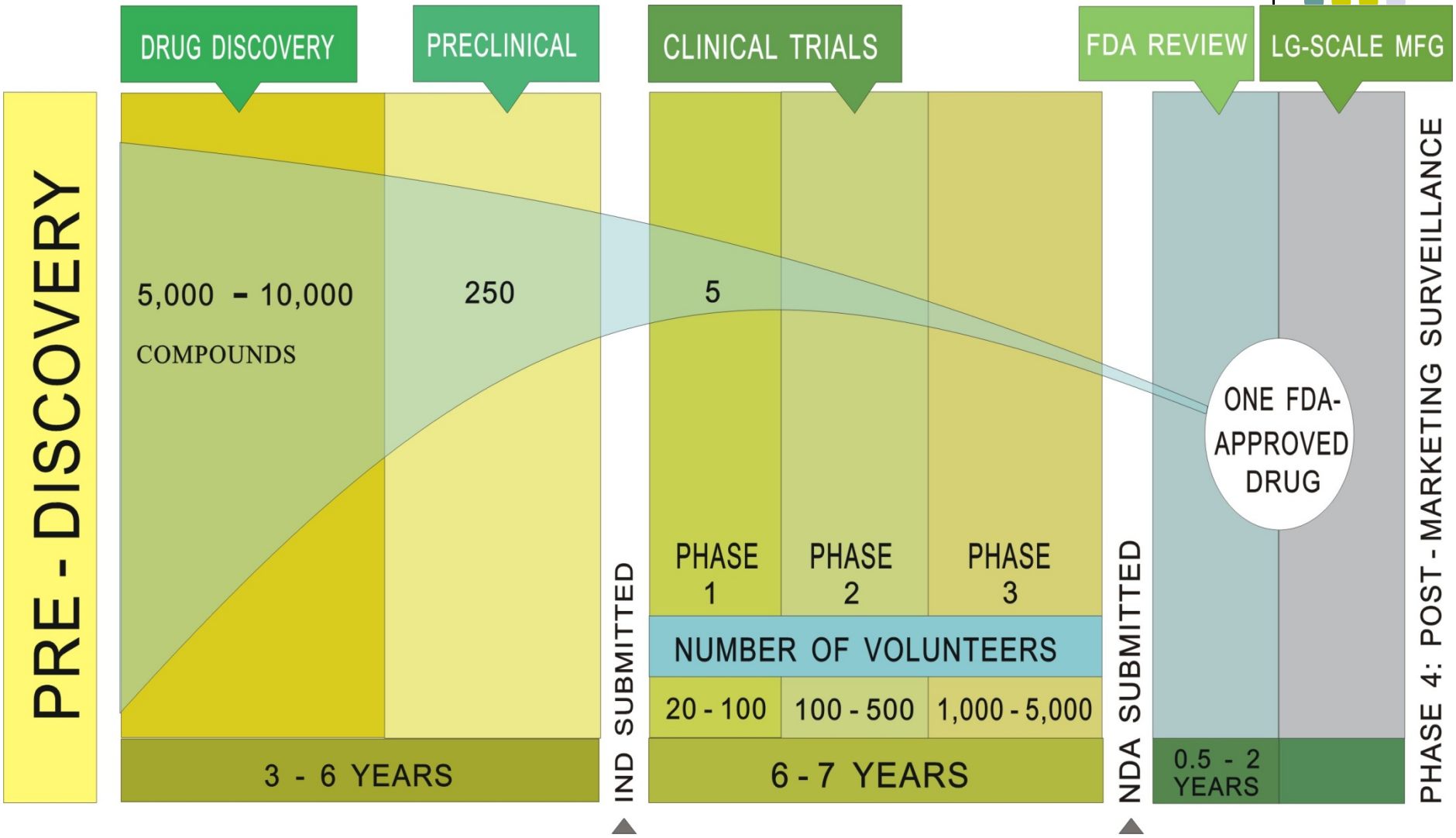
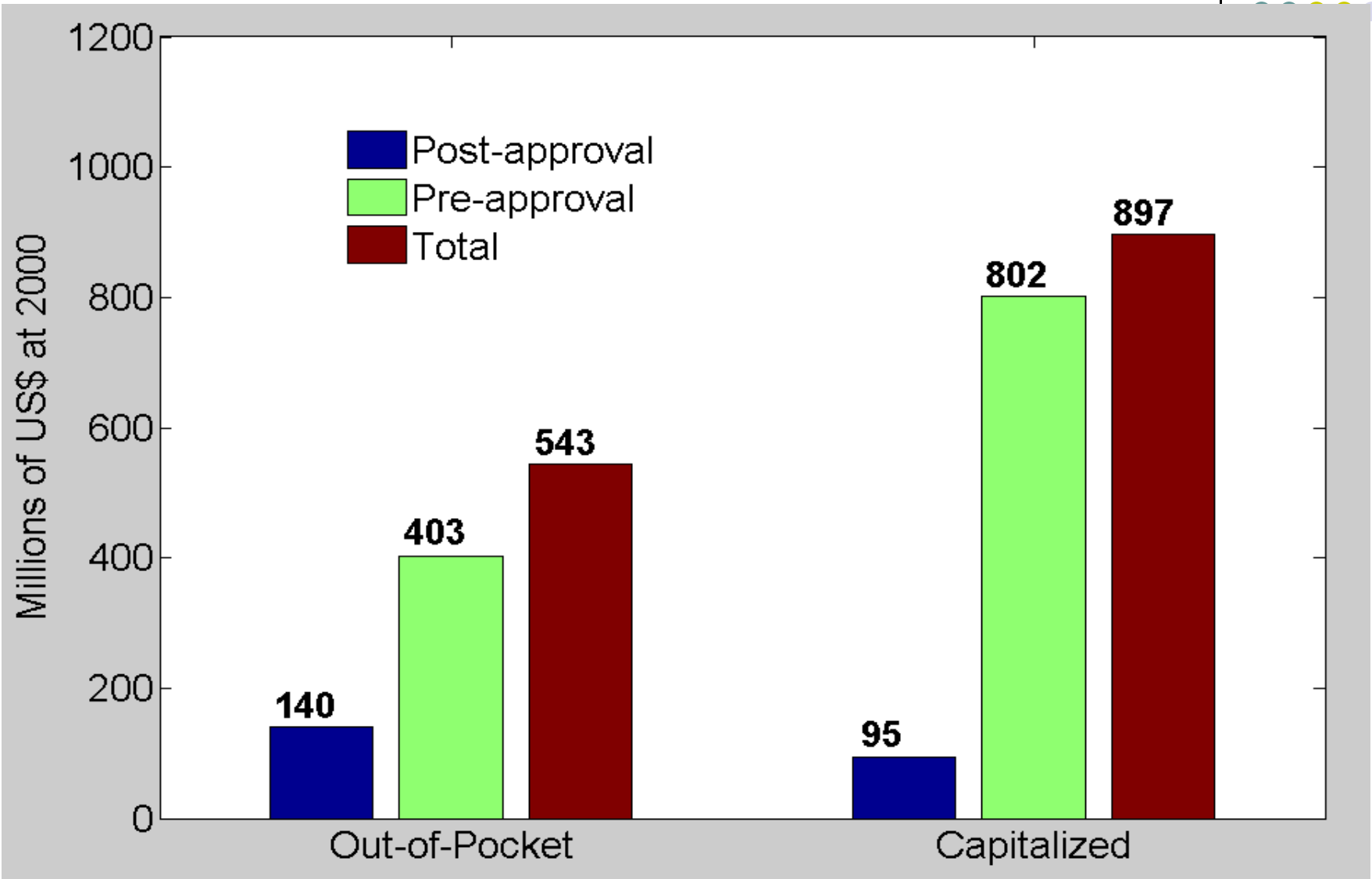


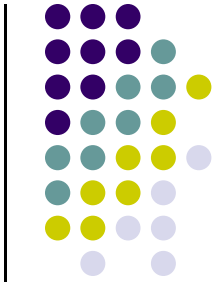
Figure adopted from the brochure of INNOVATION.ORG "*Drug Discovery and Development: Understanding the R&D Process*".

Post-approval R&D Cost



Data source: DiMasi et al., J Health Economics 2003;22(2):151-185

Post-approval Clinical Trial: Phase IV



- Research on a new medicine continues even after approval.

- As a much larger number of patients begin to use the drug, companies must continue to monitor it carefully and submit periodic reports, including cases of adverse events, to the FDA.

- In addition, phase V clinical trials can be set up to evaluate long-term safety or how the new medicine affects a specific subgroup of patients.

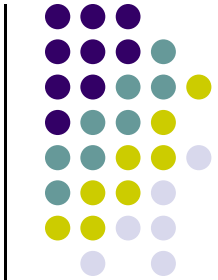
Post-approval Clinical Trial: Phase IV

- Some drugs have been withdrawn from the market because of risks to the patients, and unexpected adverse effects were not detected during Phase III clinical trials and were only apparent from the wider patient community.

Drug	*Time Withdrawn*	*Risk/Reason of Being Withdrawn*
Thioridazine	2005, U.K.	cardiotoxicity
Pemoline	2005, U.S.	hepatotoxicity
Natalizumab	2005, U.S.	Progressive multifocal leukoencephalopathy (PML). Returned to market on July, 2006
Ximelagatran	2006	hepatotoxicity (liver damage).
Pergolide	2007, U.S.	heart valve damage. Still available elsewhere.
Tegaserod	2007	imbalance of cardiovascular ischemic events, including heart attack and stroke.
Aprotinin	2007	increased risk of complications or death; permanently withdrawn except for research use
Inhaled insulin	2007, U.K.	national restrictions on prescribing, doubts over long term safety and too high a cost
Lumiracoxib	2007-2008	serious side effects, mainly liver damage
Rimonabant	2008	severe depression and suicide
Efalizumab	2009	increased risk of progressive multifocal leukoencephalopathy
Sibutramine	2010, Europe	increased cardiovascular risk. This drug continues to be available in the U.S.
Gemtuzumab ozogamicin	2010, U.S.	increased risks of veno-occlusive disease and no benefit in acute myeloid leukemia (AML)
Rosiglitazone	2010, Europe	increased risk of heart attacks and death. This drug continues to be available in the U.S.

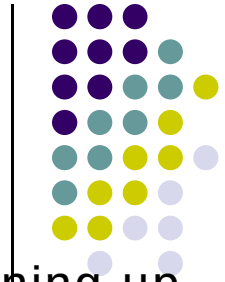
Partly cited from Wiki: http://en.wikipedia.org/wiki/List_of_withdrawn_drugs

A Recent Case



- Pfizer Prepares for Voluntary Withdrawal of U.S. New Drug Application and for Discontinuation of Commercial Availability of Mylotarg.
- “After extensive discussions with the FDA, Pfizer has decided to withdraw the NDA effective October 15, 2010.”
  - Press release from Pfizer
  - http://www.pfizer.com/home/

Concluding Remarks



- Each success is built on many, many prior failures.
- Advances in understanding human biology and diseases are opening up exciting new possibilities for breakthrough medicines.
- Researchers face great challenges in understanding and applying these advances to the treatment of diseases.

Duration	**Stage**
Various	**Pre-discovery** *Goal*: Understand the disease and choose a target molecules. *How*: Scientists in pharmaceutical research companies, government, academic and for-profit research institutions contribute to basic research.
3~6 years	**Discovery** *Goal*: Find a drug candidate. *How*: Create a new molecule or select an existing molecules as the starting point. Perform tests on that molecule and then optimize (change its structure) it to make it work better
	Preclinical *Goal*: Test extensively to determine if the drug is safe enough for human testing. *How*: Researchers test the safety and effectiveness in the lab and in animal models.

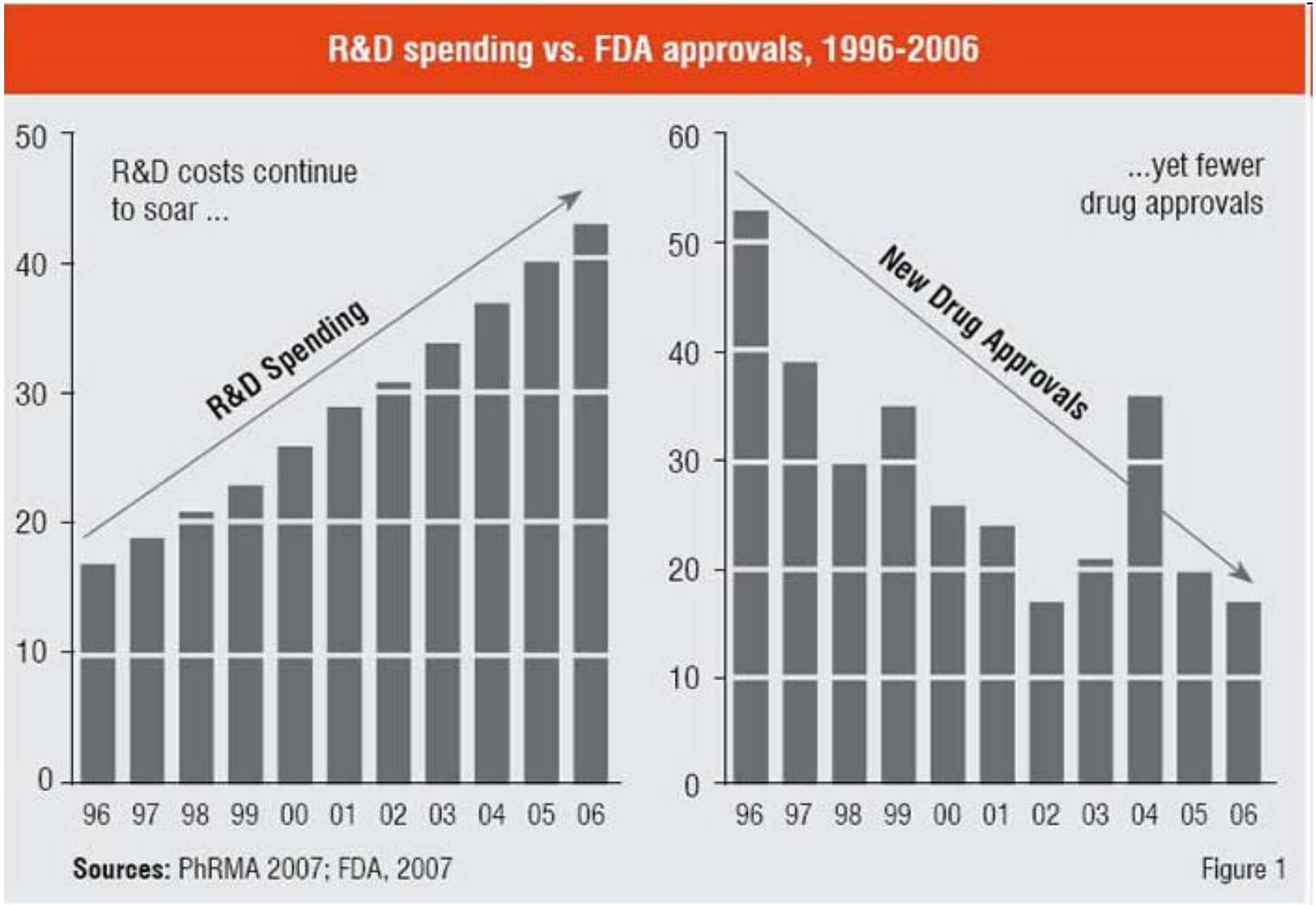
Table adopted from the online brochure from INNOVATION.ORG: "Drug Discovery and Development" 2007

Concluding Remarks

6~7 years	**IND** *Goal*: Obtain FDA approval to test the drug in humans. *How*: FDA reviews all preclinical testing and plans for clinical testing to determine if the drug is safe enough to move to human trials.
	Clinical Trials *Goal*: Test in humans to determine if the drug is safe and effective. *How*: Candidate drug is tested in clinical setting in three phases of trials, beginning with tests in a small group of healthy volunteers and moving into larger groups of patients.
0.5 ~ 2 years	**Review** *Goal*: FDA reviews all results to determine if the drug can be approved for patients to use. *How*: The FDA reviews hundreds of thousands of pages of information, including all clinical and preclinical findings, proposed labeling and manufacturing plans. They may solicit the opinion of an independent advisory committee.
	Manufacturing *Goal*: Formulation, scale up and production of the new medicine
	Ongoing Studies *Goal*: Monitor the drug as it is used in the larger population to catch any unexpected serious side effects.
	Total *How much*: $800 million - $1 billion *How long*: 10 – 15 years

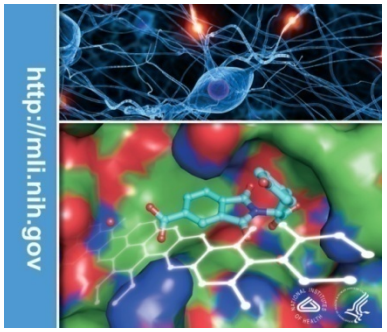
Table adopted from the online brochure from INNOVATION.ORG: "Drug Discovery and Development" 2007

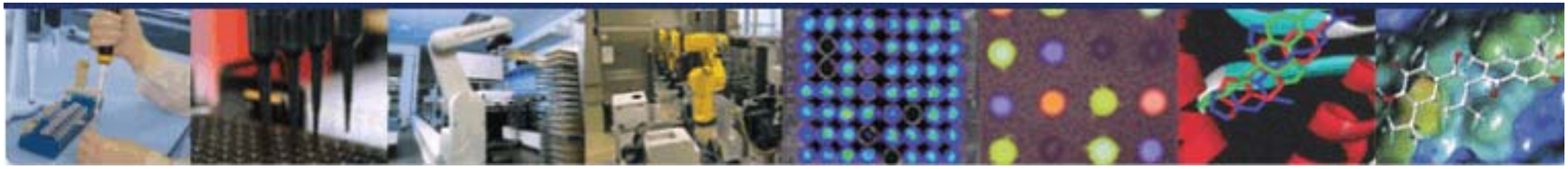
Bottleneck in Drug Discovery



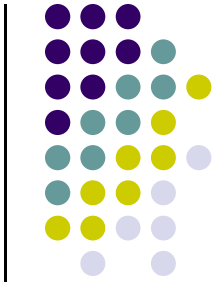
Part II: Drug Discovery Related Programs in the Public Sectors

- NIH Roadmap Molecular Libraries and Imaging project aims to profile millions of chemicals and their interactions with biological systems each year.
- EPA routinely performs testing of chemicals and evaluate their toxicities.
- Large pharmaceutical companies screening and profiling millions of chemicals each year
- FDA is investigating new technology for evaluating the interactions between chemicals and biological systems
- Results are freely available in the PubChem database.



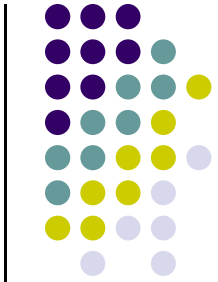


Molecular Probe Discovery



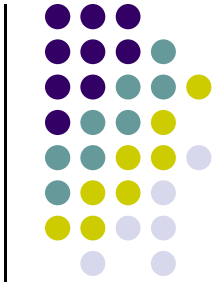
- NIH Roadmap Molecular Libraries and Chemical Probes Program
  - A research program designed to develop small organic molecules that can be used as chemical probes to study the functions of genes, cells & biochemical pathways,
  - Goal: providing new ways to explore the functions of major components of cells in the functions of major components of cells in health & disease

MLPCN



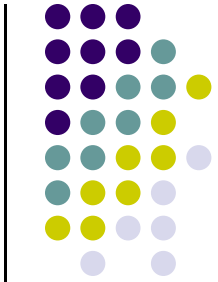
- US National Institute of Health (NIH) Molecular Libraries Probe Production Centers Network MLPCN Program with 9 centers
  - **Comprehensive Centers:** Provide all three services: assay, cheminformatics/informatics, and medicinal chemistry within a single site. Broad, Burnham, NCGC, and Scripps are comprehensive centers.
  - **Specialized Screening Centers:** Handle specialized types of assays including handling assay informatics. Johns Hopkins, Southern Research Institute, and UNM are specialized screening centers.
  - **Specialized Chemistry Centers:** Focus on providing medicinal chemistry and cheminformatics support for performing structure-activity relationships that is typically needed to produce useful chemical probes from screening hits. These are located at Kansas and Vanderbilt.
- $500M/6 years

Chemical Probe



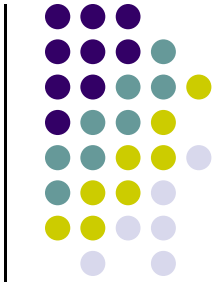
- A potent, selective, and cell-permeable small molecule that modulates a specific biochemical or cellular functions and provides a useful tool for biomedical and biological research.
- Comparing to gene knock-out/in techniques and RNAi techniques, small molecule probes can target a specific site of a cell's chemical machinery, thus provides information on a specific step in a network of cell functions.

Probe & Drug?



- Ideal Probe? (S. Frye, NCB, pp. 159-162, March 2010)
  - Target selectivity: paralogs, orthologs, genes in the same pathway, genes important for pharmacodyhnamics
  - Connection between the cellular phenotype and the molecular mechanism: pharmacology,
  - Toxicity and stability
  - Availability and synthesis feasibility

Probe & Drug?

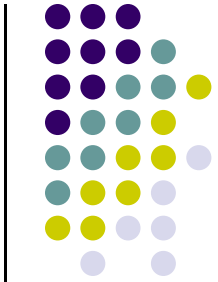


- Not as top priorities:
  - Oral bioavailability
  - Tendency to be metabolized
  - Half-time
  - Cost of manufacture
  - …

PubChem Web Portal

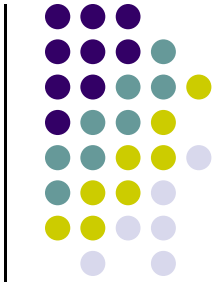
- All screening and compound data from the MLI phases are freely available to the public via a web portal called PubChem
  - Annotated information about the bioactivities of small molecules
  - Chemical structures and compound probe information
  - A fast chemical structure similarity search tool.

Exploratory Analysis



- Exploring the utility of MLPCN data (screening results, target proteins, and small molecules) in the future therapeutic exploration
- Comparison and analysis of MLPCN targets and drug targets
  - Novelty of MLPCN targets
  - MLPCN targets are a promising source for new drug targets
- MLPCN screening compounds *vs.* approved drugs, metabolites, and natural products
  - Increase its drug-likeness and biogenic bias

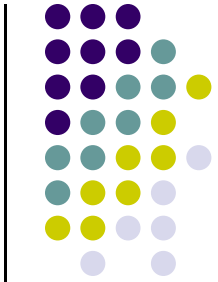
Fact Sheets (as of Jan 2009)



Total Number of Bioassays	**1,306**
Number of Target-based Bioassays	**672**
Number of Cell-based Bioassays	**634**
Number of Bioassays with Active Compounds	**1,126**
Number of Active Compounds in all assays	**151,930**
Number of Bioassay-Compound Pairs	**555,859**
Number of Bioassay Pairs with at least one shared compounds	**124,442**

Zhang, Lushington, and Huan, Characterizing the Diversity and Biological Relevance of the MLPCN Assay Manifold and Screening Set, *Journal of Chemical Information and Modeling*, Vol. 51, No. 6, pp. 1205-1215, 2011

Some Terminologies

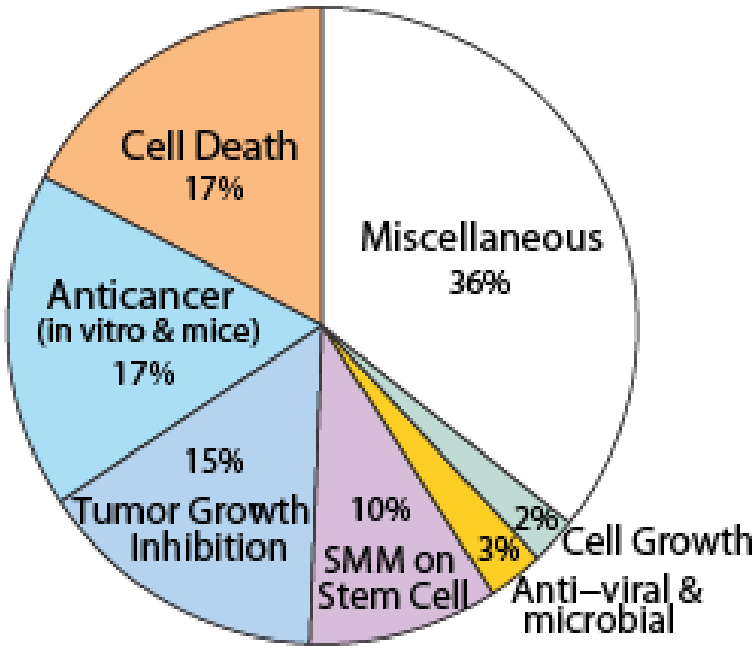


- MLPCN targets: 200 distinct protein extracted from 680 target-based bioassay from MLPCN screening

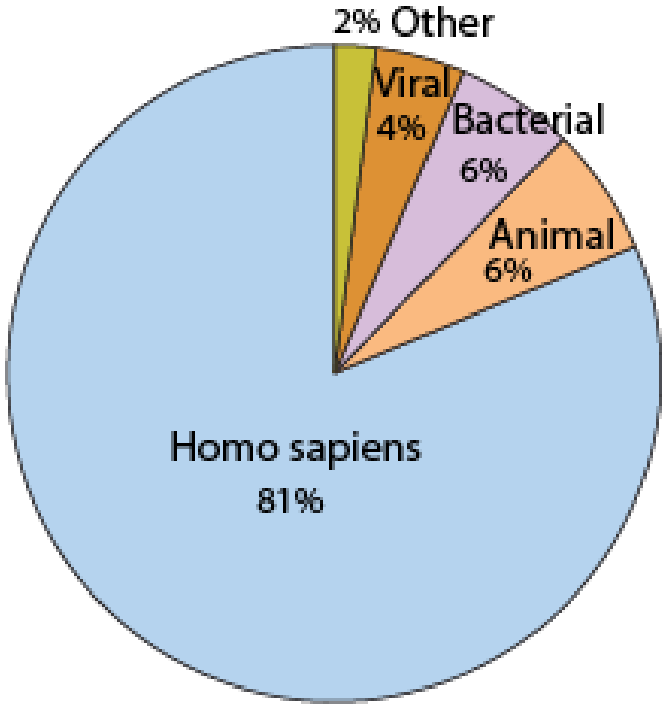
- MLPCN screening set: a compound set collected from 23 bioassays deposited between May 1 – July 22, 2009
  - A compound is selected if it was tested in 21 of the 23 assays (i.e. 90%)
  - 279,768 compounds obtained

- Random ChemNavigator set: 279,768 compounds randomly extracted from the ChemNavigator compound collections
  - ChemNavigator: a library of commercially available small molecules

PubChem BioAssays



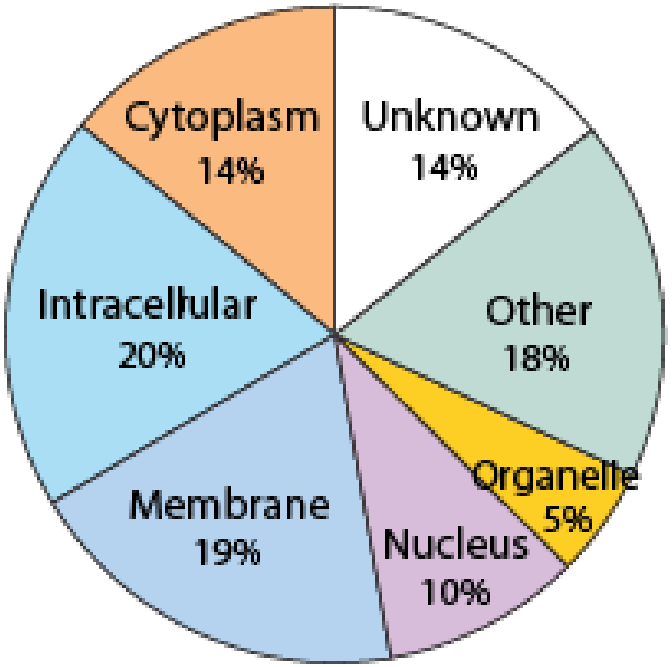
Purposes of Cell-based Assays



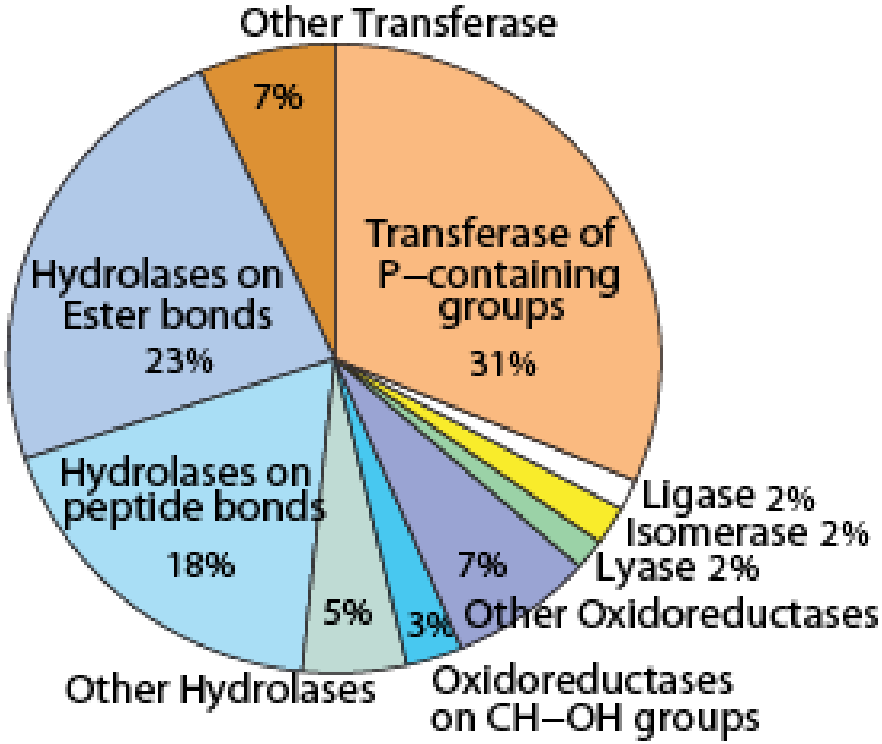
Organisms of MLPCN targets

MLPCN Targets

- 289 target proteins are extracted from 680 target-based bioassays
- 200 distinct proteins are obtained from converting 215 gene symbols
- 113 MLPCN targets are identified as enzymes

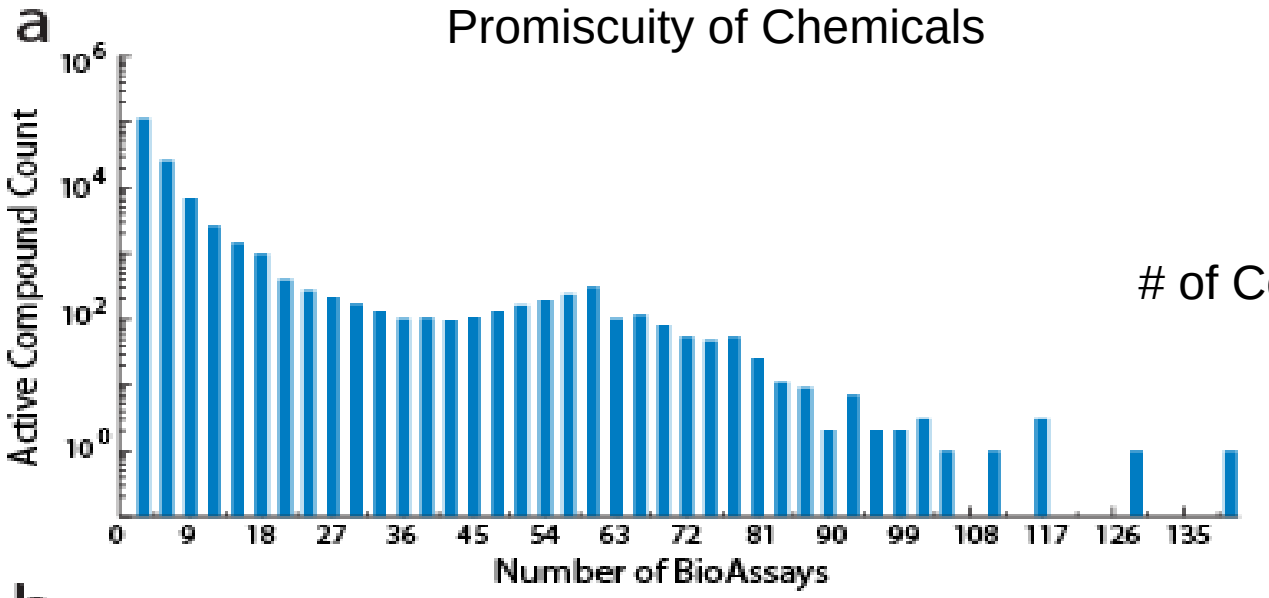


Subcellular locations

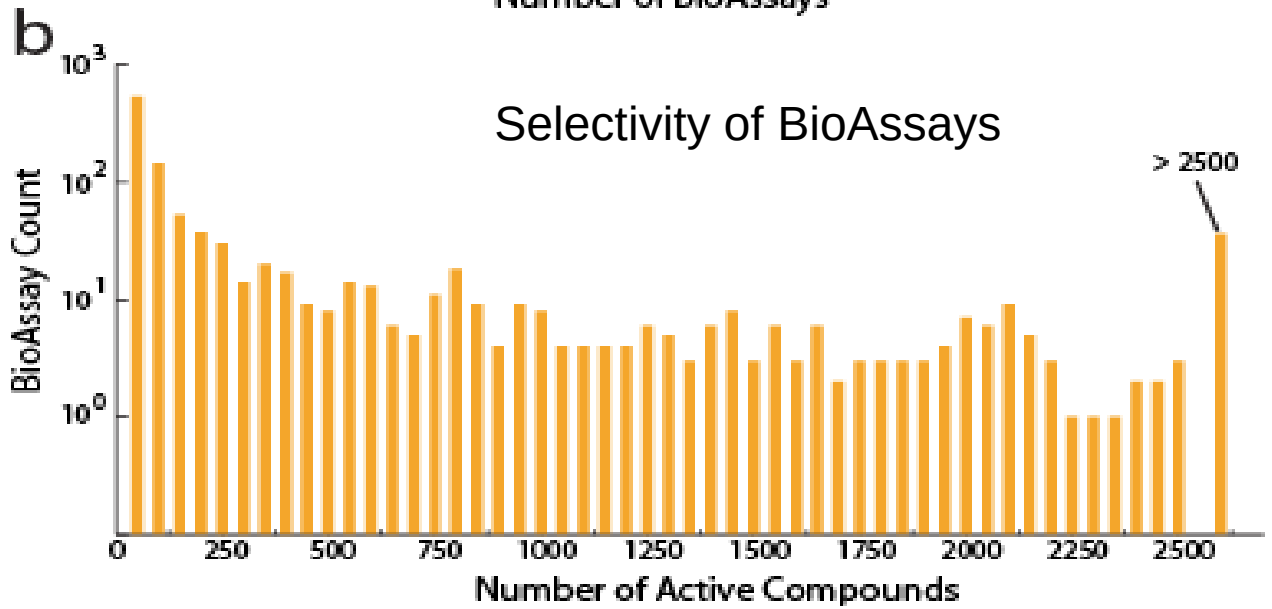


Cellular functions of 113 targets

Statistics of PubChem BioAssays



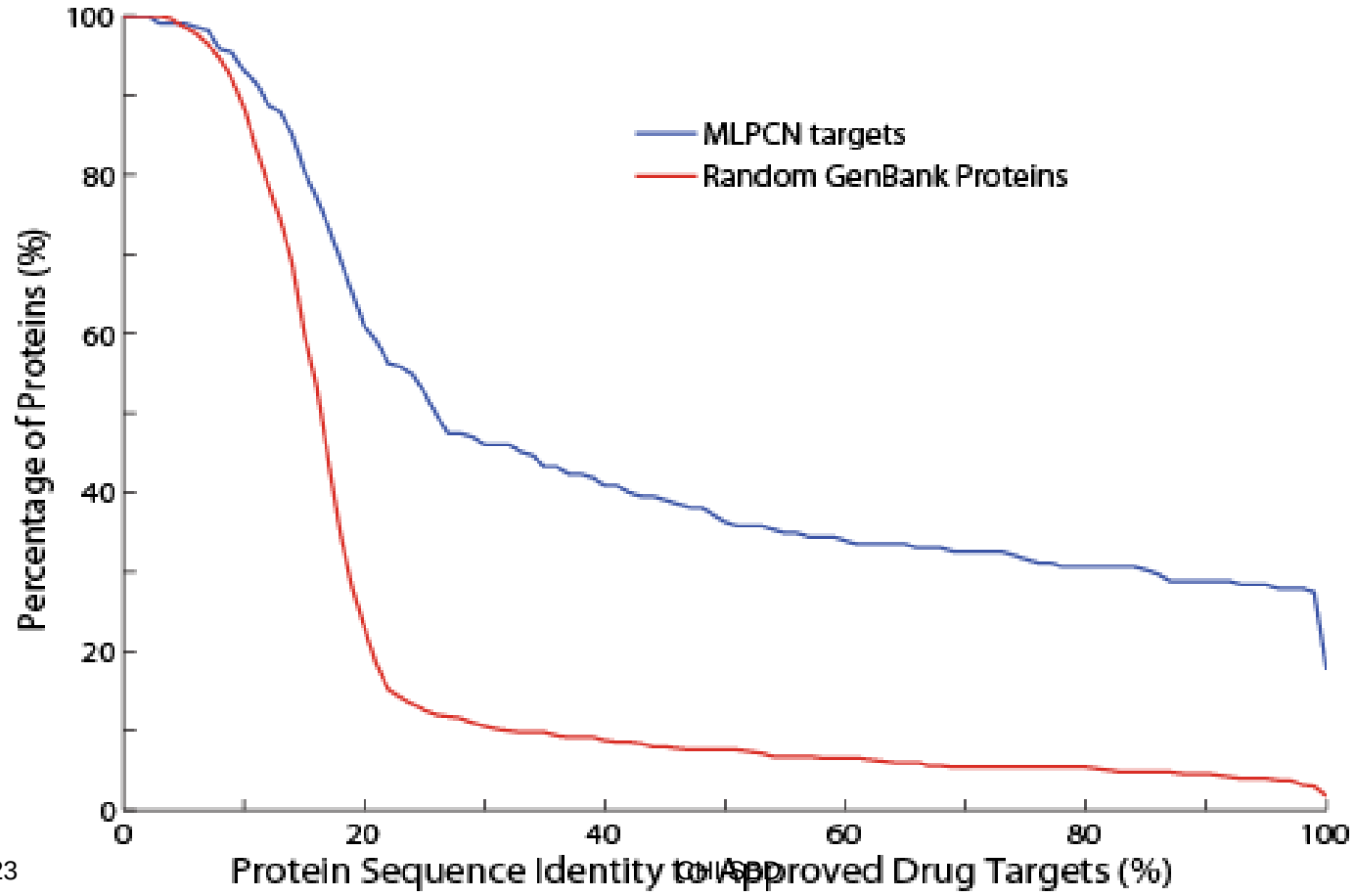
# of Compounds Active in X BioAssays



# of BioAssays with X Active Compounds

MLPCN Targets and Drug Targets

- Needleman-Wunsch global alignment (gap open =11, extension = 1) between MLPCN targets and drug targets
- 500 human proteins randomly selected from GenBank as control set

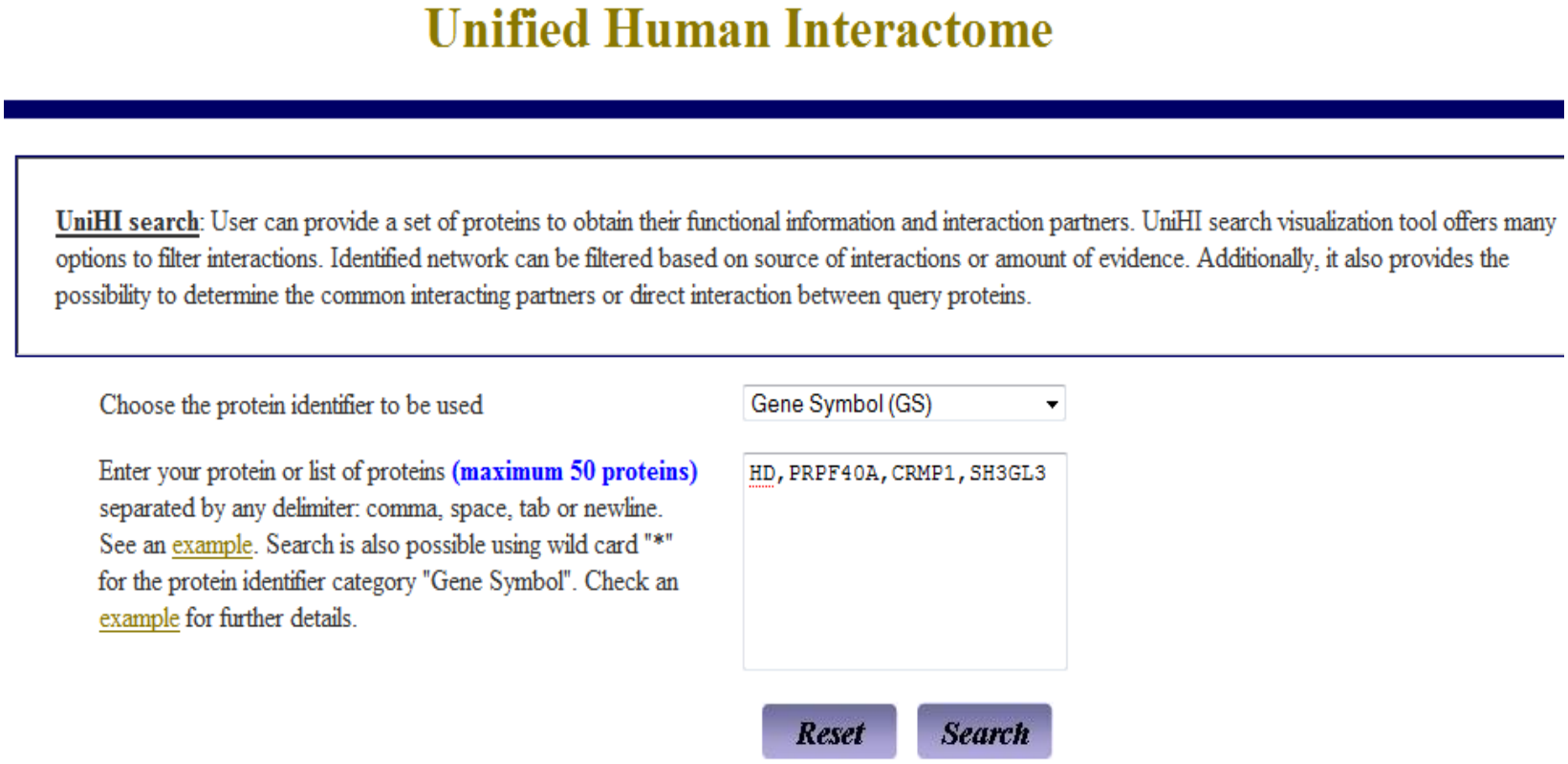


MLPCN Targets and Drug Targets

- A MLPCN target is defined as “similar” to drug targets if its sequence identity to at least one drug target is >= 30%
- MLPCN target weighted counts vs. total number of targets in PubChem

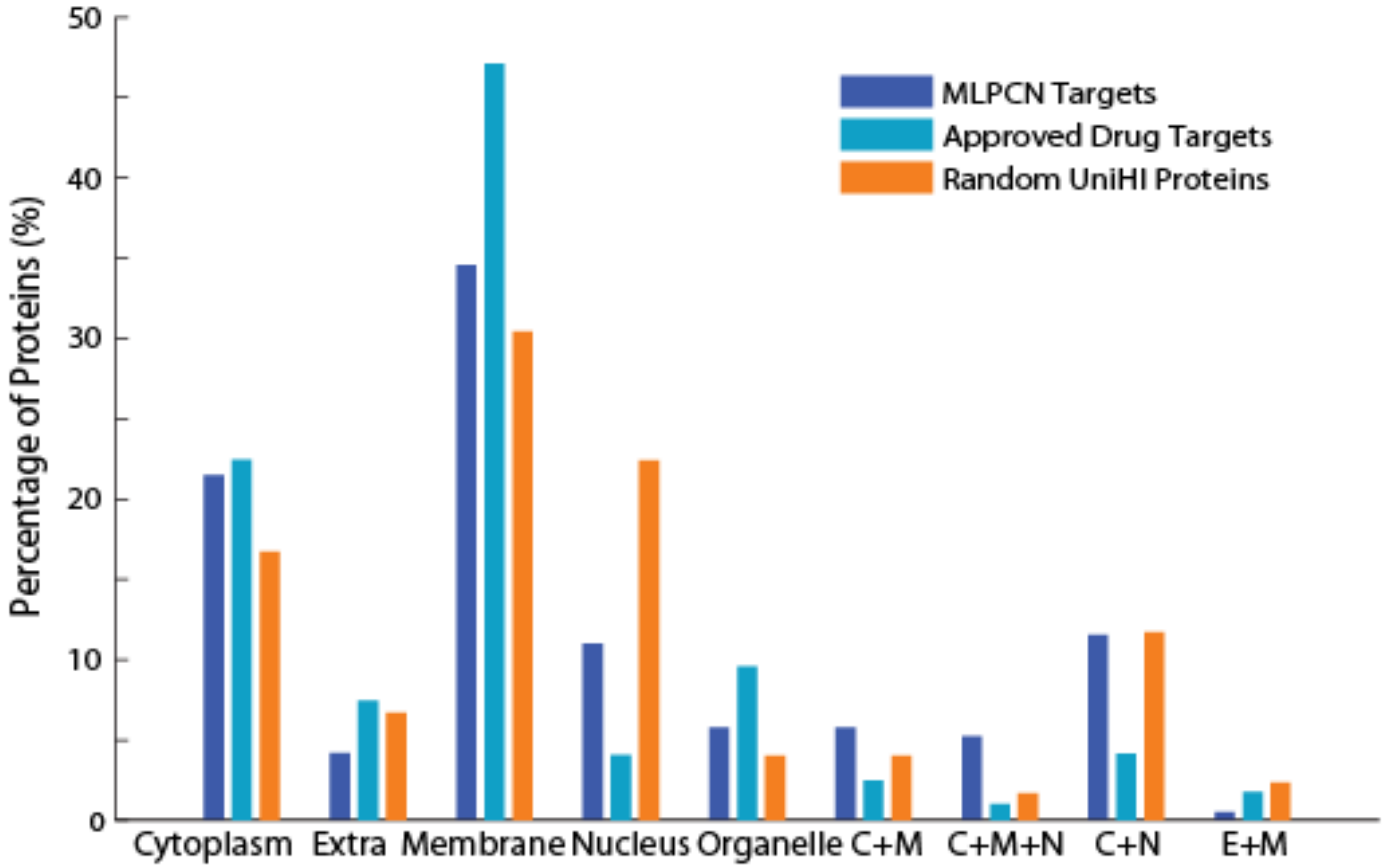
UniHI: Human Protein-Protein Interaction Network

- UniHI is a unified human PPI network containing over 250,000 human PPIs collected from 14 major PPI sources with careful data integration and literature curation.
- One of the largest human PPI networks, with various confidence scoring systems for each PPI



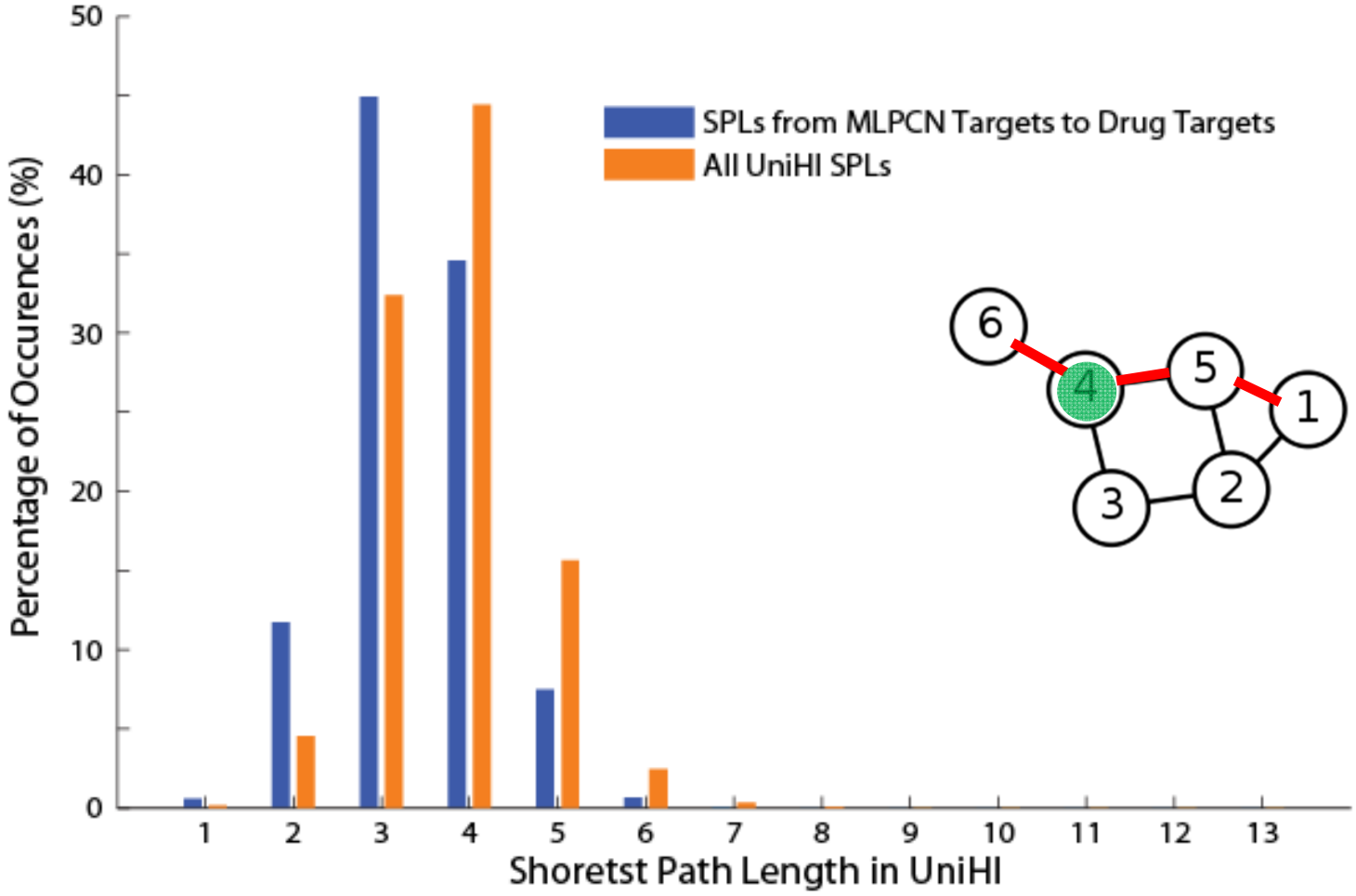
Subcellular location of MLPCN and Drug Targets

- Subcellular location of 182 MLPCN targets, and 1035 drug targets from NCBI Entrez Gene and Gene Ontology databases
  - 347 random human proteins from UniHI as control set



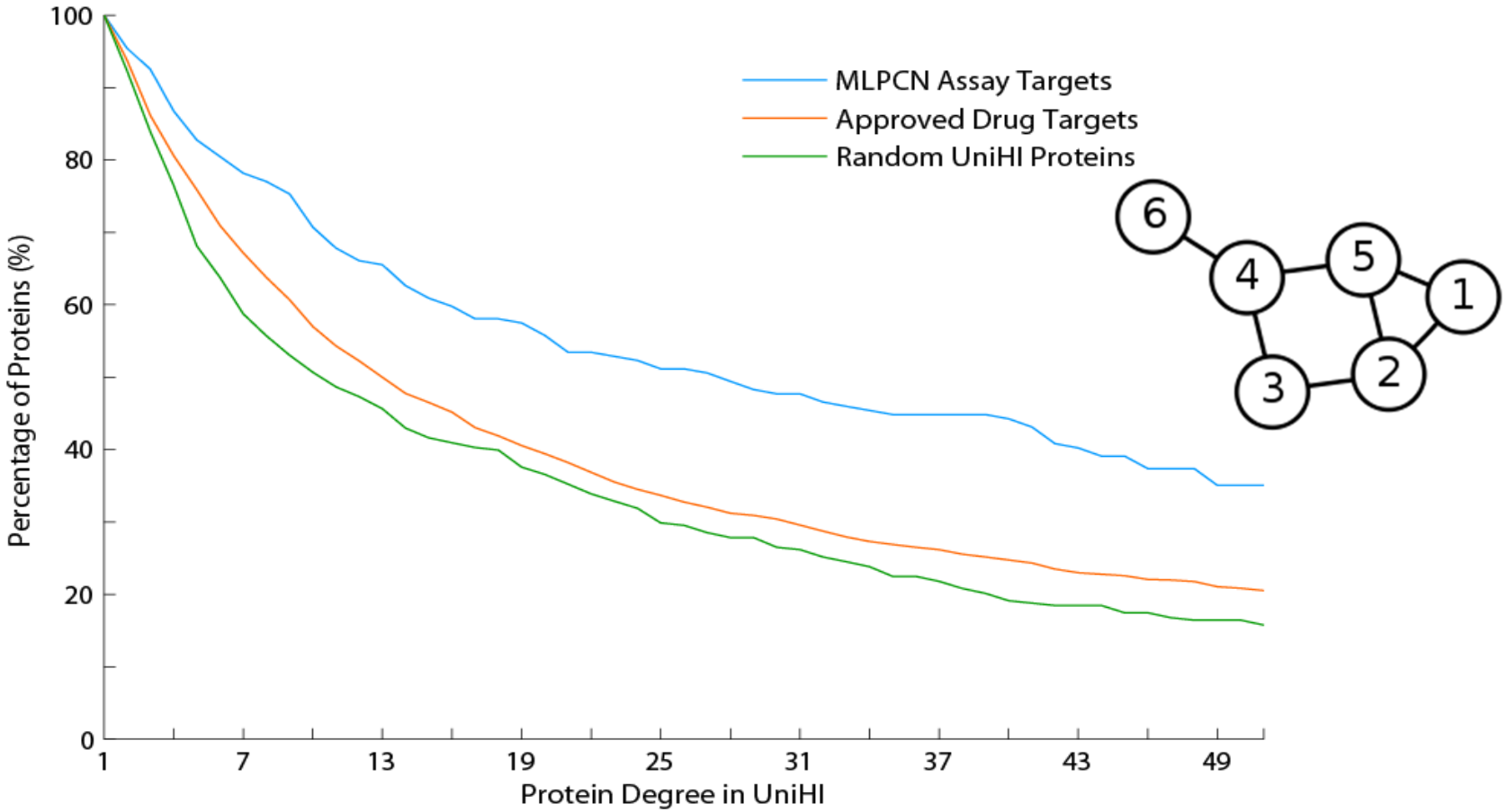
Distribution of Shortest Path Lengths

- Shortest Path Length: the smallest number of PPIs between any MLPCN target to any drug target in the UniHI network (graph)
  - Control set: shortest path length between any two proteins in UniHI

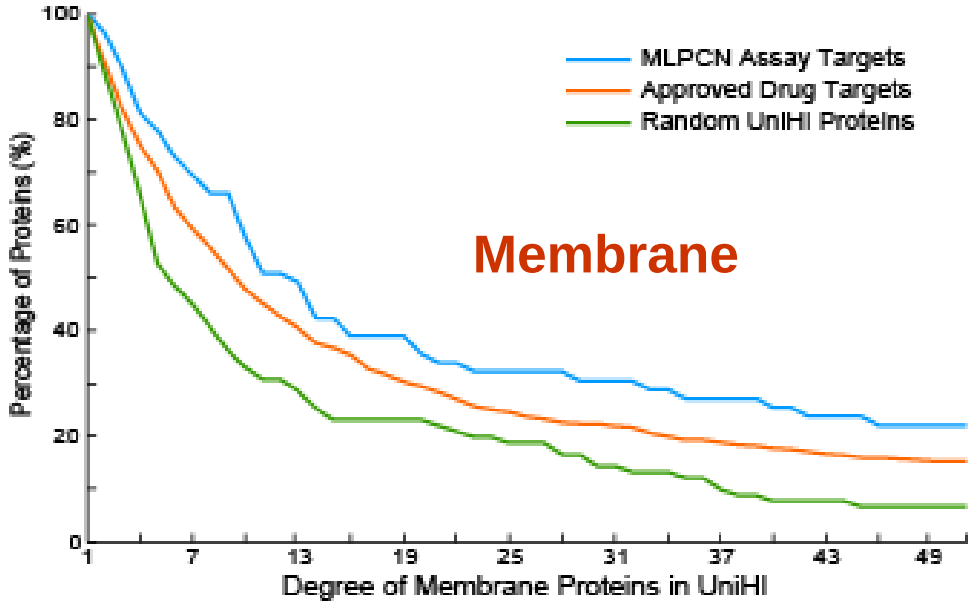


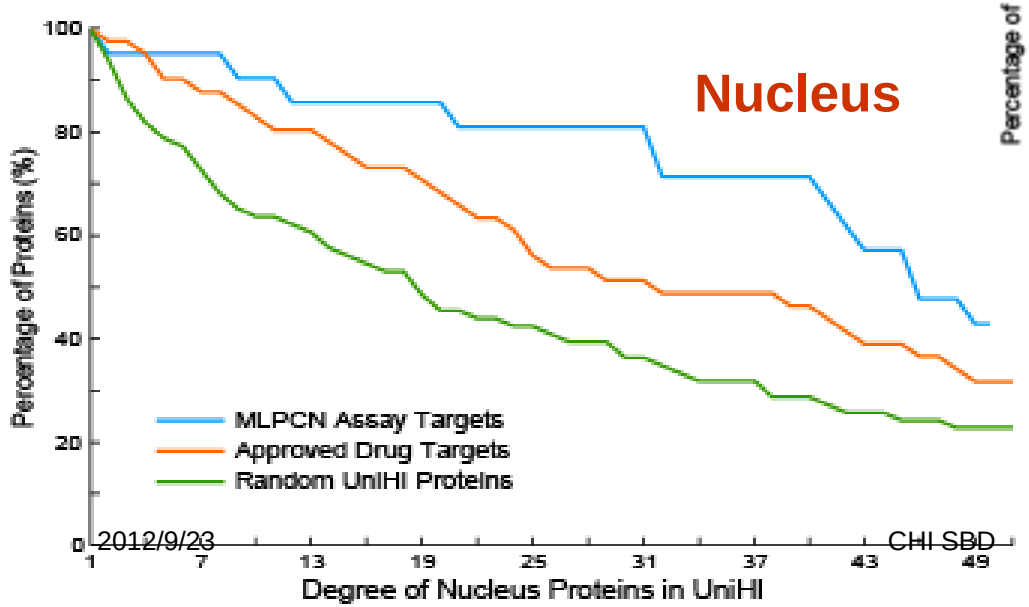
Degree Distribution in UniHI

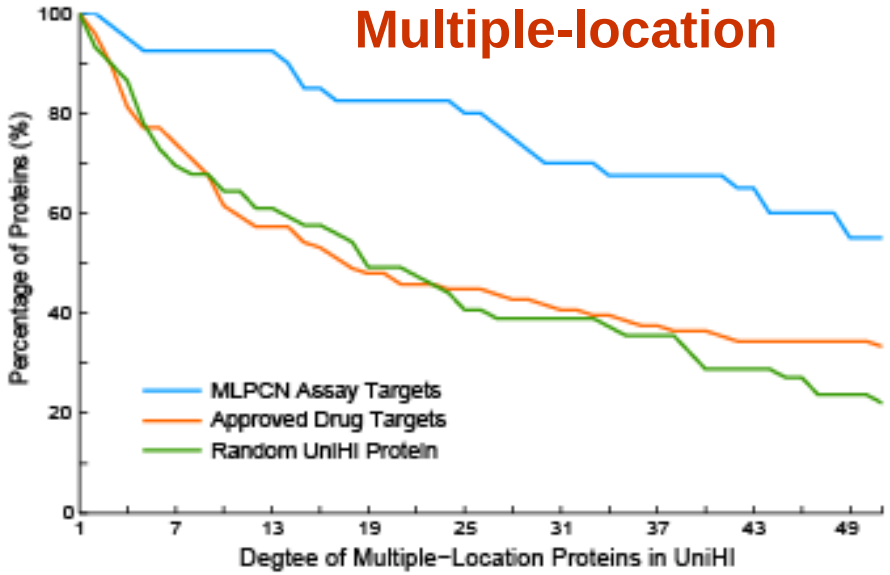
- Network degree: the number of interacting proteins of a given protein in the UniHI network (graph)
- Control: 347 random human proteins in UniHI



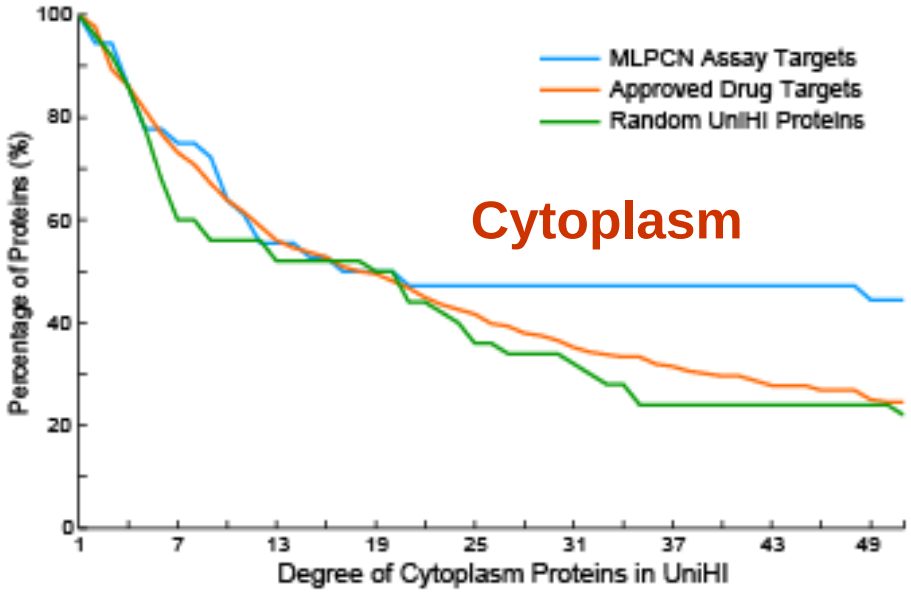
Degree Distribution and Subcellular Localization

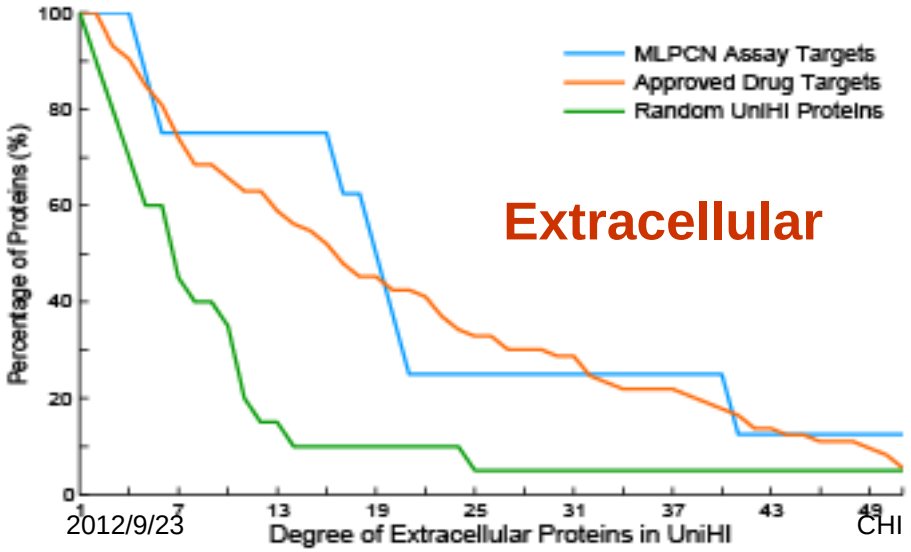


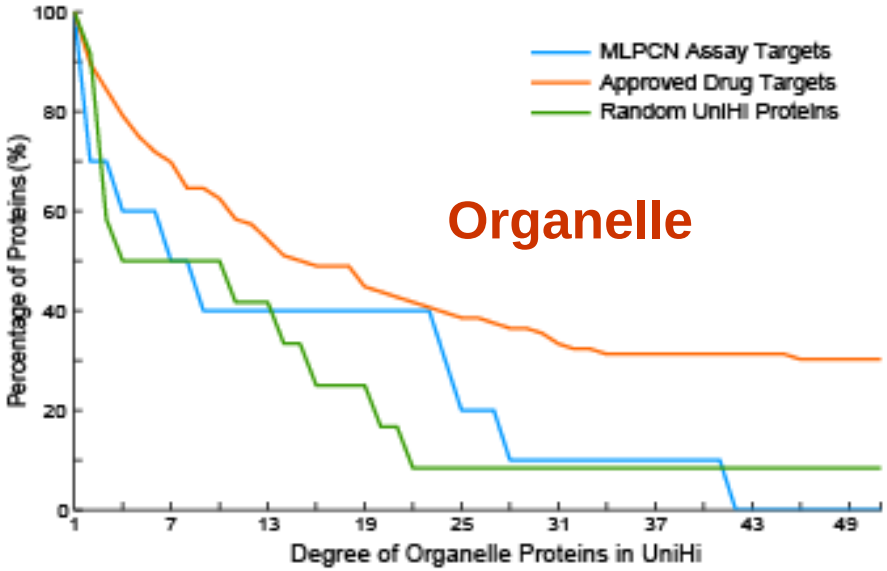




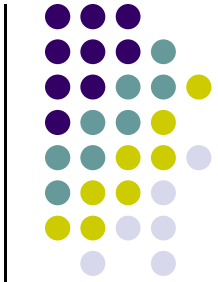
Degree Distribution and Subcellular Localization





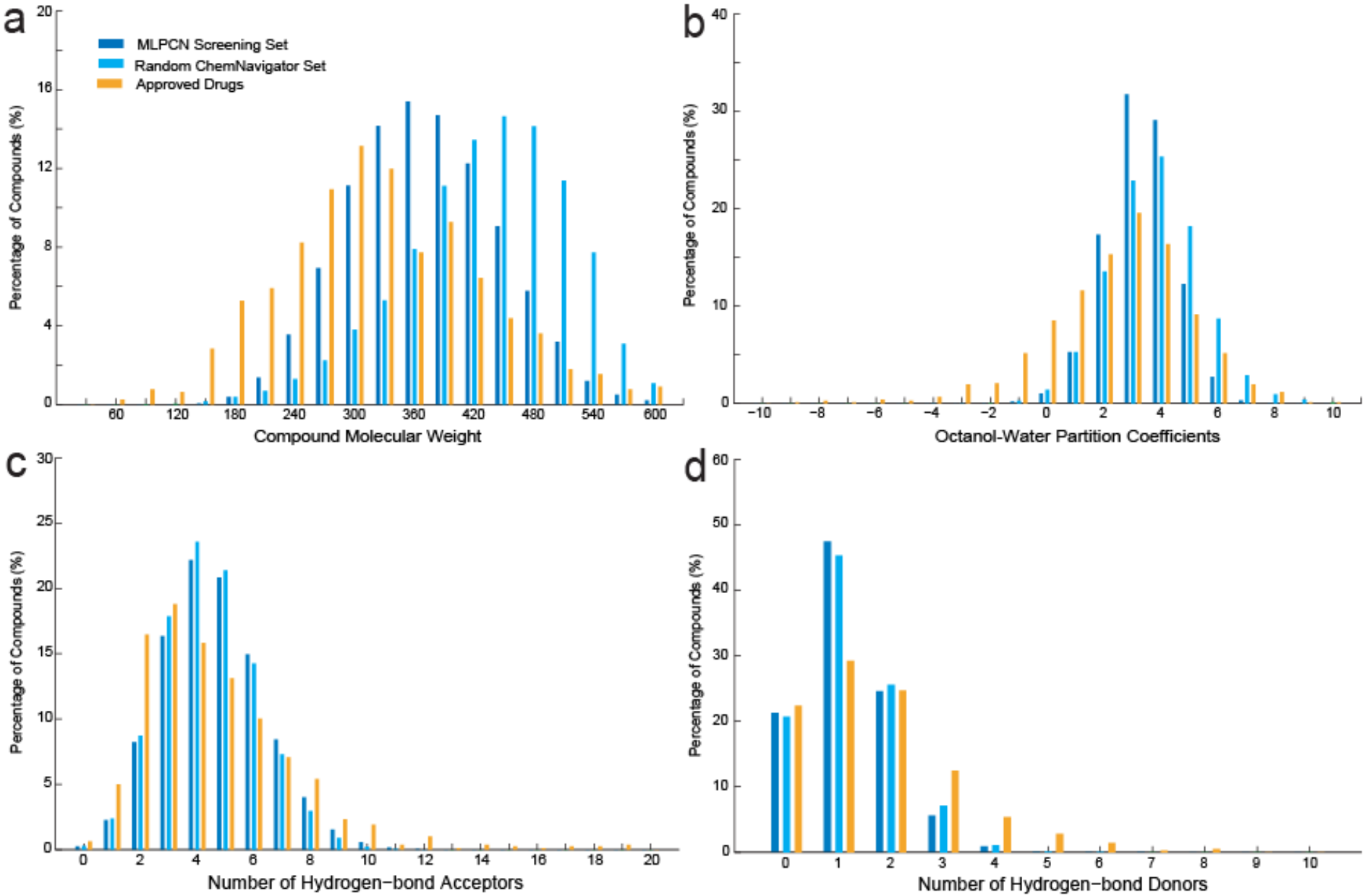


A Few Remarks

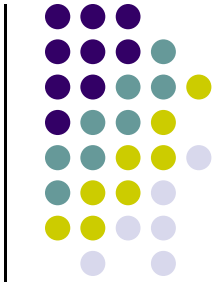


- Shortest path length analysis shows that the MLPCN targets are clustering around drug targets, and revealed that the MLPCN tends to sample pathways that have already been therapeutically targeted.
- This significant difference in median degrees of MLPCN and known drug targets implied that MLPCN targets are somewhat distinct relative to current drug targets, and thus may theoretically afford novel avenues for eventual therapeutics development.

Drug Likeness of MLPCN Compounds



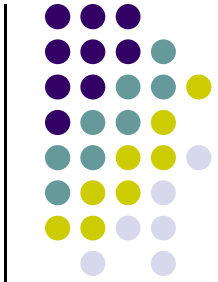
Compound Diversity Analysis



- When designing an optimal screening library for MLPCN bioassays, a crucial step is to assess its chemical space coverage, structural novelty, pharmaceutical and biological relevance compared to other important compound collections

- Characterizing the chemical space defined by a compound set
  - Extract features (descriptors) from each compound
  - Map each compound into an *N*-dimensional space consisting of *N* molecular structural features and properties
  - This descriptor set enabled us to compare how two compound sets distribute in the same chemical descriptor space.

Compound Diversity Analysis (2)

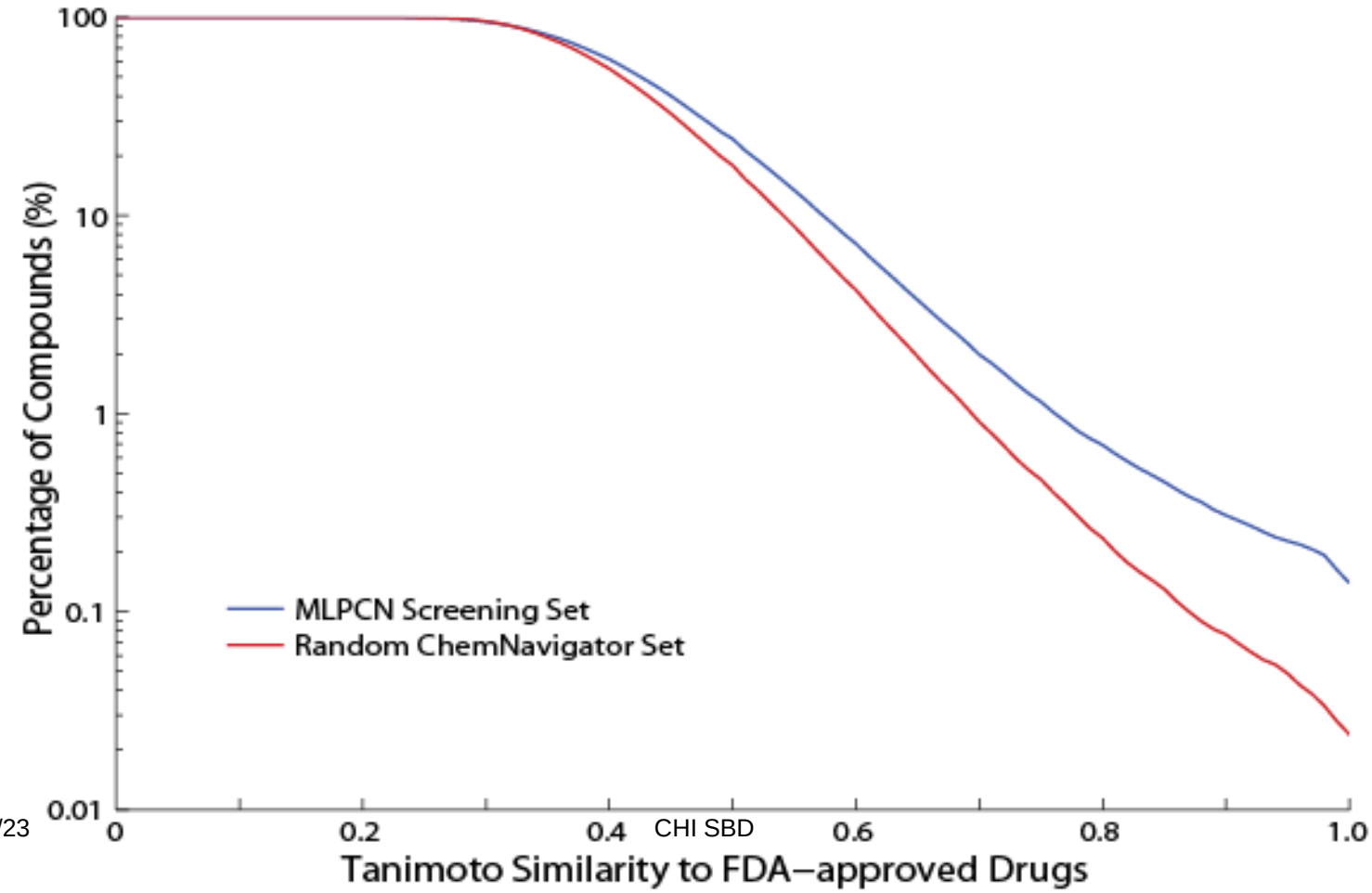


- Software: Tripos DiverseSolutions program
  - Calculate the BCUT descriptors for all sets of compounds,
  - Auto-select three descriptors to best define a 3D chemical space for the MLPCN screening set according to optimal compound dispersion across Cartesian space.
  - Use the first two descriptors to make a 2D chemical descriptor space, and map different sets of compounds into this space.
  - This MLPCN descriptor space was then partitioned into 600 equal bins in each axis (i.e. 360,000 cells).

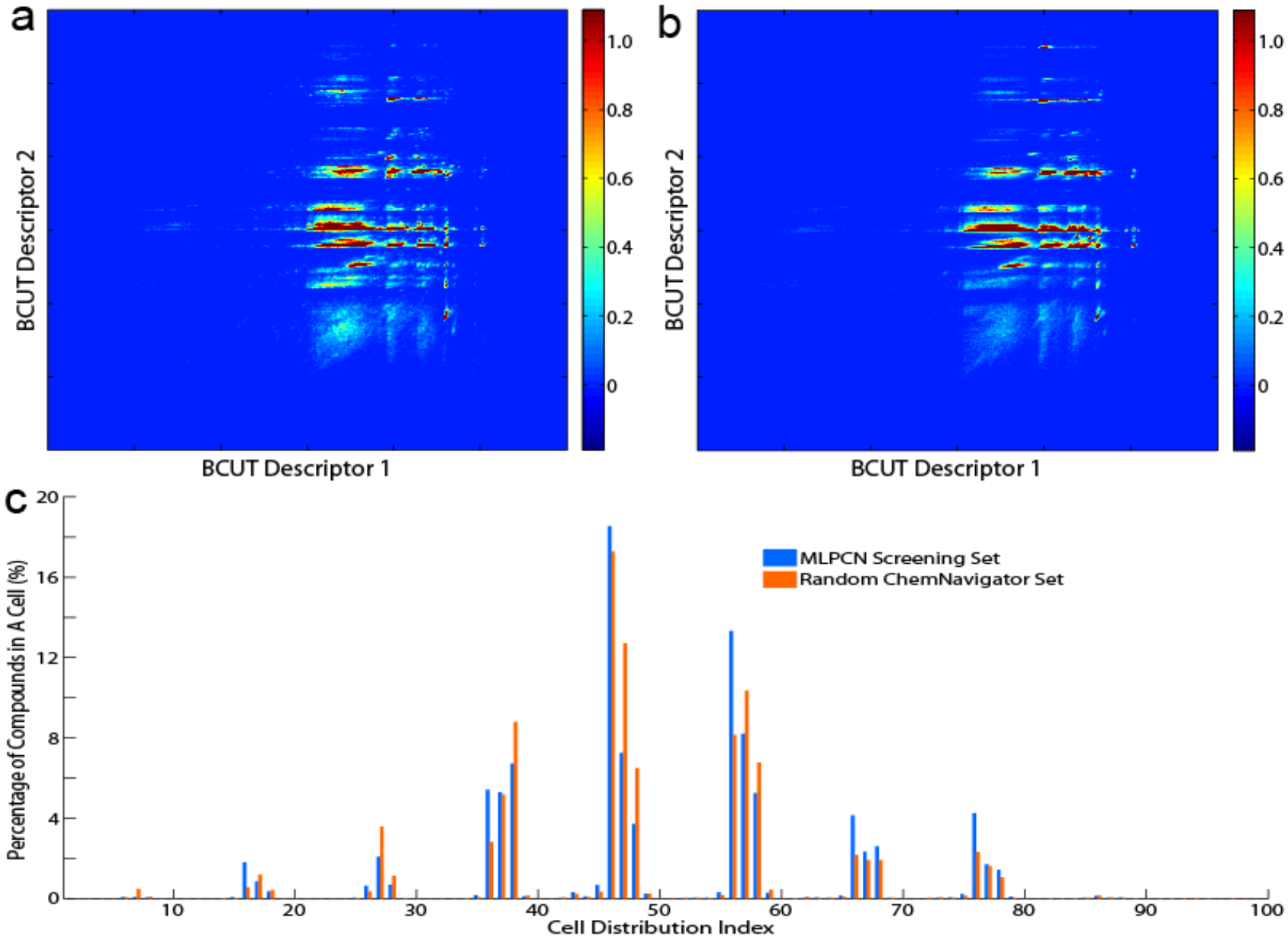
- BCUT Descriptors are obtained from the positive and negative eigenvalues of the adjacency matrix of a compound, weighting the diagonal elements with atom weights.

Drug Likeness of MLPCN Compounds

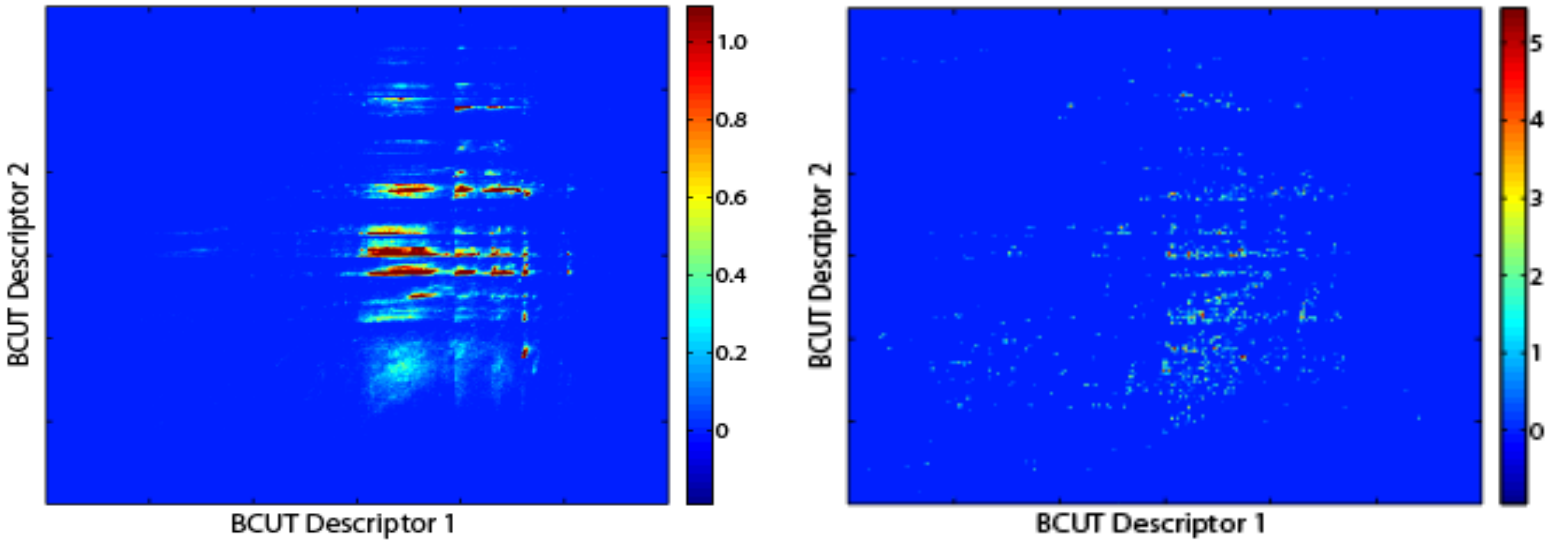
- For each compound in MLPCN or ChemNavigator, identify its nearest neighboring compounds in approved drugs (most similar)
- Compound similarity: Daylight fingerprint FP2 and Tanimoto coefficients

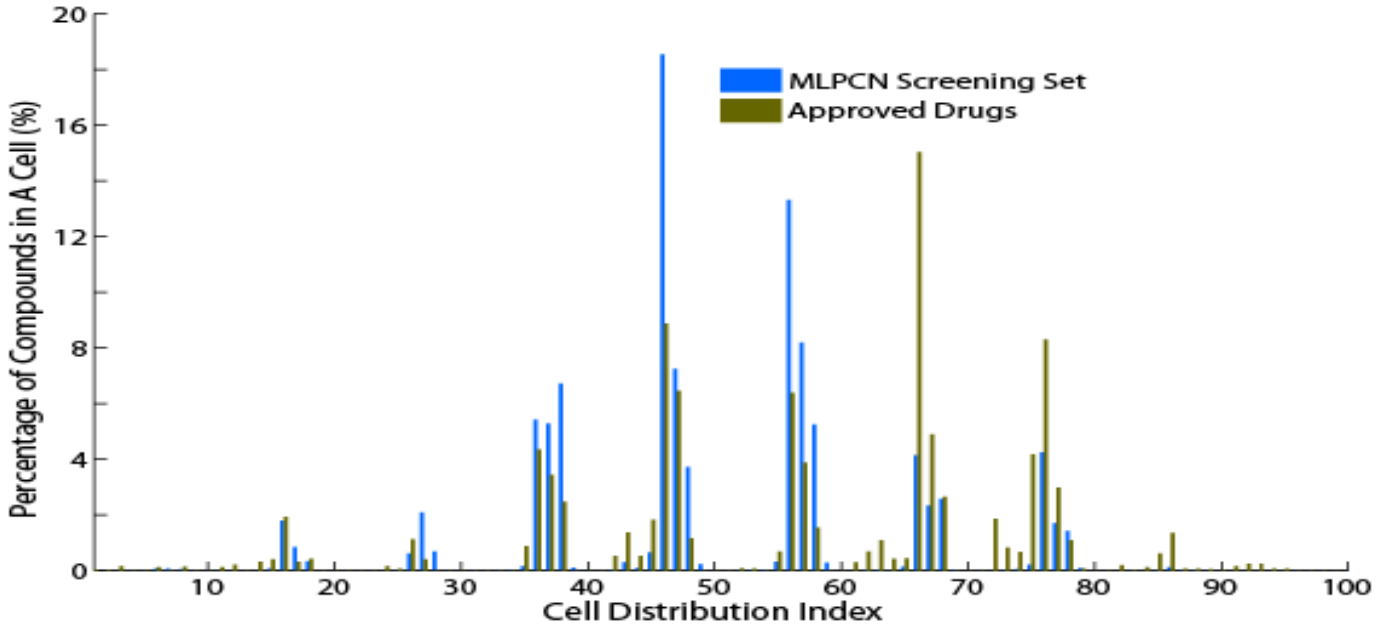


MLPCN Compounds *vs.* ChemNavigator Compounds

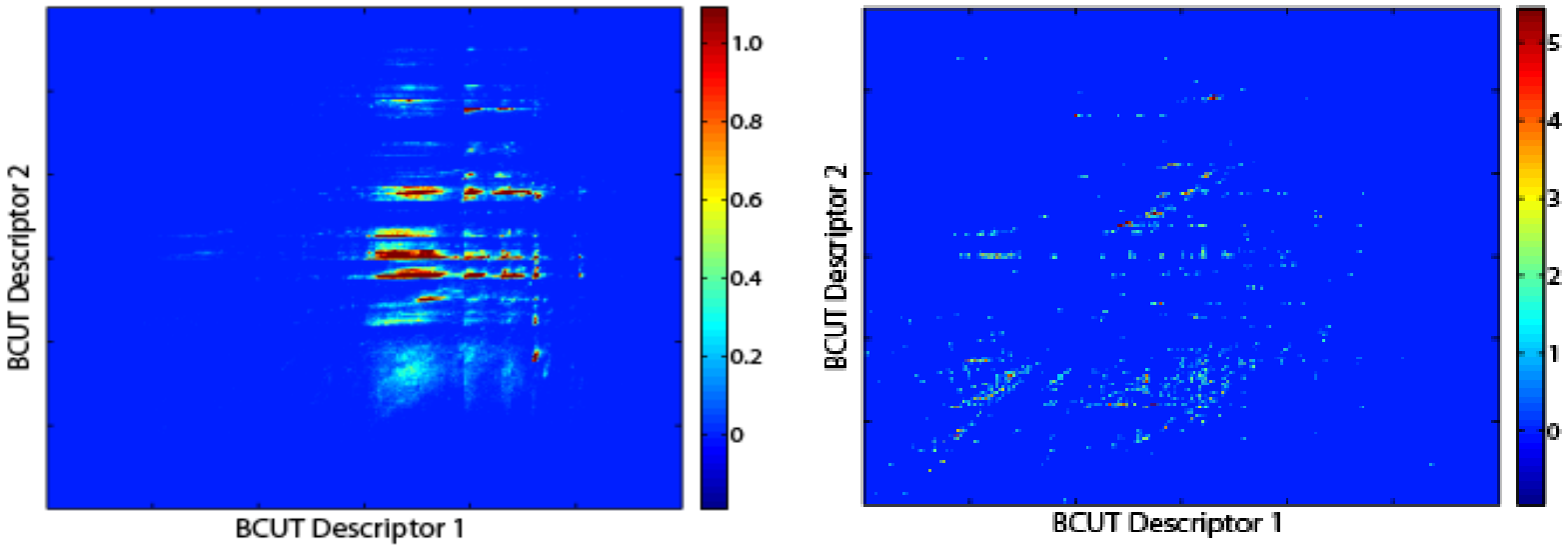


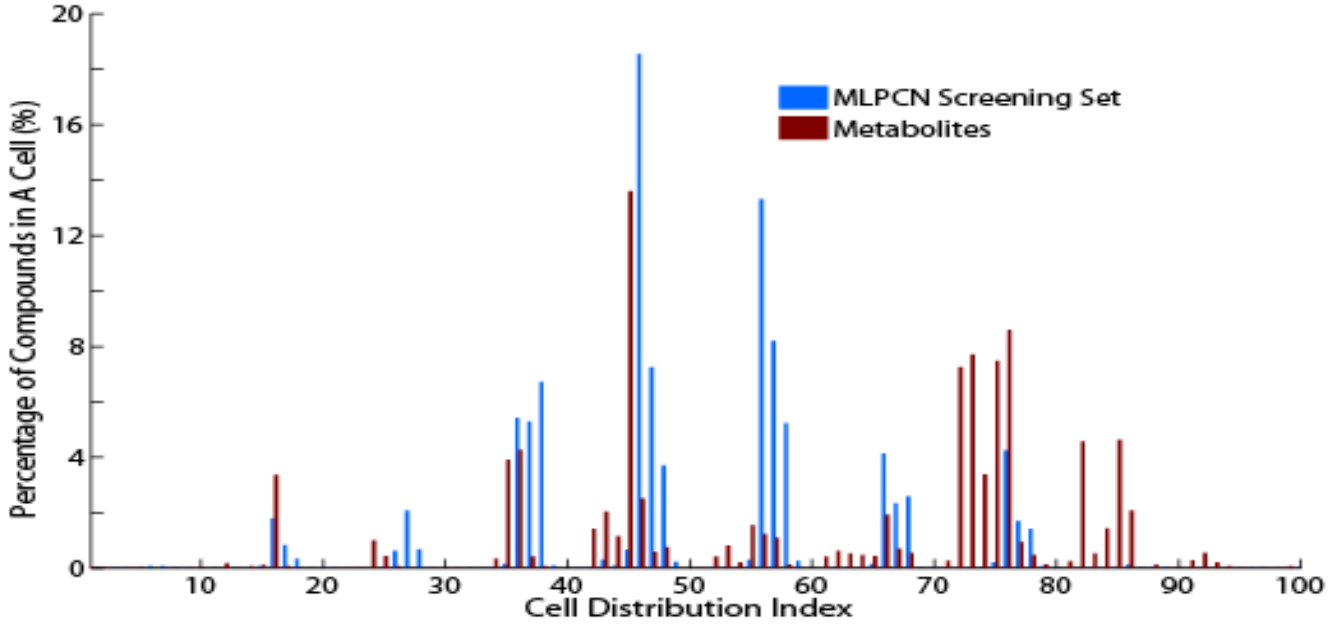
MLPCN Compounds *vs.* Approved Drugs



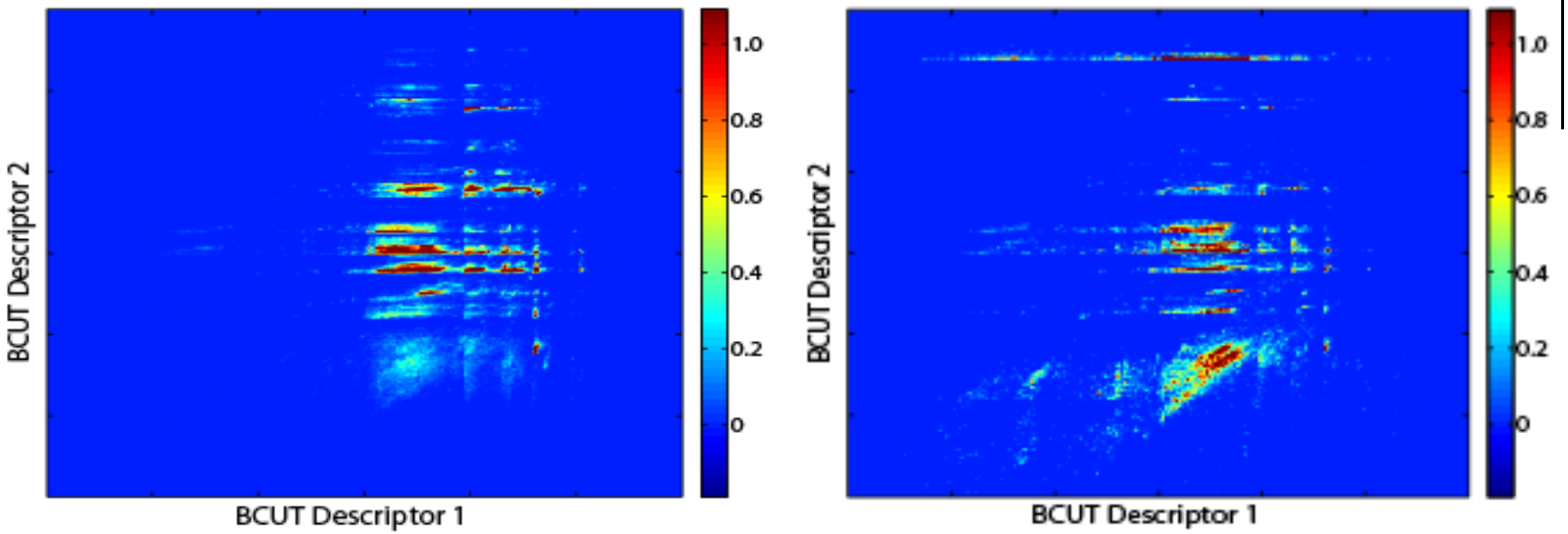


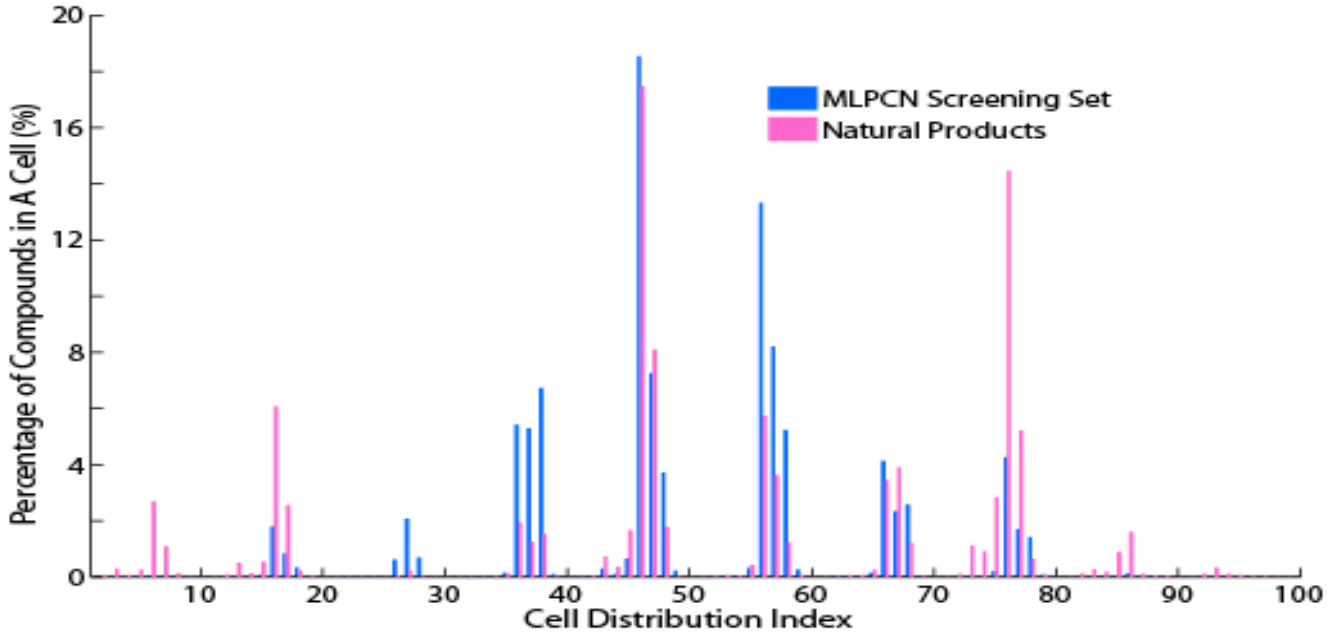
MLPCN Compounds *vs.* Metabolites



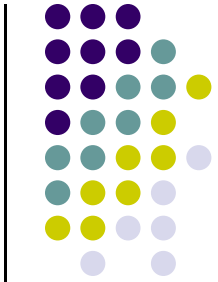


MLPCN Compounds *vs.* Natural Products





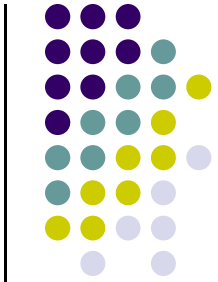
Summary (not in a definite sense)



- The MLPCN screening set is found to be a well-chosen subset of
  - Available drug-like small molecules
  - A highly diverse compound collection with greater biogenic bias than a comparable-sized set of commercially available compounds,
    - Incorporation of more metabolite-like chemotypes.

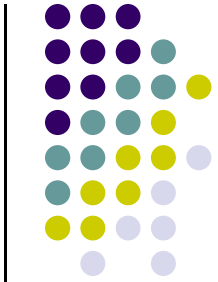
- Enhance the screening set diversity by exploring regions of chemical space that are under-populated in the MLPCN set relative to other biogenic compound collections
  - Potentially enhance the quality of resulting bioassay data in ways suitable for advancing both basic research and rational drug discovery.

Part III: Kernels for Chemical Activity Prediction



- Chemical Activity Prediction
- Chemical Graphs and Features
- Kernels for Structured Data
- Kernels for Chemical Graphs
  - Path-based: random or all sequences of specific length
  - Semi-structured: subtrees and cycles
  - General subgraphs, alignment

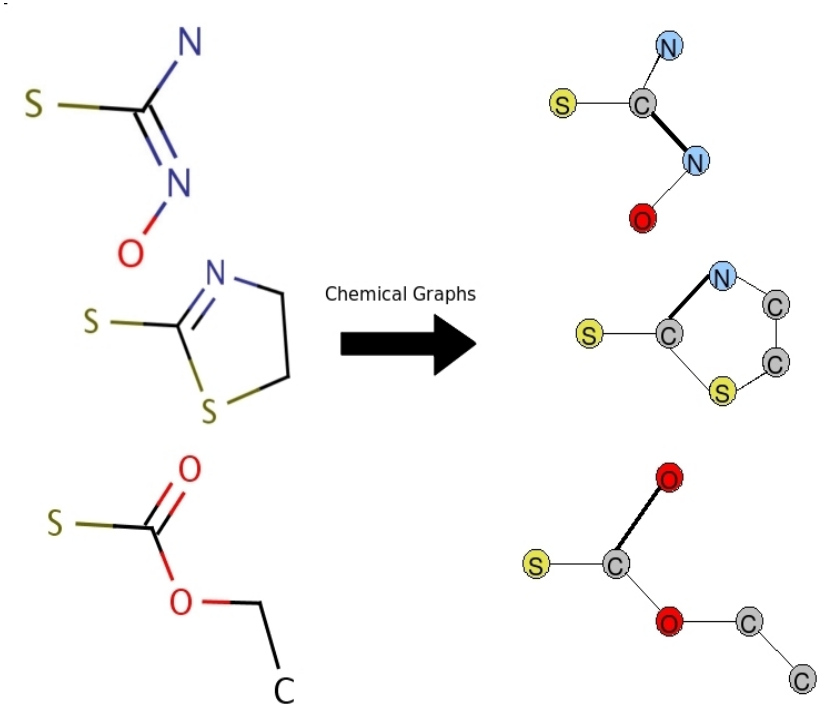
Chemical Activity Prediction



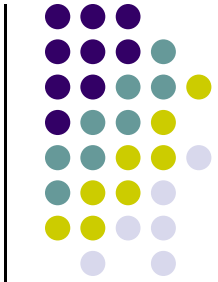
- Activity is observed chemical function
  - Toxicity, binding affinity, intestinal absorption, etc.
  - Important for screening candidate drugs
- Functional activity depends on structure
  - Compounds with 'similar' structure might have similar function
  - ‘Similar' structures, with similar activity, may share common structure features

Chemical Graphs

- Use graph representation for chemical activity prediction to retain rich expressivity
- Transformation of chemicals to graphs is straight forward.
  - Atoms correspond to vertices.
  - Bonds correspond to edges.
  - Vertices and edges are labeled with atom element and bond type, among other properties.

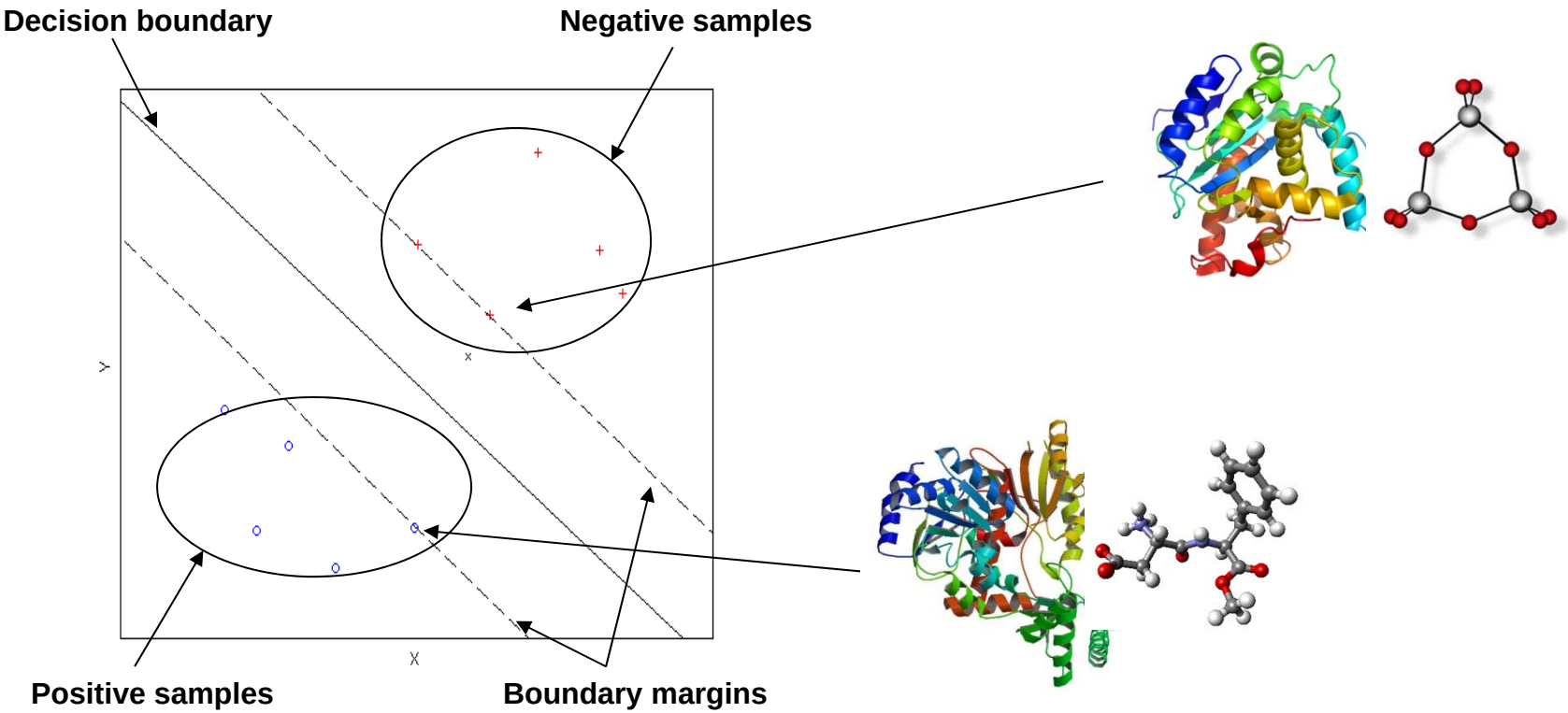


Chemical Classification

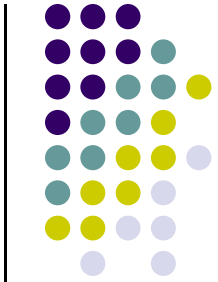


- Machine learning tools typically require a numeric sample-feature matrix as the input representation.
- The classification of chemical graphs requires some way to embed them in a suitable space, either explicitly or implicitly.

Embedding Graphs for Classification

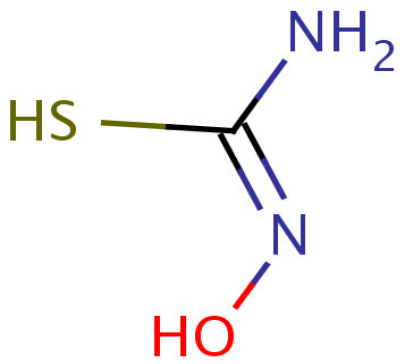


Chemical Features



- The features describing a chemical graph embed it in a spatial representation.
- Chemical features take many forms, such as those describing an entire molecule, or those describing particular atoms.

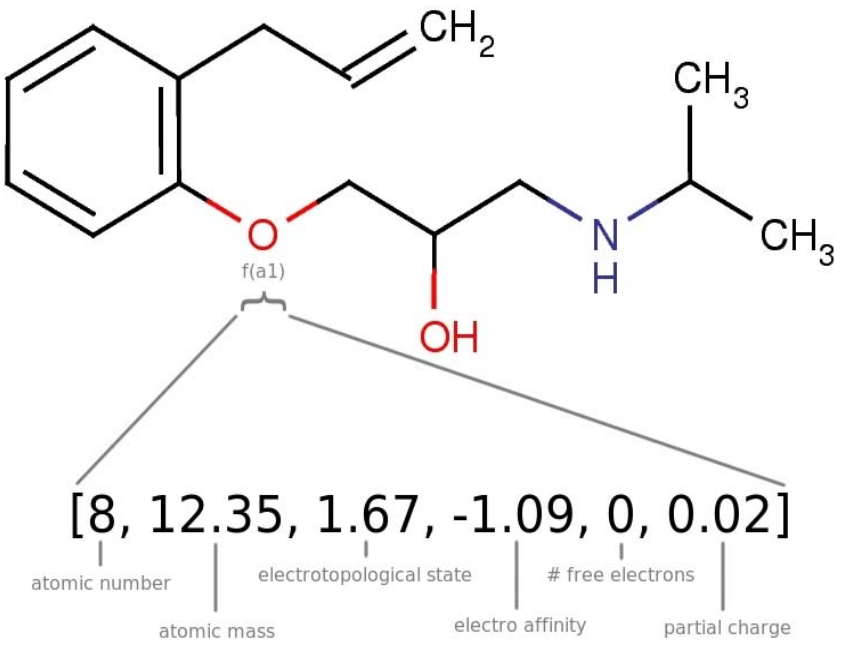
Examples of Features



[9, 5, …]

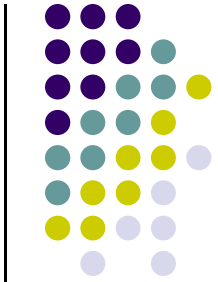
# of atoms # of bonds . . .

Molecular



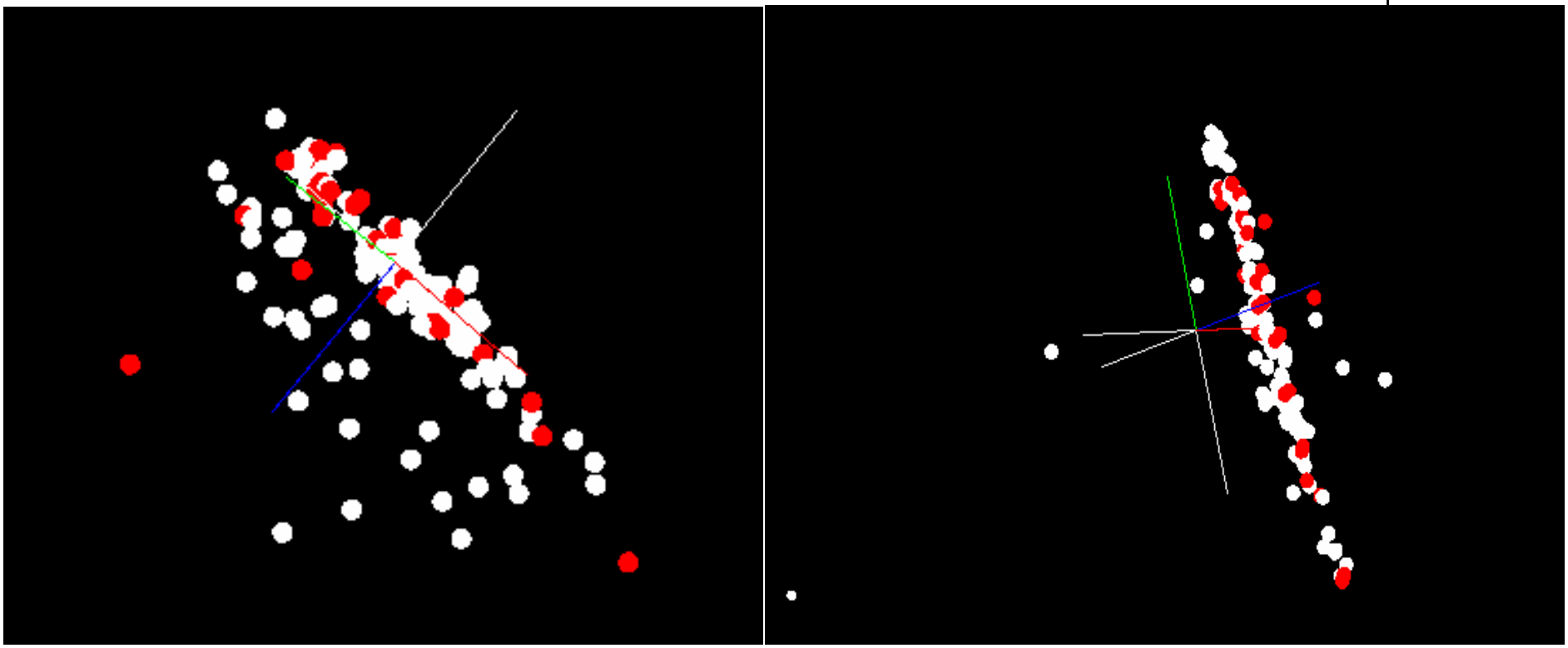
Atomic

Kernel Methods

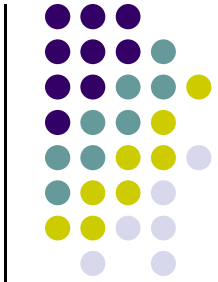


- Instead of explicitly computing features, compare chemical graphs using a kernel function.
- This kernel matrix of pair-wise similarities embeds chemical graphs into a space suitable for classification.
- The kernel function between two objects replaces the inner product of two feature vectors in the classifier optimization problem.
- Shift from finding good classifier to finding good kernel function.

Kernel Space Visualizations

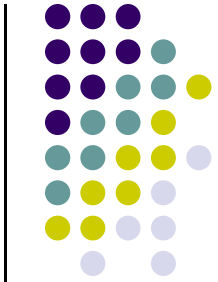


R-Convolution Kernel



- Kernels between chemical graphs are defined as cases of a general kernel between structured data, the R-Convolution Kernel.
- The difference in kernel functions depends on the method used to decompose complex graph structures into simpler ones.
- See Haussler, D. Convolution Kernels on Discrete Structures. Technical Report UCSC-CRL099-10, Computer Science Department, UC Santa Cruz, 1999

R-Convolution Definitions



$x \in X$ be a composite object

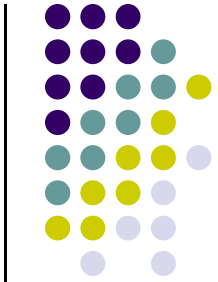
$\vec{x} = x_1, ..., x_D \in X_1 \times ... \times X_D$ are its parts

$\mathcal{R}$ defined on $X_1 \times ... \times X_D \times X$

$\mathcal{R}(\vec{x}, x)$ true iff $\vec{x}$ are the parts of x

$\mathcal{R}^{-1}(x) = \{\vec{x} : \mathcal{R}(\vec{x}, x)\}$

R-Convolution Equation



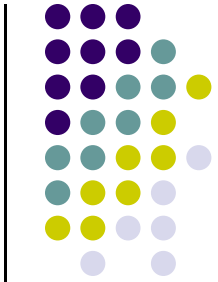
$$k_{\mathcal{R}}(x,y) = \sum_{\vec{x}\in\mathcal{R}^{-1}(x),\vec{y}\in\mathcal{R}^{-1}(y)} \prod_{d=1}^{D} K_d(x_d, y_d)$$

Kernel between composite objects *x* and *y*

Sum over the parts of *x* and *y*

Product of kernels between parts of *x* and *y*

Recursive Decomposition

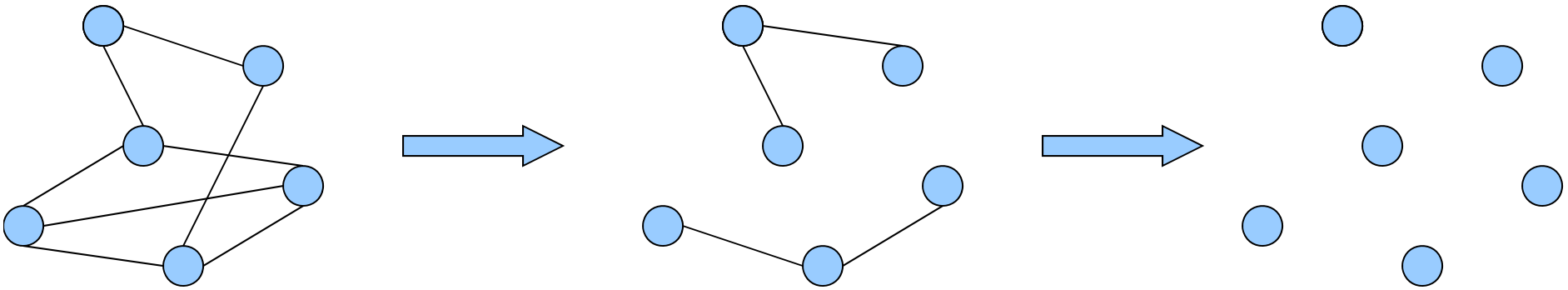


- The R-convolution kernel framework allows for recursive application.
- For example, a kernel between chemical graphs may depend on a kernel between linear molecular fragments, which may in turn depend on a kernel between individual atoms.

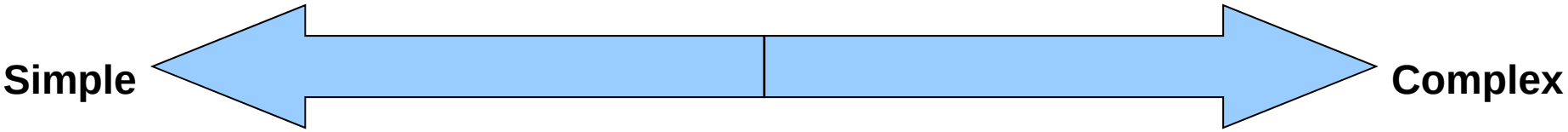
Recursive Decomposition

$$k_{\mathcal{R}}(x,y) = \sum_{\vec{x}\in\mathcal{R}^{-1}(x),\vec{y}\in\mathcal{R}^{-1}(y)} \prod_{d=1}^{D} K_d(x_d, y_d)$$

$$k_{\mathcal{R}}(x,y) = \sum_{\vec{x}\in\mathcal{R}^{-1}(x),\vec{y}\in\mathcal{R}^{-1}(y)} \prod_{d=1}^{D} K_d(x_d, y_d)$$



Range of Decompositions

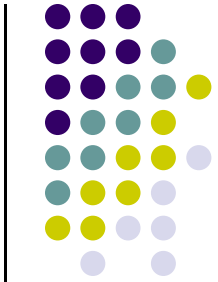


- numeric vectors

- strings
- sets
- multi sets

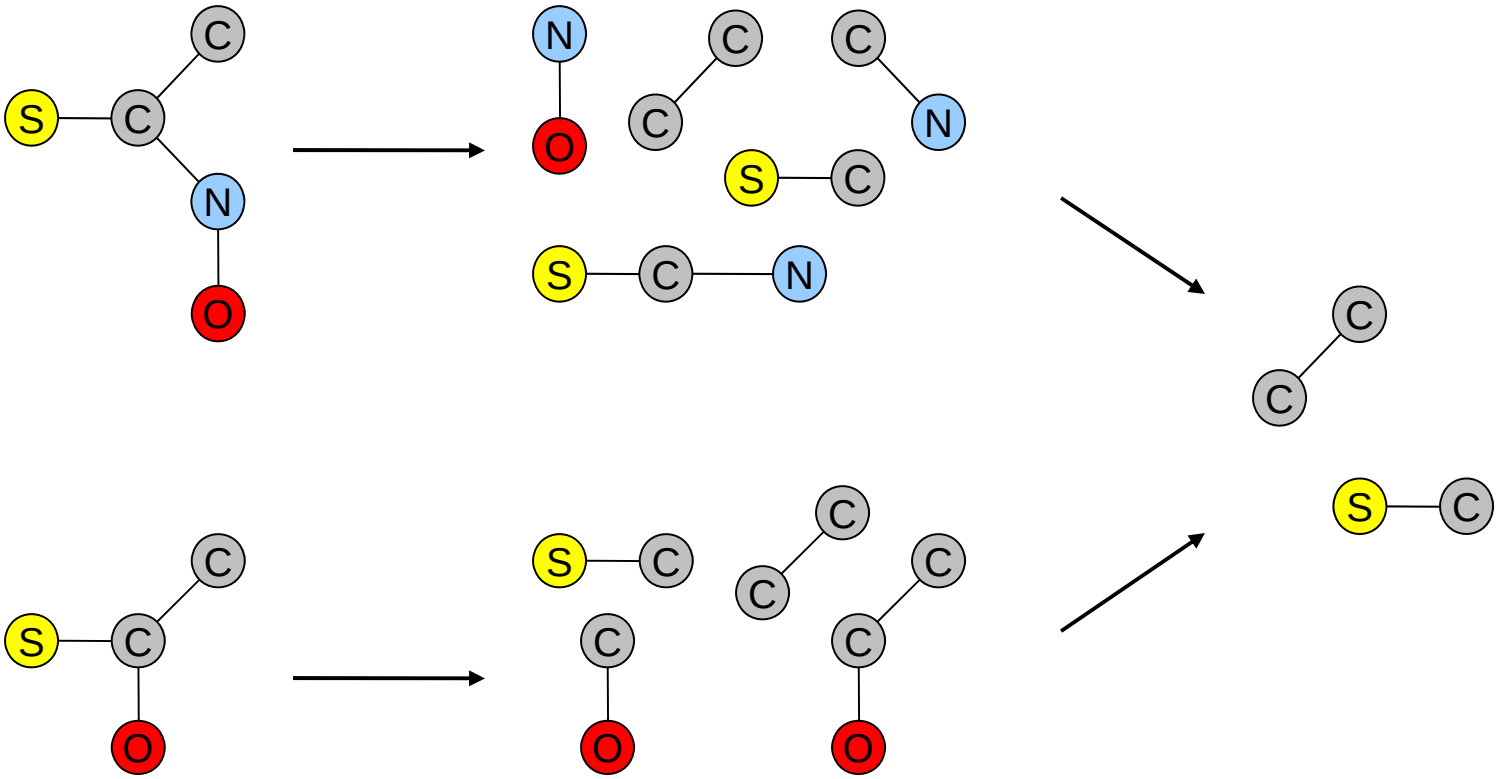
- graphs
- trees
- cycles

Path-based Kernels

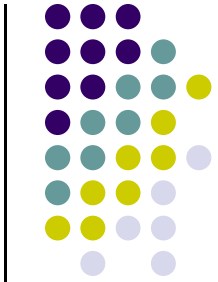


- Construct kernels by computing shared path substructures
- Examples:
  - Marginalized kernel - P. Mahé, et al. Graph kernels for molecular structure-activity relationship analysis with support vector machines. J Chem Inf Model, 45(4):939–51, 2005.
  - Spectrum kernel - C. Leslie, E. Eskin, and W.S. Noble. The spectrum kernel: a string kernel for SVM protein classification. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Kevin Lauerdale, and Teri E. Klein, editors, Proceedings of the Pacific Symposium on Biocomputing 2002.
  - Perret, Mahe, Vert. Chemcpp: an open source C++ toolbox for kernel functions on chemical compounds. Software available at http://chemcpp.sourceforge.net 2007.

Finding Shared Paths

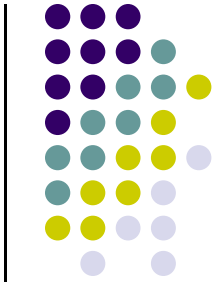


Marginalized Kernel (Mahé et al. 2005)



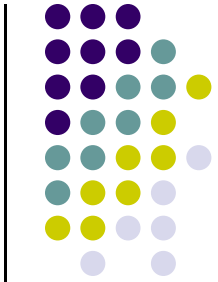
- Randomly generate a set of paths of a specified length from a chemical graph.
- Compute similarity for two chemical graphs based on the number of shared random paths.

Spectrum Kernel (Leslie 2002)



- Generate all paths in a chemical graph up to or exactly equal to a specified length.
- Again, compute the similarity between two chemical graphs according to the number of common paths.

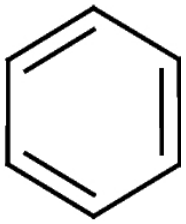
Kernels with non-path Features

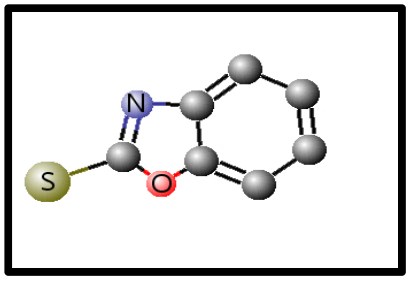


- Kernels with non-path features
- Examples:
  - Cyclic patterns - Horvath, Gartner, Wrobel. Cyclic pattern kernels for predictive graph mining. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004.
  - Subtree kernel - P. Mahé and J.P. Vert. Graph kernels based on tree patterns for molecules. Technical Report HAL:ccsd-00095488, Ecoles des Mines de Paris, September 2006.

Cyclic Kernel (Horvath 2004)

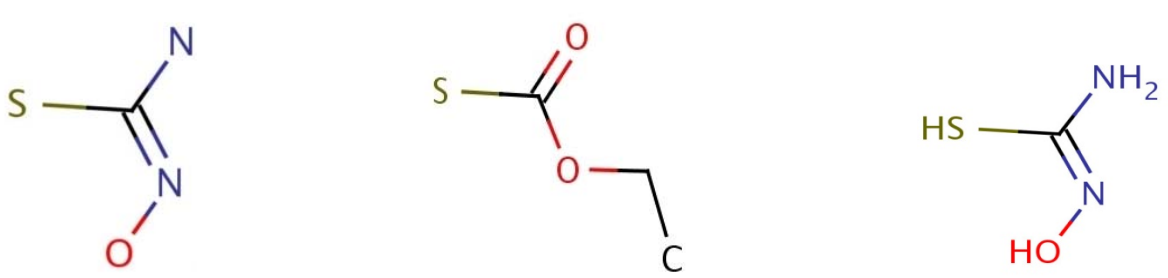
- Index chemical graphs as a set of cyclic patterns.
- Such patterns are common in organic molecules.
- Example cycle patterns:



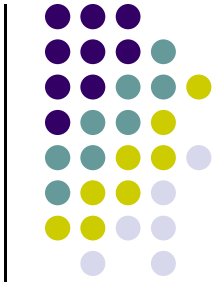


Subtree Kernel (Mahé 2006)

- Like cycles, subtrees or branching patterns are common in biology, particularly in lipid-type molecules.
- Subtrees are mined and chemical graphs are indexed by their presence or absence.
- Many small molecules are already trees:

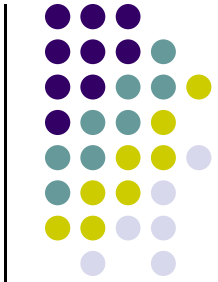


General Graph Kernels



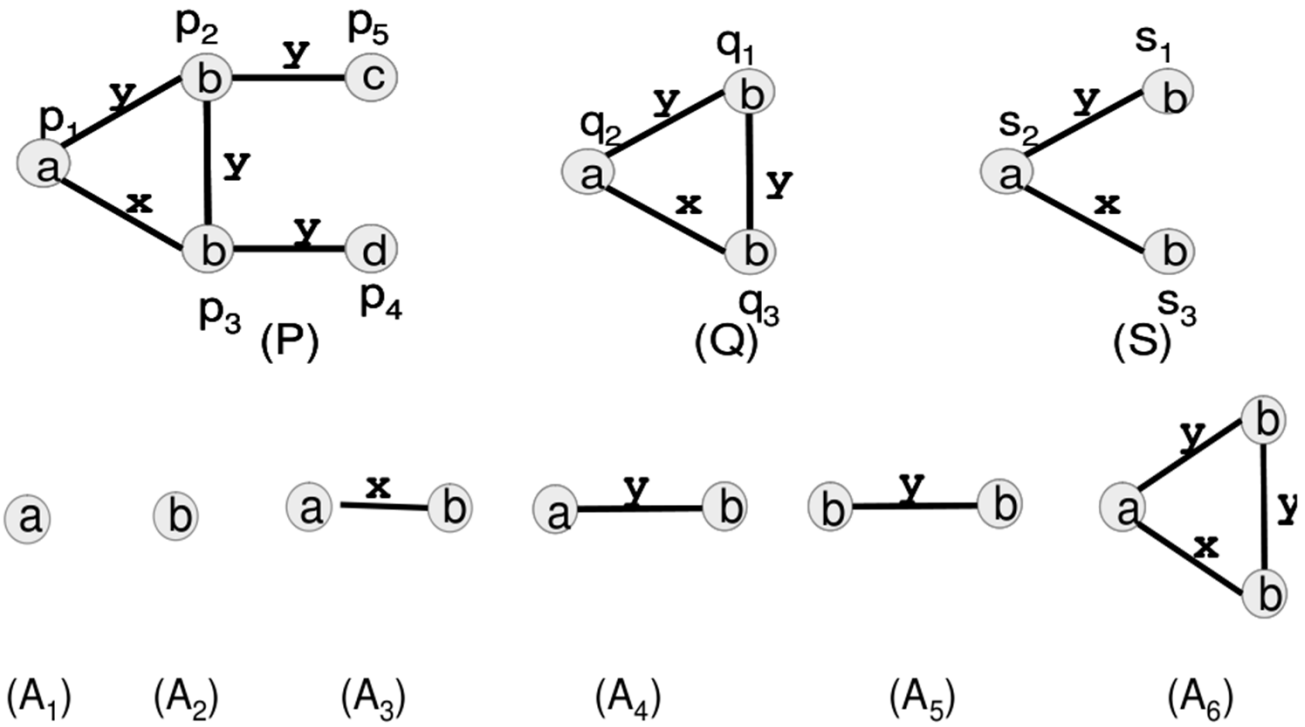
- Some kernels take advantage of the rich chemical graph structure and perform as little decomposition as possible.
- Examples:
  - Subgraph kernel - Mahé, Ralaivola, Stoven, and Vert. The pharmacophore kernel for virtual screening with support vector machines. Technical Report Technical Report HAL:ccsd-00020066, Ecole des Mines de Paris, march 2006.
  - Optimal Assignment kernel – Frohlich et al. Kernel Functions for Attriubted Molecular Graphs - A new Similarity-Based Approach to ADME Prediction in Classification. QSAR & Combinatorial Science 25(4), 2006

Subgraph Kernels

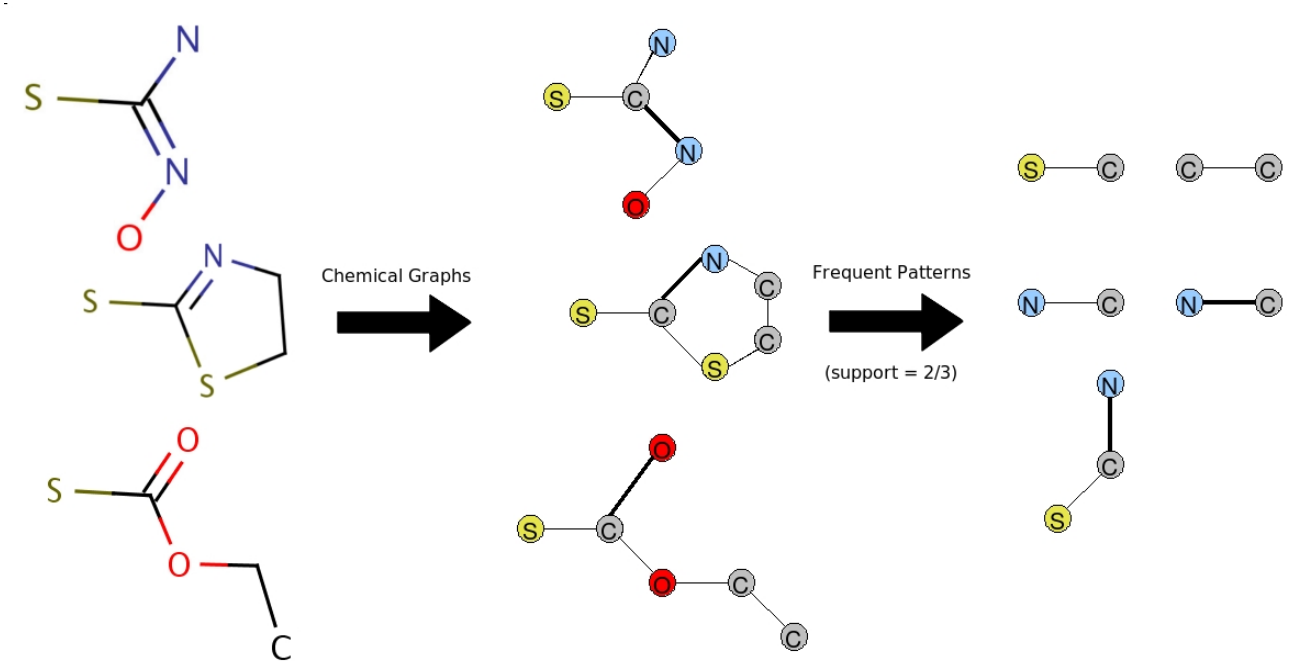


- Decompose chemical graphs into the most general substructure.
- Can mine patterns and compute similarity based on shared patterns.
- Many aspects of chemical activity are determined by functional groups or *pharmacophores* that can be represented as subgraphs and incorporated into a kernel. (Mahé 2006)

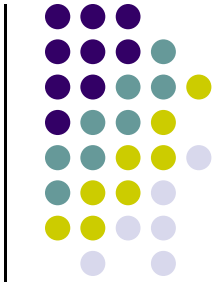
Subgraph Examples



Frequent Patterns in Chemical Graphs

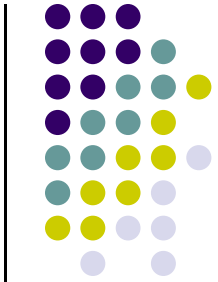


Pharmacophore Kernel (Mahé 2006)



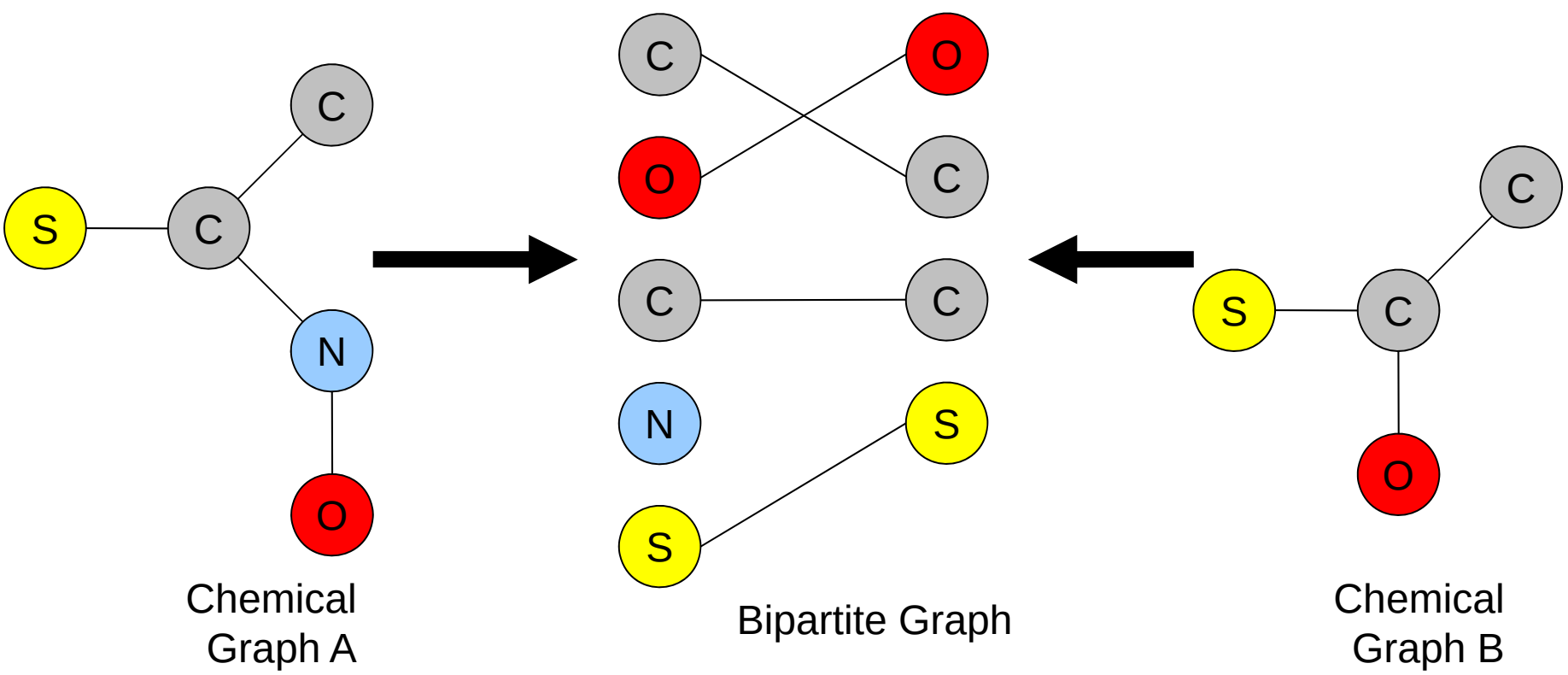
- Many molecular properties are determined by the existence of specific patterns that can attach to chemical scaffolds in a modular way.
- The 3-dimensional arrangement of these *pharmacophores* is also incorporated for chemical activity prediction.

Optimal Assignment Kernel (Frohlich 2006)

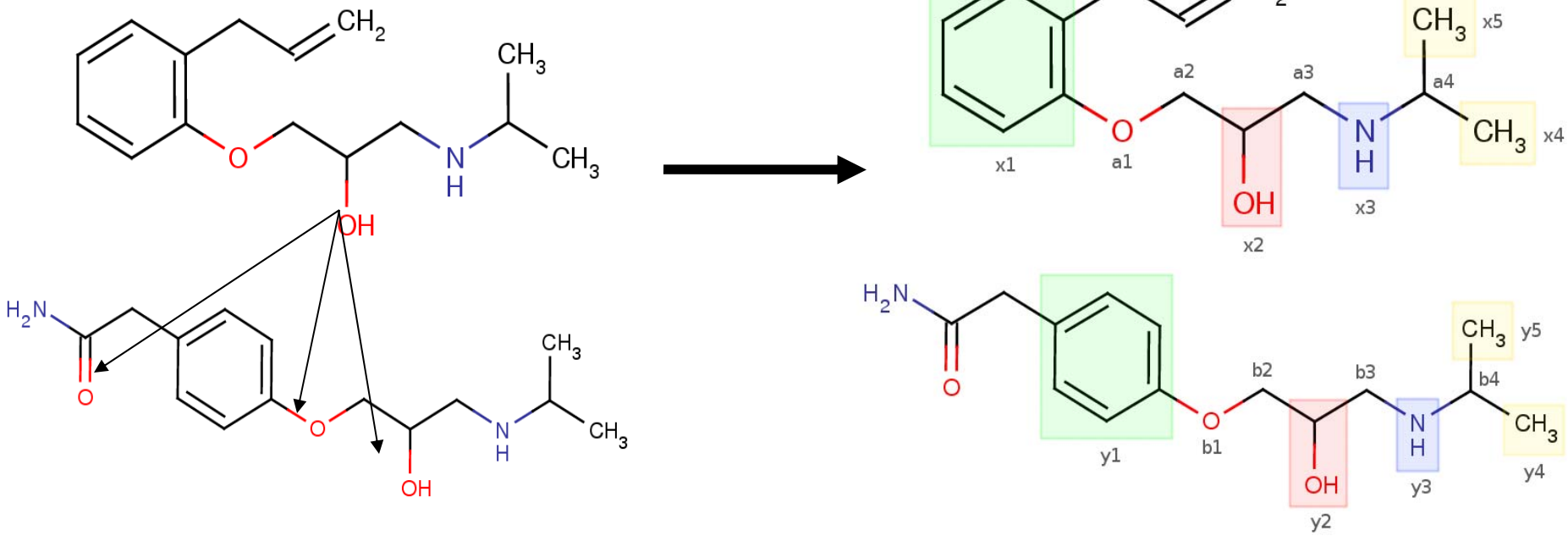


- Instead of using a decomposition, two chemical graphs are aligned by matching vertices from one graph to the other.
- Computes a maximum weighted bipartite graph between to sets of vertices, but is not positive semi-definite as originally published.
- Uses a recursive matching function to align groups of vertices.

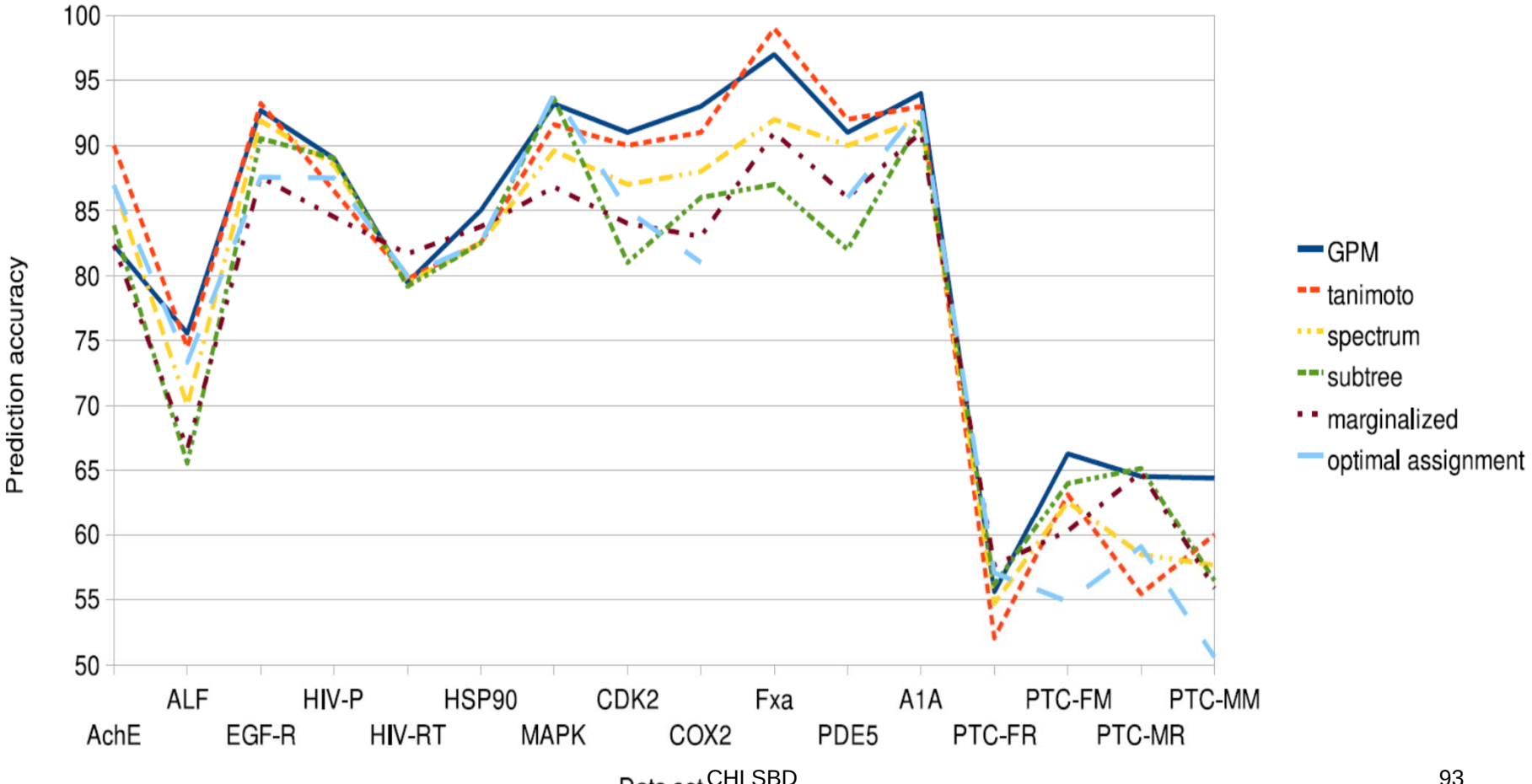
Bipartite Graph



Matching Vertices and Patterns

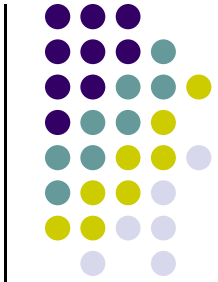


Protein-Chemical Interaction



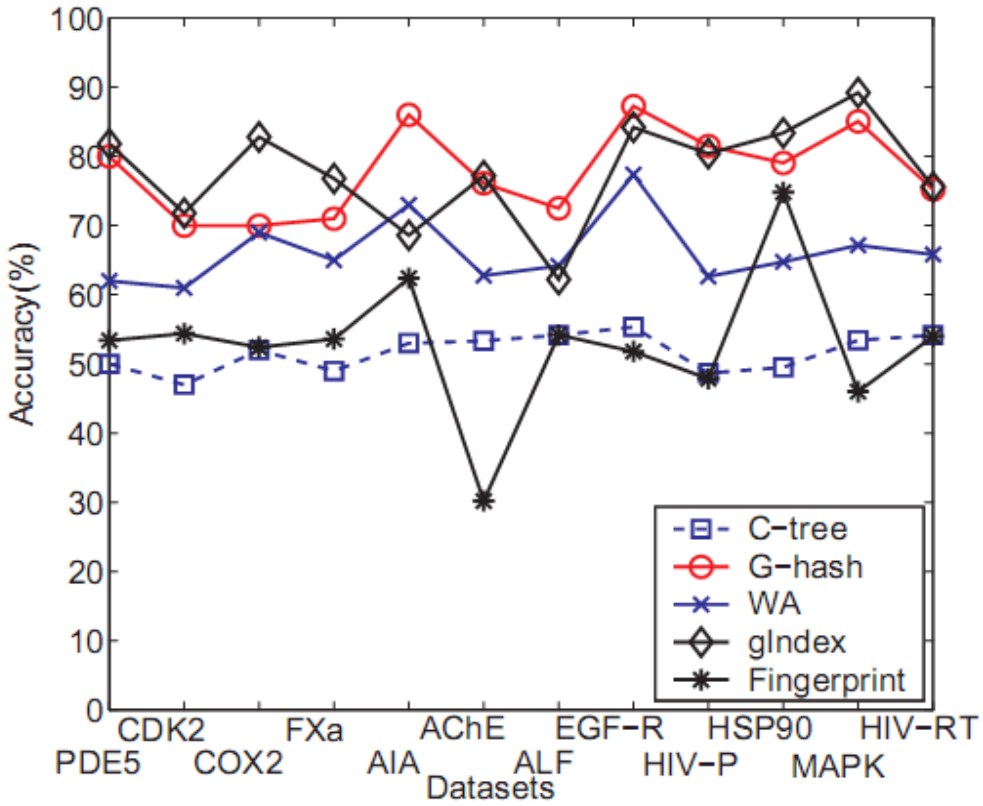
Samlter & Huan, IEEE/ACM Transaction of Bioinformatics and Computational Biology, 2010

Kernel Based Similarity Search



- Using the kernel functions to define similarity
- Scale up those kernel functions to chemical structure database search

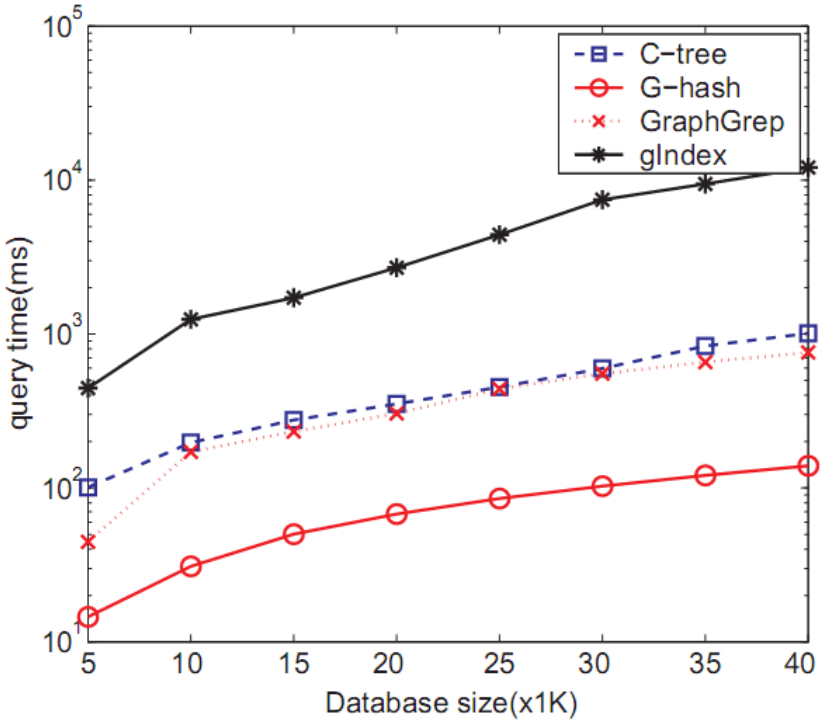
k-NN Classification Results



Wang et al., Application of Kernel Functions for Accurate Similarity Search in Large Chemical Databases, *BMC Bioinformatics* Vol. 11 (Suppl 3):S8, 2010

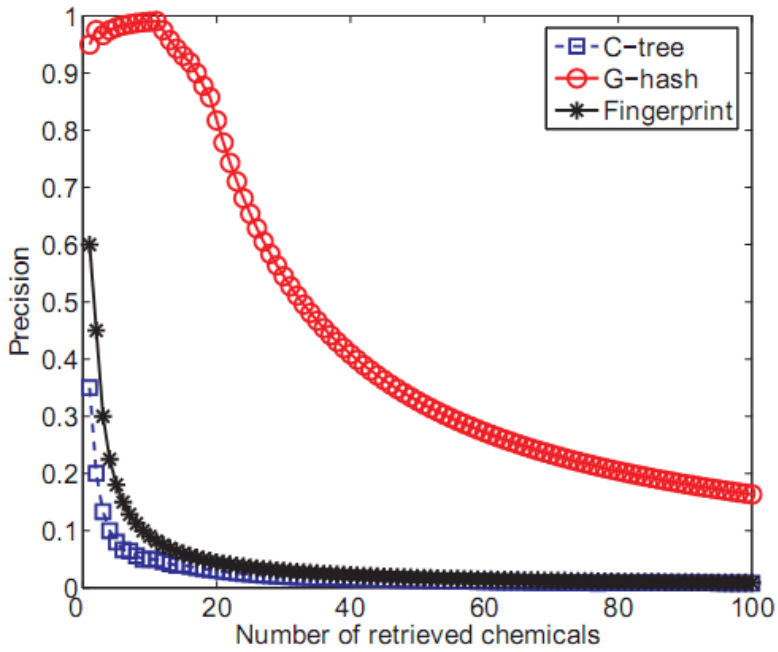
k-NN Query Processing Time

- We applied a novel kernel-based similarity measurement to measure similarity of chemicals.
- In our method, we utilize a hash table to support new graph kernel function definition, efficient storage and fast search.

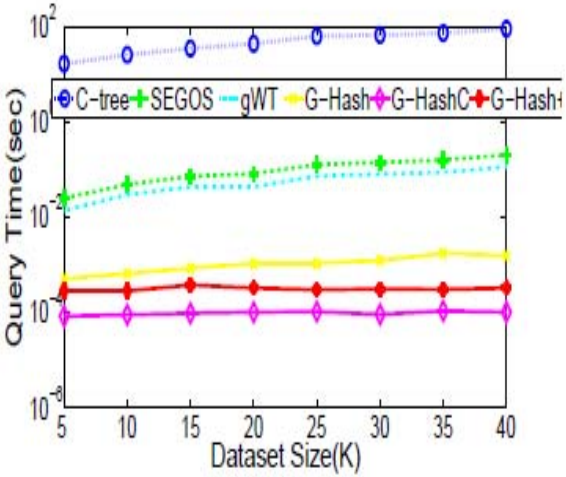


Chemical Enrichment Study

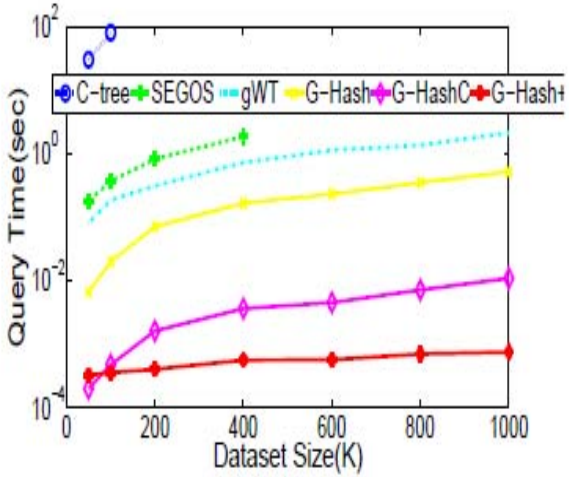
- Obtained 110 inhibitors of focal adhesion kinase 1 (FADK 1) with AID810 from PubChem
- Randomly picked 20 chemical compounds. Augmented them to the NCI/NIH AIDS data set to form a new database
- Picked one chemical from these 20 chemicals as the query chemical to search the new database and retrieve 100 nearest neighbors
- Computed precision as the percentage of chemicals in the top k compounds belongs to the true 19 hits



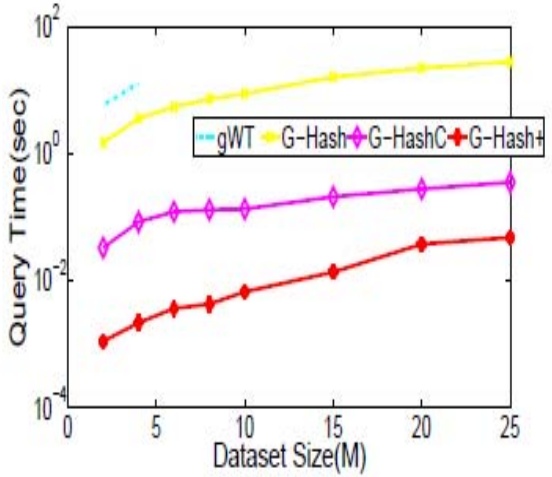
Scalability of the Algorithm



(a) The AIDS Database

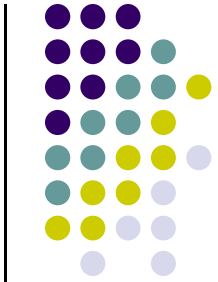


(b) A Small PubChem Database



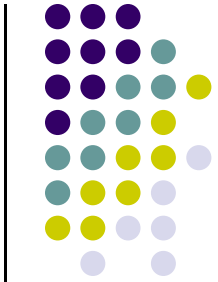
(c) A Large PubChem Database

Additional Chemical Features



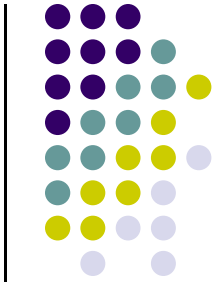
- Not limited to structure – much more data sometimes available
  - In fact common structure will not always reveal common effect
    - Whole picture of biological systems needed in reality
- Different ways of characterizing a drug (chemical) based on its effects and interactions

Additional Chemical Features



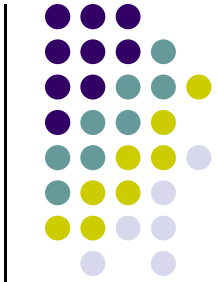
- Not limited to structure – much more data sometimes available
  - In fact common structure will not always reveal common effect
    - Whole picture of biological systems needed in reality
- Different ways of characterizing a drug (chemical) based on its effects and interactions

Part IV: Advanced Topics of Data Analysis in Drug Discovery



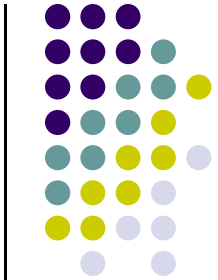
- Use Quantitative Structure Activity Relationship models
- Use machine learning, data mining, information retrieval, text mining, image analysis to understand information in a wide range of data types
- Modeling a variety of end-points
  - Protein-chemical interaction
  - Gene-chemical interaction
  - Chemical toxicity
  - Absorption, distribution, metabolism, and excretion (ADME) properties

Additional Chemical Features



- Different ways of characterizing a drug:
  - Interaction Profiles
    - Chemical-protein interactions, chemical-genetic interactions
  - Drug Effects (Phenotypical – *text mining*)
    - Side effect profile
    - Pharmacological effects
  - In Vitro/ In Vivo test effects
    - Genetic profiles, screening profiles

Interaction Profiles



- Characterize an object based on its interactions (interactome) with another set
  - Similar idea to kernel methods

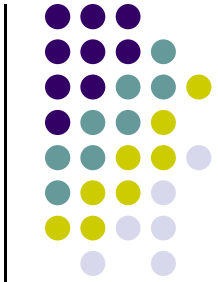
Proteins of Interest

Compounds tested against proteins

	P1	P2	P3	P4	…
C1	1	0	0	1	…
C2	0	1	1	1	…
…					…

  - Interaction networks (graph), binary vector or real-valued activity/interaction strength

Chem.-Protein Interaction Profile Example

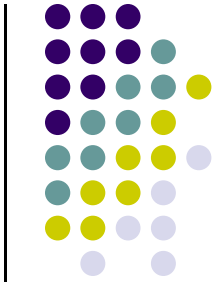


- Chemical effects are usually result of multi-protein interactions (Hopkins 2008)
  - Proteome similarity good indicator of common effects
- (Yang 2009) exploited protein interactome of chemicals using data mining techniques for exploring Severe Adverse Drug Reaction (SADR)
  - Determine common protein sub-groups
  - Classify SADR using profile

A. L. Hopkins, “Network pharmacology: the next paradigm in drug discovery.” Nature Chemical Biology 4, 682 – 690, 2008.

L. Yang, J. Chen, and L. He, “Harvesting Candidate Genes Responsible for Serious Adverse Drug Reactions from a Chemical-Protein Interactome.” PLoS Comput Biol 5(7), 2009.

Chem.-Genetic Interaction Example

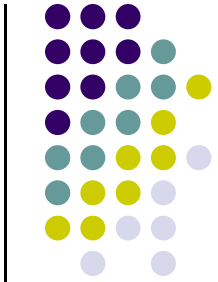


- Chems. with different structure can still share common effects! (*structure isn't always enough)*
- (Parsons 2004, Parsons 2006) used chemical-genetic profiles (a.k.a. hyper-sensitivity profiles)
  - To infer protein or pathway targets and
  - To identify pathways protecting against toxic effects of a drug
  - Potentially providing info. about compound's mode of action

Parsons et al., "Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways." Nat. Biotechnol. 22:62–69, 2004.

Parsons et al., "Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast." Cell 126:611–625, 2006.

Chem.-Genetic Interaction Example

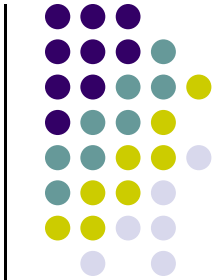


- Compounds with very different structures can have similar modes of action, captured by chemical-genetic profile
  - E.g. two highly selective inhibitors of Hsp90, highly unrelated structurally, similar chemical-genetic profiles (Parsons 2006)
- Looked at inhibitors with yeast and gene knockout

Parsons et al., "Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways." Nat. Biotechnol. 22:62–69, 2004.

Parsons et al., “Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast.” Cell 126:611–625, 2006.

Chem.-Genetic Interaction Example

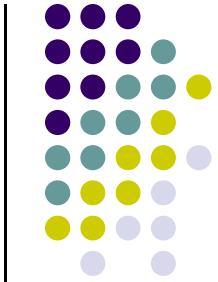


- Chemical-genetic profile:
  - Interaction is characterized by combination of chemical with gene knockout leading to cell death (or defects)
    - Emerging high-throughput method
  - ~5,000 yeast deletion mutants and up to 82 compounds tested
- 2-D hierarchical clustering and probabilistic sparse matrix factorization for visualization and to identify compounds with similar modes of action

Parsons *et al.,* "Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways." Nat. Biotechnol. 22:62–69, 2004.

Parsons *et al.,* "Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast." Cell 126:611–625, 2006.

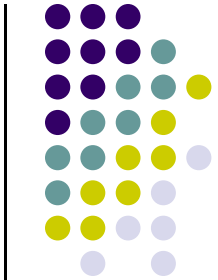
Side Effects as Features



- Adverse Drug Reactions (side effects) used to predict drug-target interactions
  - (Campillos 2008) demonstrated how side-effects could reveal unknown interactions
  - Drugs with similar (phenotypic) side-effect profiles used to predict common targets
    - Reveal existing FDA-approved drugs for one disease could be used for a different one
    - E.g.: Rabeprazole (protein-pump inhibitor) used to treat stomach ulcers and pergolide (dopamine receptor agonist) have common side-effect profile – rabeprazole shown to bind to dopamine receptos (Campillos 2008)

Campillos *et al*., "Drug target identification using side-effect similarity." Science, 321, 2008.

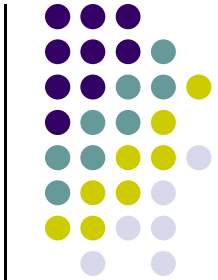
Side Effects as Features



- (Kuhn 2008) provides side-effect database free for academic use - SIDER:
  http://sideeffects.embl.de/
  - 1,450 side effects, 888 drugs
  - Drug side-effects were collected using text mining approach from package inserts of drugs –e-format
  - From public sources such as FDA
  - Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART) - side effect lexicon

Kuhn *et al*., “A side effect resource to capture phenotypic effects of drugs." Molecular Systems Biology, 6:343, 2009.

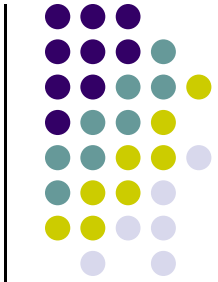
Pharmocological Effects



- (Yamanishi 2010) used chemical structure, protein sequence, and general phenotypic effects of the chemicals to predict chemical-protein interaction network
- Pharmocological effects:
  - keywords for drugs were obtained from the JAPIC (Japan Pharmaceutical Information Center) database - 18,653 keywords in total
  - Grouped: pharmaceutical effects, adverse effects, caution, usage, properties, etc. (general text info.)

Yamanishi *et al*., “Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework" ISMB, 2010.

Pharmocological Effects



- "Pharmaceutical effects" key words used as binary features
- Two step process:
  - Use known pharmacological effects to predict unknown ones in chemicals (regression model)
  - Use known and predicted pharmacological features to predict drug target interaction network
    - Embed drugs in targets into a unified space, and use distance threshold to determine interaction

Yamanishi *et al*., "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework" ISMB, 2010.

Relationship Between Chemical and Pharmacological Spaces w.r.t Drug Targets

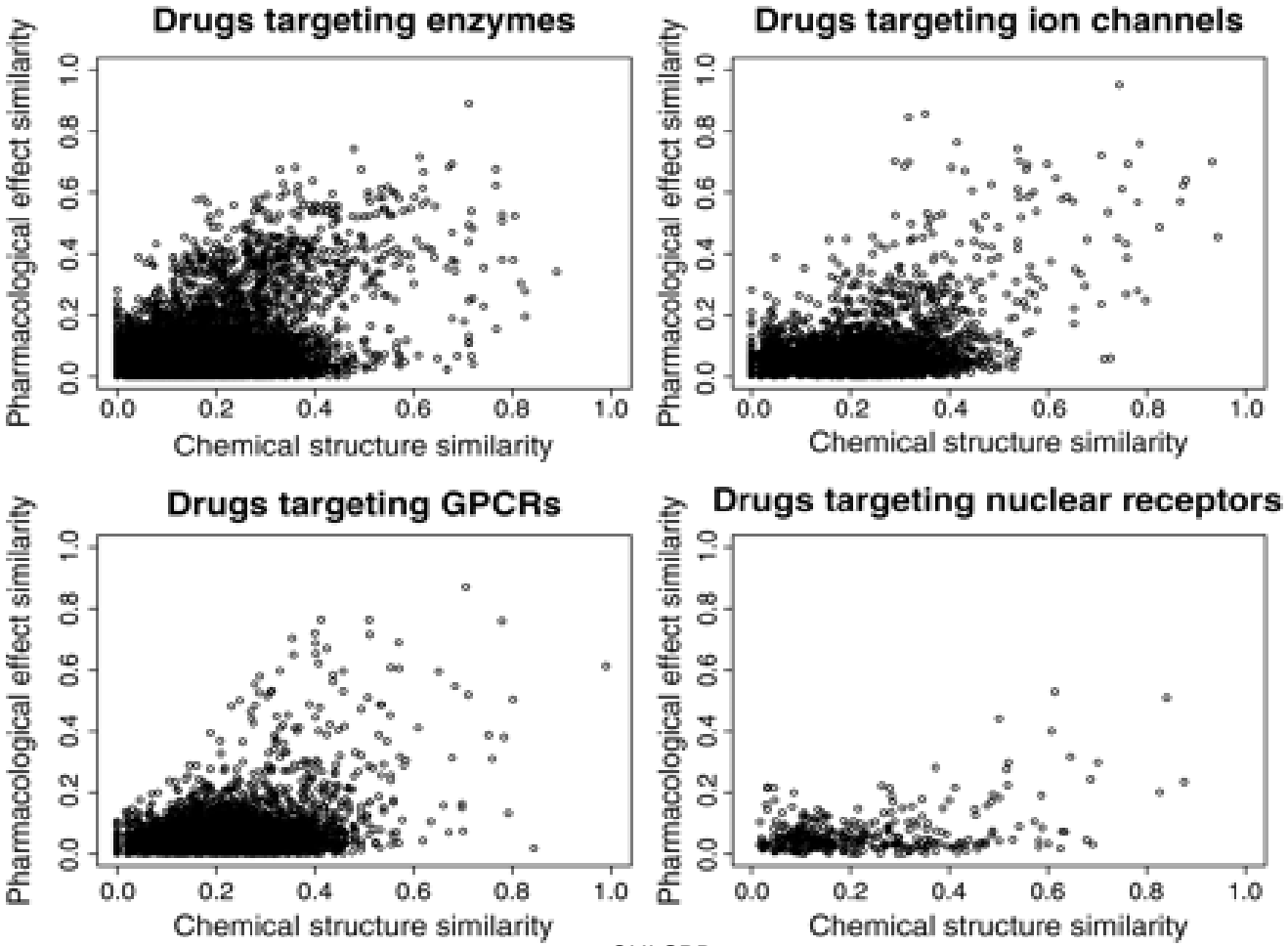
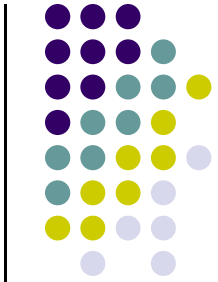


*Figure taken with permission from (Yamanishi 2010)

In Vitro Screening

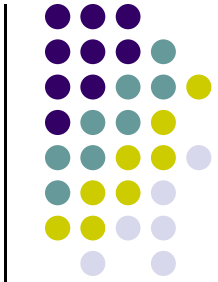


- In vitro (test tube) experiments can be designed to measure indicators of a drug's effects – features
  - E.g. measure gene expression, transcriptional responses, protein function, etc. of samples/cells of interest combined with drugs
  - E.g. (Iorio 2010), (Judson 2010)
  - High-throughput screening approach (HTS) quicker, less expensive then obtaining end-points

Iorio *et al.,* "Discovery of drug mode of action and drug repositioning from transcriptional responses." PNAS 107(33), 2010.

Judson *et al., In Vitro* Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project." Environ Health Perspect 118(4) 2010.

In Vitro Screening



- U.S. Environmental Protection Agency's (EPA) ToxCast Program (Judson 2010)
  - Phase I profiled >300 chiefly pesticide chemicals
  - Over 400 HTS endpoints collected - biochemical assays of:
    - protein function
    - cell-based transcriptional reporter and gene expression
    - cell line and primary cell functional
    - developmental endpoints in zebrafish embryos and embryonic stem cells

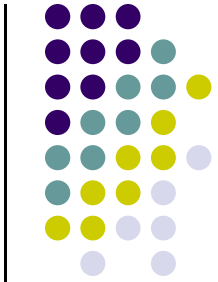
Judson *et al., In Vitro* Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project." Environ Health Perspect 118(4) 2010.

In Vitro Screening

- U.S. Environmental Protection Agency's (EPA) ToxCast Program (Judson 2010)
  - ~$2 billion in animal toxicity studies
    - Battery of toxicology methods to obtain reliable toxicity end-points:
      - Developmental toxicity, multi-generation reproductive studies, sub-chronic and chronic rodent bioassays, etc.
  - Phase II will expand chemicals tested – more and wider variety

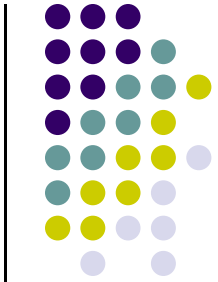
Judson *et al., In Vitro* Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project." Environ Health Perspect 118(4) 2010.

Nature of Non-Structure Features



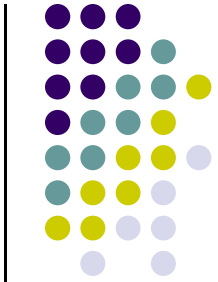
- Typically all *expensive* to obtain
  - Unlike structure-based chemical descriptors that are fast, inexpensive, and easy to obtain
  - Missing values – not all information is available for all drugs, all features, etc.
    - e.g. missing protein interactions, only tested for some compounds, some proteins, different proteins for different compounds
    - In vitro and in vivo tests expensive, not likely to get for all data e.g. EPA data – time and cost
    - Missing pharmacological effects

Nature of Non-Structure Features



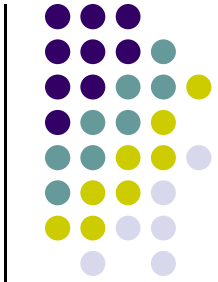
- Typically all *expensive* to obtain
  - Transfer learning:
    - As a result, must make use of what labeled data available
      - Expensive and time-consuming to obtain end-points and additional features for specific set of chemicals or e.g. targets
    - However chemical space is huge, we must consider effects of selection bias when using existing available data to reduce time and cost
      - Different targets, sets of chemicals, different marginal or conditional distributions – *transfer learning*

Nature of Non-Structure Features



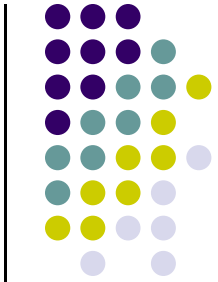
- Another potential solution: adaptive data mining techniques
  - E.g. active learning
  - Adaptively determine what information is most necessary (which compounds to test, etc.) to achieve some goal, e.g. elucidate chemical activity model

Adaptive Approaches to Drug Discovery



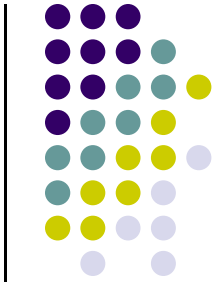
- Computational methods could make drug development process more adaptive
  - Adaptive techniques could improve efficiency and success (reduce costs) of drug discovery process
    - Model P(drug high success | drug descriptor, drug combinations/conditions, sample indicators, etc.)
    - To better understand P(Y|X), choose most informative test
  - Active learning
  - Bayesian Clinical Trials

Adaptive Approaches to Drug Discovery



- Active learning with computational models of activity can aid in hit/lead identification
  - Drug Discovery Process:
    - Identify Target
    - Test an initial set of chemicals against target (HTS)
    - Based on results **refine activity model** (chemist or **machine**)
    - Suggest next set of chemicals to test
    - Repeat…

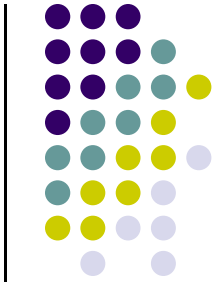
Adaptive Approaches to Drug Discovery



- Active learning:
  - Repeated tests – but tests cost
  - Ability to choose which instances to obtain label
  - Exploit choice to *identify most hits* **or** *reveal most info. about activity model* as quickly as possible
- E.g. Warmuth *et al.* 2003, SVM approach to identify candidate drugs to screen
  - Farthest from hyperplane – most certain
  - Closest to hyperplane – most uncertain

Warmuth *et al.* "Active Learning with Support Vector Machines in the Drug Discovery Process," Journal of Chemical Information and Computer Science, 2003.

Adaptive Approaches to Drug Discovery



- Maintaining and updating model of drug success – also apply to other phases
- Highlight: “Bayesian Clinical Trials” (Berry 2006)
  - Adaptive, computational approaches successfully used to help regulate clinical trials
  - Case study of FDA approved drug

D. Berry, “Bayesian Clinical Trials.” Nature Reviews Drug Discovery, 2006

Conclusions

- Drug discovery is a very very expensive process
- Enormous opportunities for data analytics.
  - Data are increasingly becoming publically available
  - No one knows the best practice to discovery a drug (even big pharms in the business >100 years)
- Challenges:
  - Do not underestimate the beast!

Questions?

- Dr. Jun (Luke) Huan

  Associate Professor

  Department of Electrical Engineering and Computer Science

  University of Kansas

  jhuan@ku.edu

  http://people.eecs.ku.edu/~jhuan

- I thank CHI and KU Special Chemistry Center for sponsoring my talk