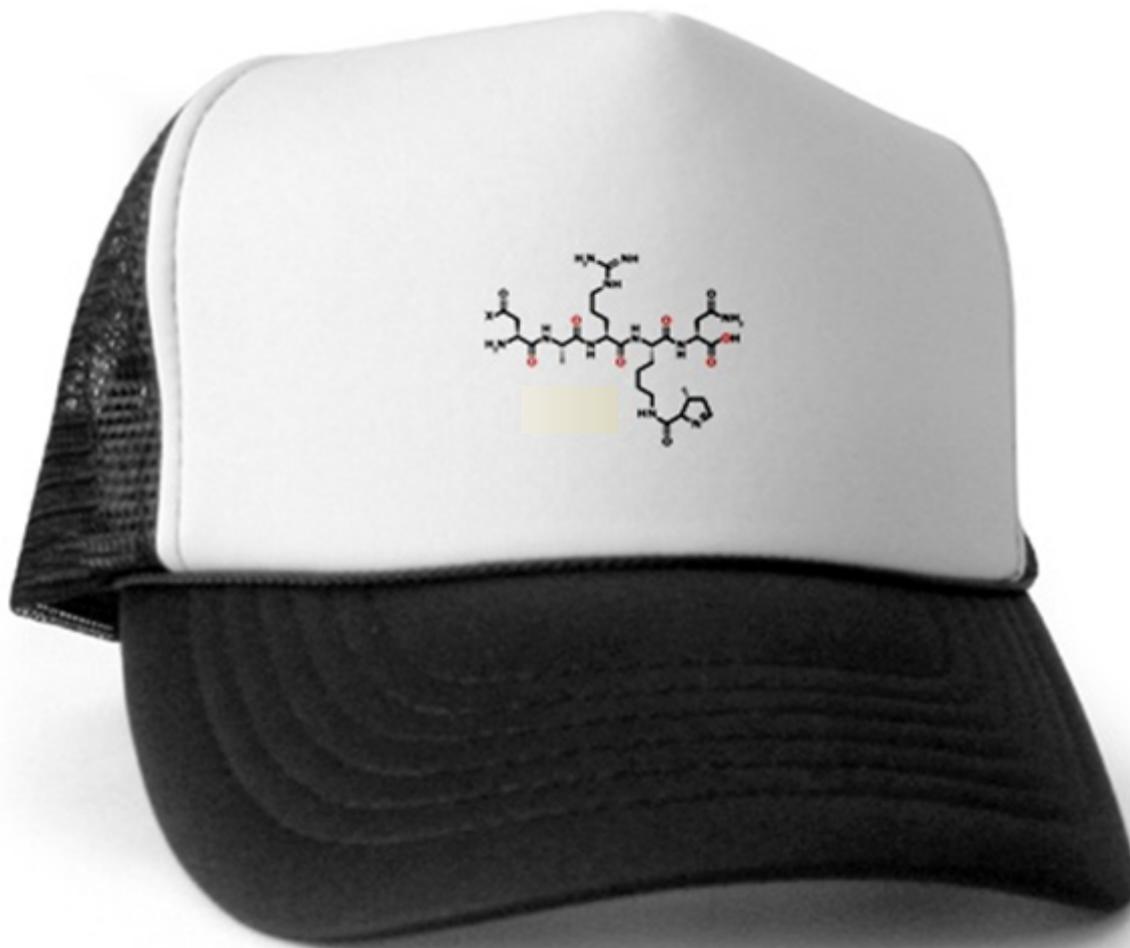


Drug and Probe Discovery and its Mathematical Challenges

George Karypis

karypis@cs.umn.edu

Department of Computer Science & Engineering
University of Minnesota



Some terminology

Assay: A biological test, measurement or analysis to determine whether compounds have the desired effect either in a living organism, outside an organism, or in an artificial environment.

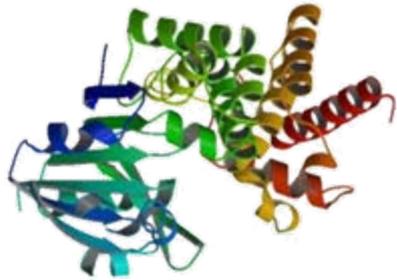
Hit compound: A biologically active compound that exceeds a certain activity threshold in a given assay.

Lead compound: A compound that exhibits pharmacological properties which suggest its value as a starting point for drug development.

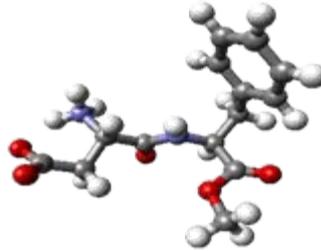
Drug like compound: Sharing certain characteristics with other molecules that act as drugs. The set of characteristics usually include size, shape solubility in water and organic solvents. These characteristics relate to absorption, distribution, metabolism, and excretion (ADME).

Chemical probe: A chemical compound with activity in the primary and any secondary assays with adequate potency, selectivity, and aqueous solubility to be useful for in vitro (cell-based) experimentation.

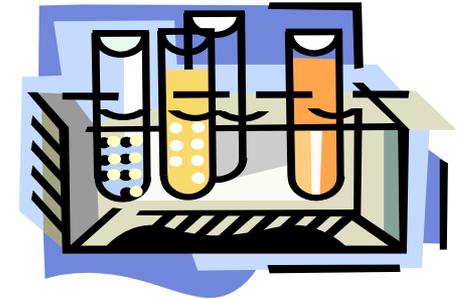
Drug development process (the cartoon view)



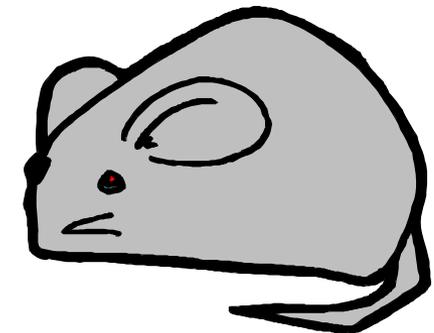
Target discovery



Hit/Lead discovery
& optimization



Small scale
production



Laboratory and
animal testing

$10^6 \rightarrow 1$

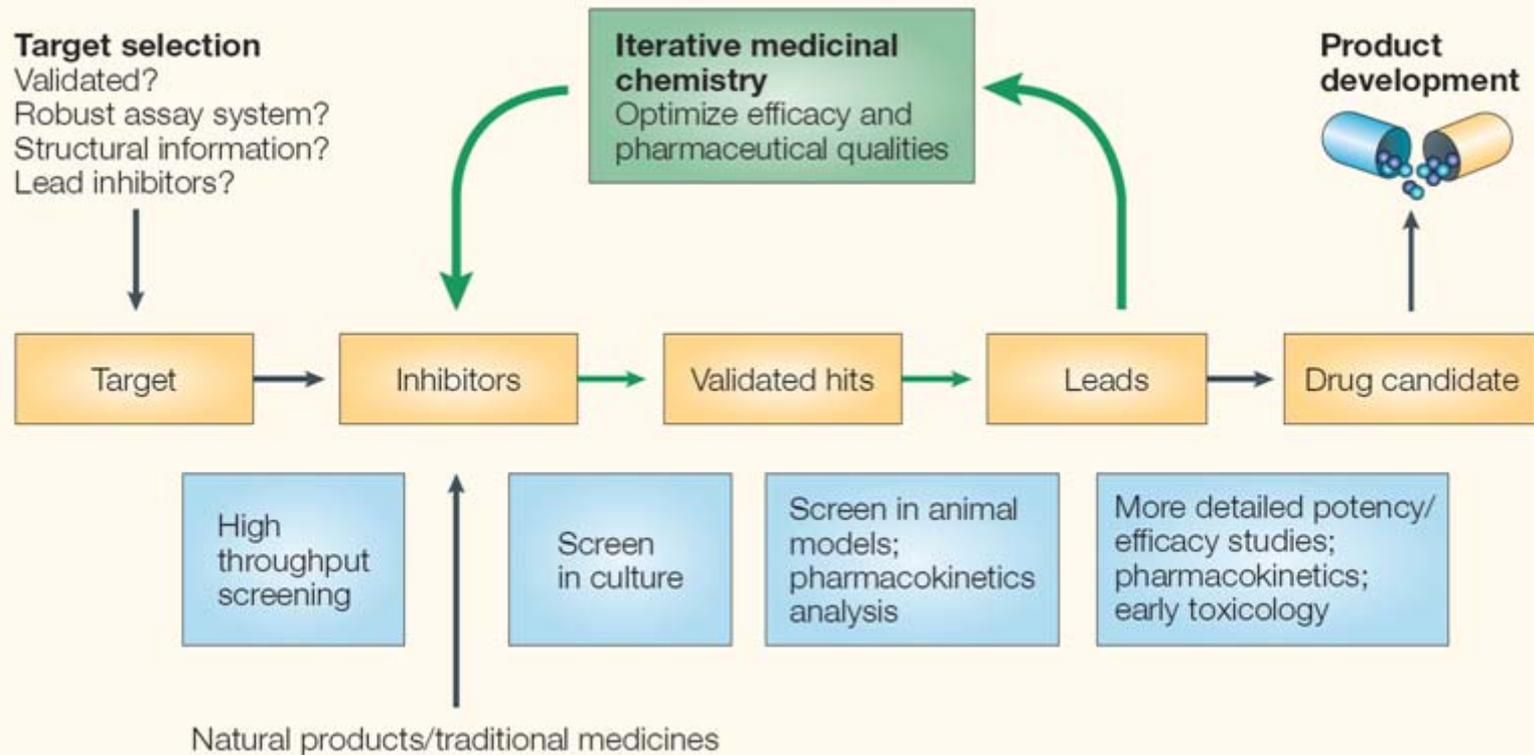


Production for
clinical trials



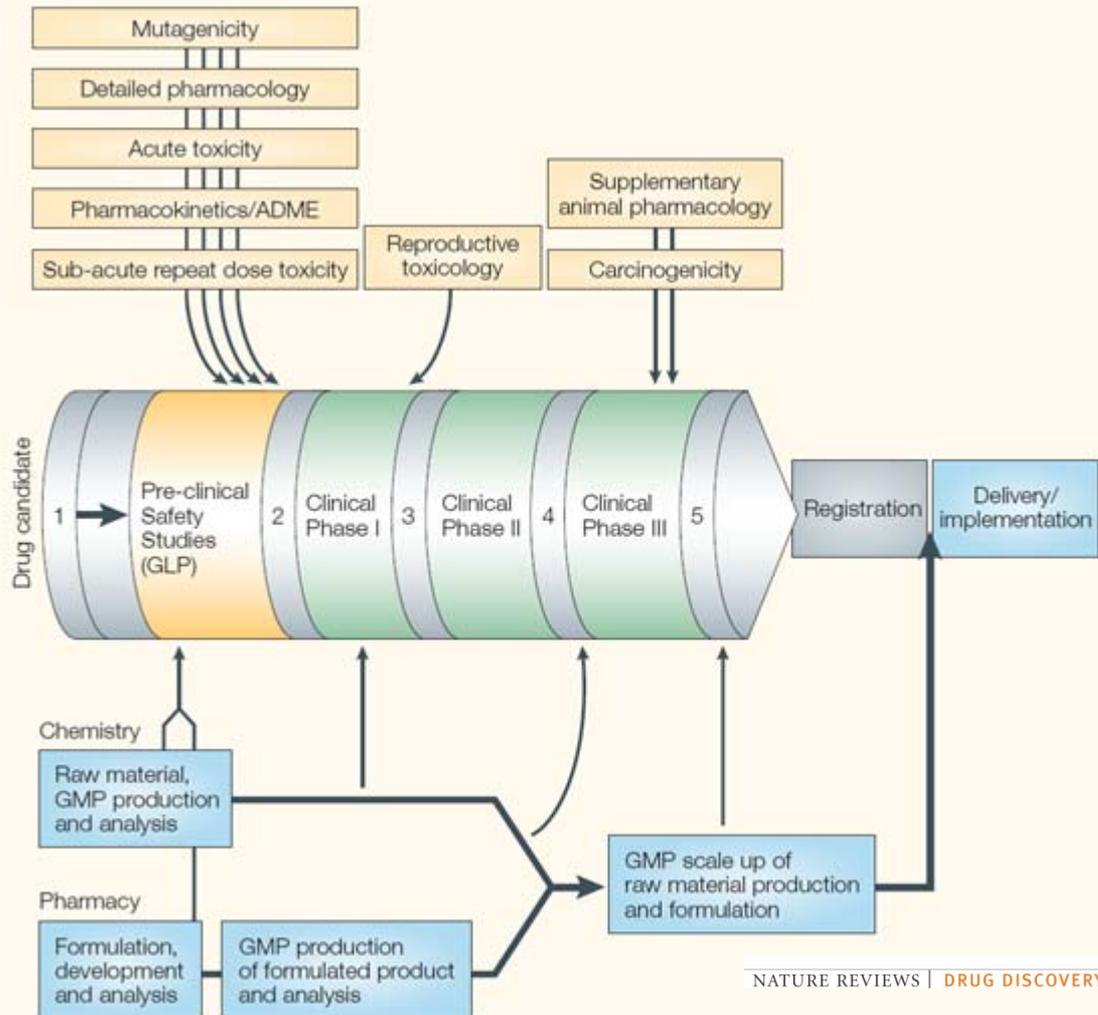
File IND

Coming up with a drug candidate



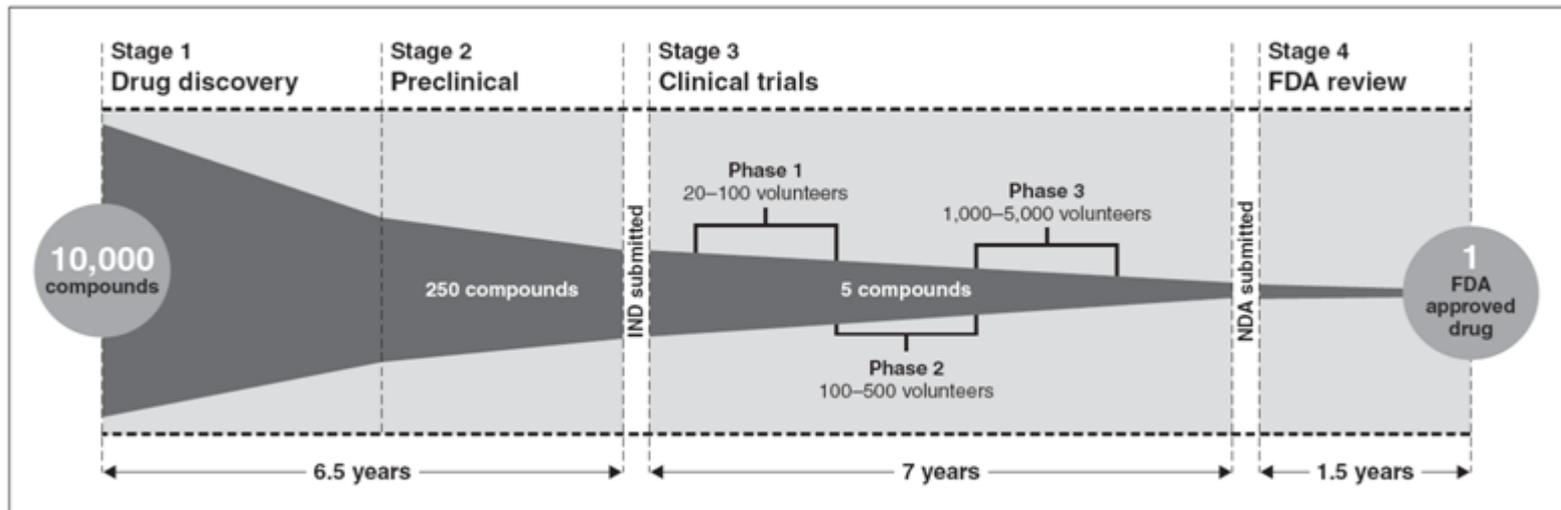
NATURE REVIEWS | DRUG DISCOVERY

Following up on that drug candidate



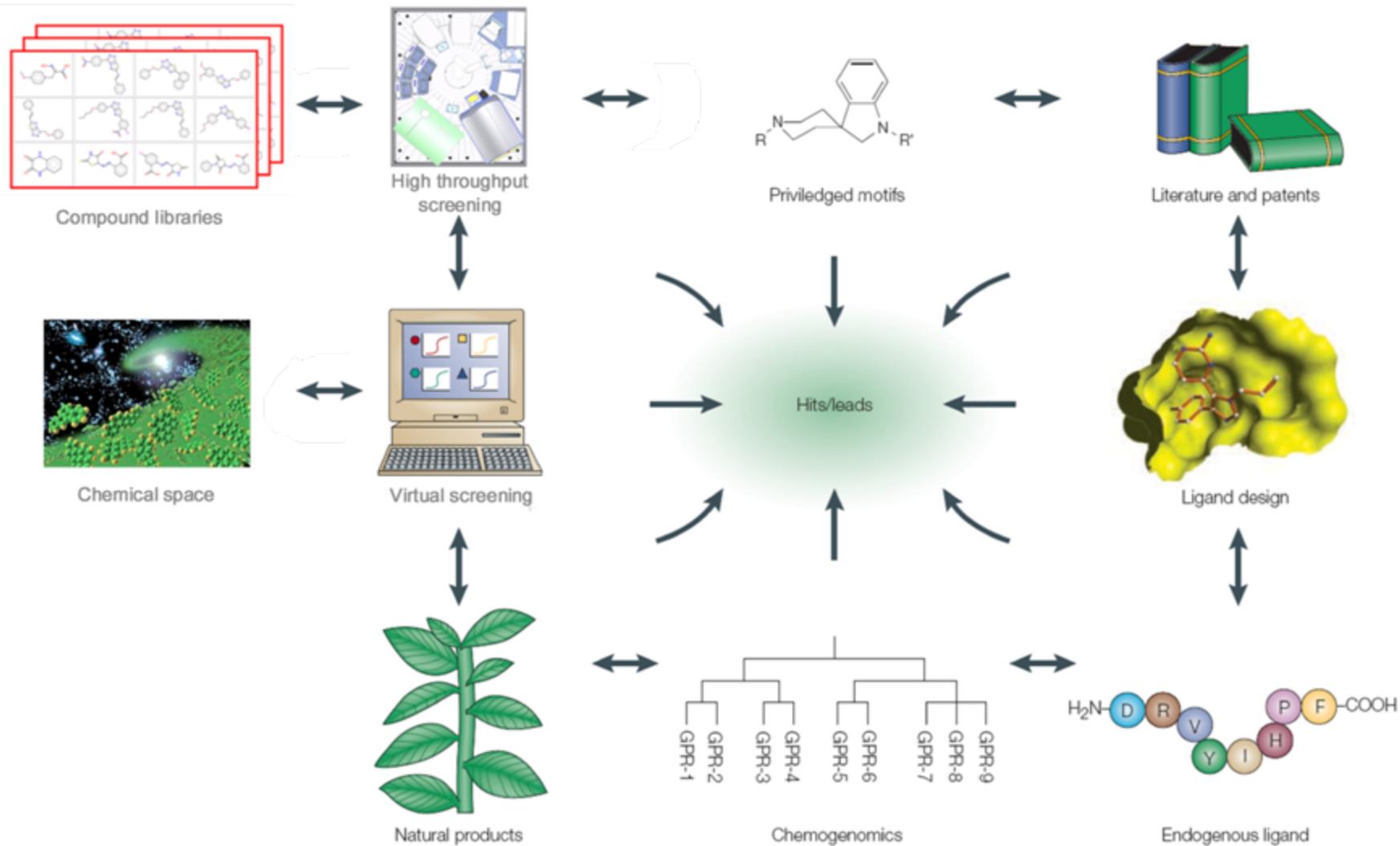
The entire process is slow and costly

Figure 1: The Drug Discovery, Development, and Review Process

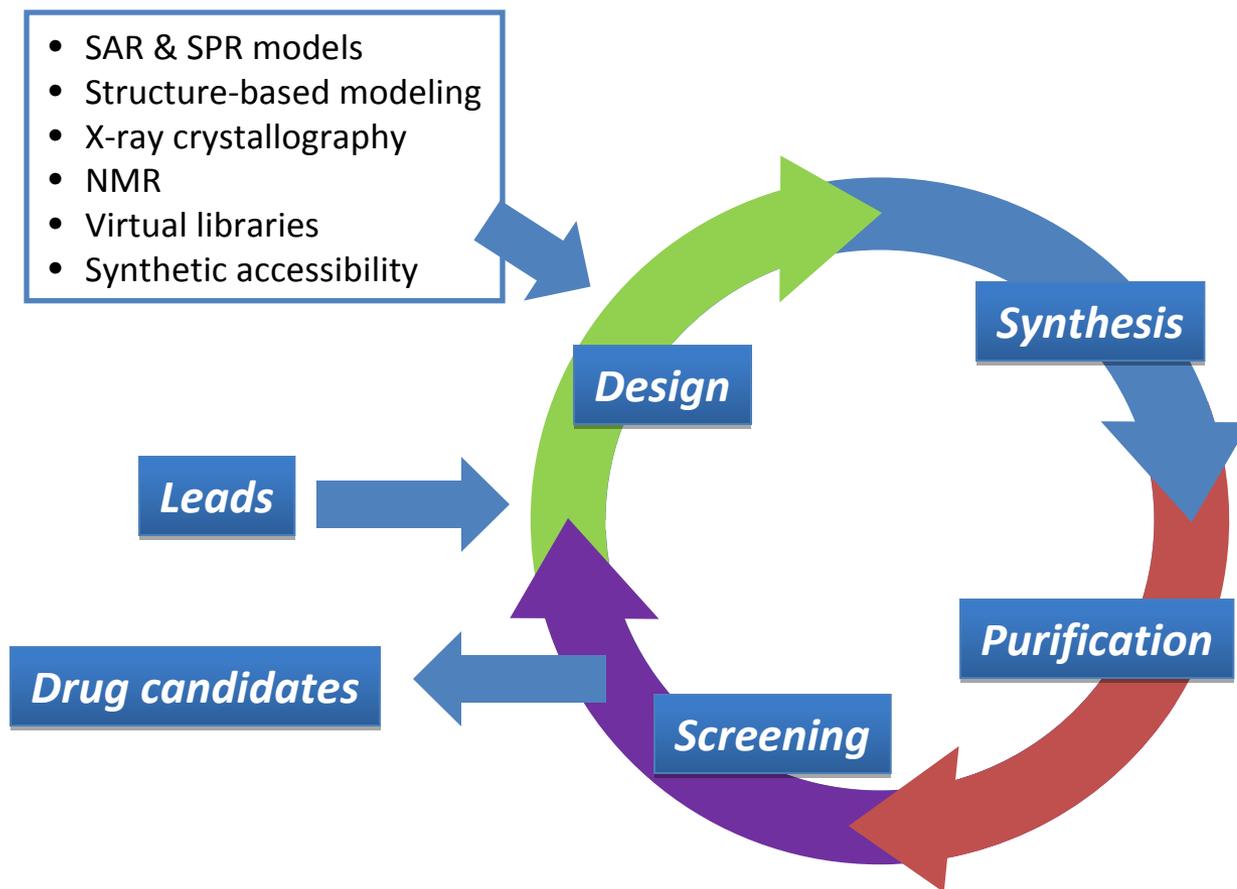


Source: Pharmaceutical Research and Manufacturers of America.

Hits and leads discovery



Lead optimization



Challenges in drug development

- Going after the wrong target
- Failure to identify hit and/or lead compounds
 - Limited chemical diversity of the screening library
 - Non-druggable targets
- Failure to optimize lead compounds
 - Hard to diversify and/or synthesize a series of analogs
 - Could not achieve the desired binding affinity/potency
 - Did not achieve the desired ADME and toxicity properties
- Low selectivity
- High off-target activity
 - Recent studies have suggested that this may not be such a bad thing (polypharmacology)
- Failure to identify a high-throughput synthesis route

Slow rate of drug discovery

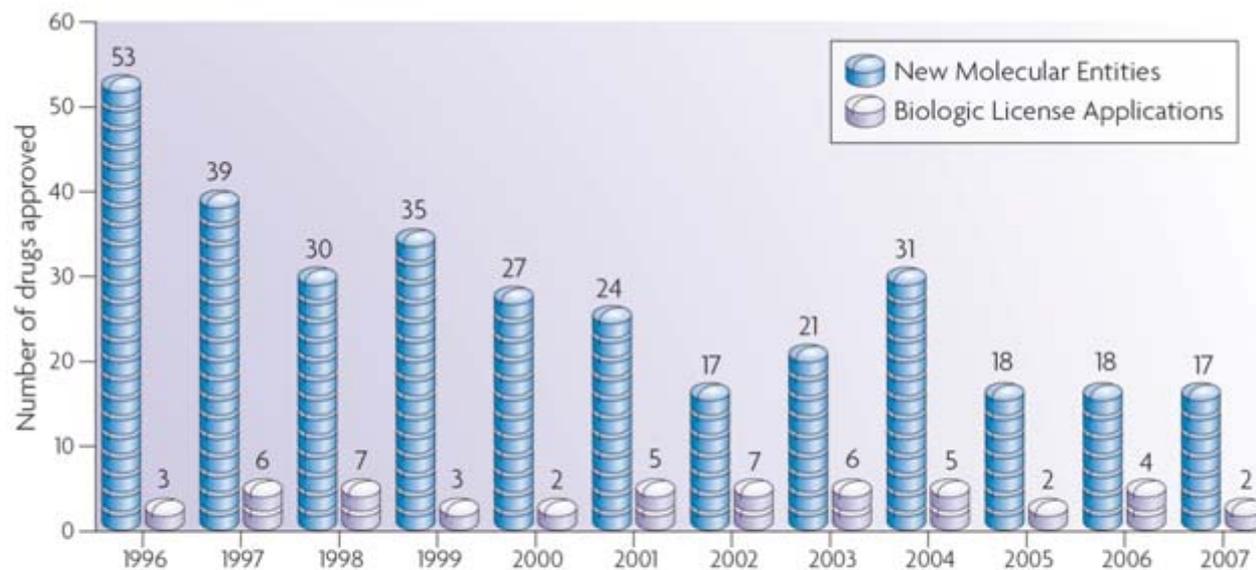


Figure 1 | **FDA drug approvals.** New molecular entities and biologic license applications approved by the US FDA by year.

NATURE REVIEWS | DRUG DISCOVERY

The druggable genome

- Studies have pessimistically estimated the size of the “druggable genome” to about 3,000 proteins.

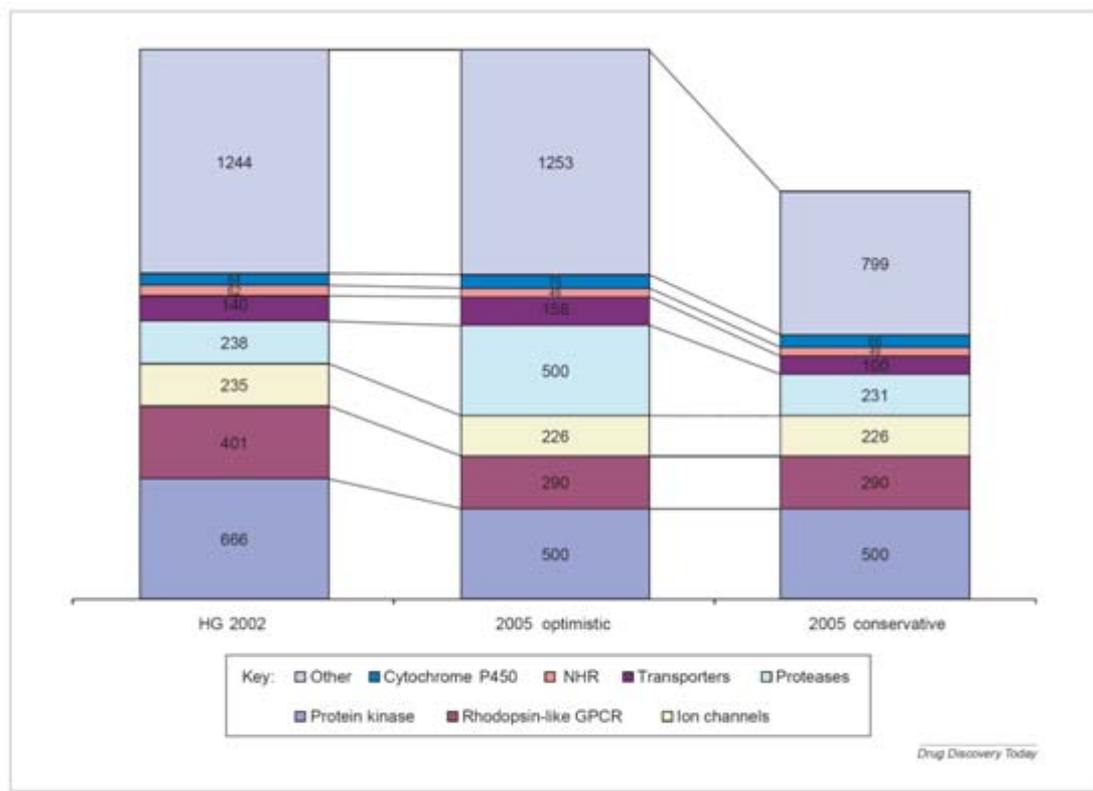


FIGURE 2

Optimistic and conservative estimates of the druggable genome compared with previous predictions [2]. Abbreviations: HG, Hopkins and Groom [2]; NHR, nuclear hormone receptors.

Chemical genetics (genomics)

Chemical Genetics (Genomics): The research field that is designed to discover and synthesize protein-binding small organic molecules that can alter the function of all the proteins and use them to study biological systems.

(National Institute of General Medical Sciences)

- Chemical genetics has emerged as a promising new approach for studying biological systems.
- It has a number of key advantages over existing approaches based on molecular genetics:
 - small molecules can work rapidly,
 - their action is reversible,
 - can modulate single functions of multi-function proteins,
 - can disrupt protein-protein interactions, and
 - if the target is pharmaceutical relevant, it can lead to the discovery of new drugs.

Where we are and where we will like to go

- Recent studies have catalogued selective small molecule modulators for about 1,500 protein targets.
 - This represents a fraction of the estimated 100,000 protein functions in human cells.
- Analysis of FDA databases has found that small molecule modulators exist for about 800 pharmaceutical relevant proteins.
 - This pails in comparison to the size of the druggable genome.
- Finding selective small-molecule modulators (probes) is a laborious and multi-step process whose success depends on a number of different factors.
 - Discovery and optimization process is similar in nature to that used for drug development
 - This needs to change in order to realize the promise of chemical genetics

Differences between drug and probe development

- Drug molecules are designed for *in vivo* use, whereas probe compounds are designed for *in vitro* use
 - Reduces the ADME-type properties that the probe compounds must satisfy
- Probe compounds must be selective with limited (ideally none) off-target activities
 - Drug compounds can have off-target activities as long as they do not lead to undesirable side effects (e.g., toxicity)

Challenges in probe development

- Finding hits for *novel* protein targets
 - Existing chemical libraries are biased/optimized for drug discovery
 - They are designed to cover only a small number of protein families
 - Pharmaceutically relevant targets (druggable & therapeutically relevant)
 - Little is known as to what compounds will bind to proteins outside these sets of targets
- Lack of 3D structural information for most of the proteins
 - Limits the set of methods that can be used to optimize leads
- Need for selectivity
- Analysis of phenotypic/cell-based assays to identify the targets of small molecules
 - This approach represents a ligand-focused target discovery (*target fishing*)

Data sources

- Chemical compound libraries
 - Several millions of compounds (PubChem contains ~19M compounds)
 - Virtual libraries can be easily generated that have several million of compounds
- Screening results
 - High-throughput screening assays, each containing initial screening results for 20K-200K compounds at a single concentration
 - Confirmatory assays usually involving less than a few thousand compounds
 - Dose-response assays for relatively small number of compounds (< 500) at different levels of concentrations
 - High-content screening assays providing spatial and quantitative information (via microscope images) of the cell's phenotype
- Target-ligand affinity information extracted from publications
- Crystallographic & NMR information from *in vitro* experiments
- High-throughput synthesis reactions

Analysis requirements & challenges

- Methods capable of analyzing the topological and/or geometric nature of the compounds in order to build accurate models to relate the biological activity of a compound to its own structure
 - Structure-activity-relationship (SAR) models
- Effective *in silico* docking-based virtual screening methods
 - Accurate scoring functions and docking protocols that allow for ligand and target flexibility and can account for genetic variations of the protein targets
- Methods to predict the synthetic accessibility of compounds and identify efficient synthesis paths
- Computational methods that can deal with noisy, sparse, and incomplete data
 - Screening results are inherently noisy and only available for a tiny fraction of the existing compounds
 - Negative information is often not available and/or non-reliable whereas positive information is available for a small number of compounds
- Methods to analyze the diversity of existing libraries and their relation to the areas of the chemical space that are of relevance for drug and probe compounds
- Methods to analyze high-content screening assays to identify the protein targets response for phenotypic changes being induced by small molecules

Ultimate goal:

In silico bridging of chemical and biological spaces

- Novel methods are needed to establish the connections between proteins (biological space) and their ligands (chemical space) in order to accelerate drug/probe development
 - Improve the accuracy of structure-activity-relationship models
 - Compound selection for screening library design
 - *De novo* ligand design guided by these connections
- Ultimate goal:
 - Given a novel protein target, construct a focused screening library containing either existing or easy to synthesize *de novo* ligands that contains a potent and selective ligand
 - An assay will only be used to identify/validate the best ligand and that's it!

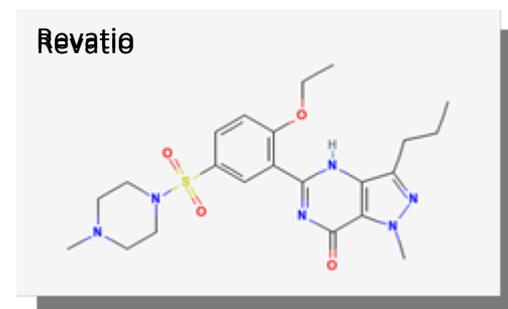
Underlying Hypothesis: Proteins that have similar binding sites should have ligands that are similar and similar ligands should bind to proteins with similar binding sites.



MATHEMATICAL VIEW

Where are the graphs/networks?

- Molecular graphs
 - The function of a compound is largely determined by its structure
 - Even though different approaches have been developed for capturing/modeling the structure of a compound, the most effective approaches focus on the compound's molecular graph
 - Atoms become vertices and bonds become edges
 - Thus, from a data representation and computational stand-point, a library of compounds is nothing more than a set of (small) graphs
- Compound co-activity network
 - Formed by connecting the compounds that share targets
- Target co-ligand network
 - Formed by connecting via an edge the targets that share ligands
- Target-ligand network
 - This is a bipartite network formed by viewing the target-ligand activity matrix in the form of a bipartite graphs
 - The targets and the ligands from the two set of vertices and for each target-ligand pair there is an edge between the corresponding target and ligand



How molecular graphs are used

- Retrieve compounds that are similar to a query compound
 - Query is performed by comparing the structure of a pair of molecules and assigning a quantitative score that captures their degree of similarity
- Build SAR models based on their structure
 - Based on the presence/absence of certain chemical fragments (substructures/subgraphs)
 - Based on kernels functions defined on their graph's structure
- Coming up with new compounds by combining a set of compounds (usually two) or by breaking it apart along an edge
 - Often synthesis and fragmentation is driven by reactions executed forward or backwards

Molecular graphs: Mathematical/Computational challenges

- Determining the similarity between small graphs
 - Maximal common subgraph
 - This is the traditional mathematical approach for comparing two graphs; however, its use for molecular graphs is somewhat limited
 - It has been found to be too rigid
 - This can change if better methods are developed for determining approximate common subgraphs (e.g., graph edit distance)
 - Descriptors
 - Each graph is represented as a set/vector of substructure descriptors and two graphs are compared using set/vector similarity measures
 - Enables the easy identification of the fragments that are associated with activity
 - It is the most widely used approach
 - Random-walk based similarity methods
 - E.g.,

$$\text{sim}(G_1, G_2) = \sum_{(v_1, v_2) \in V_1 \times V_2} p_1(v_1) p_2(v_2) \text{sim}_L(l(v_1), l(v_2)),$$

where $l(v_1)$ is a sequence of vertex labels of a random path starting at node v_1 , sim_L is a similarity function defined on the label sequences, and $p(v_1)$ is the probability of starting the random walk at vertex v_1 .

Challenges with descriptors

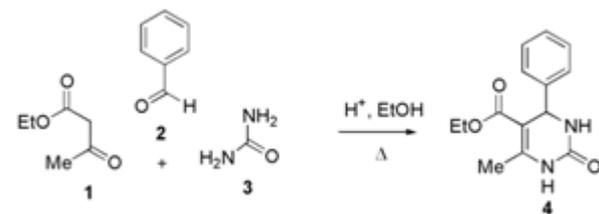
- What should be used as descriptors?
 - Bounded-length paths, atom-centered bounded-length rings, bounded-size trees, cycles, frequent subgraphs, bounded-size subgraphs, etc.

	AF	TF	PF	
GF	=	>	>	GF: Graph fragments, AF: Acyclic fragments, TF: Tree fragments, PF: Path fragments
AF		>	>	
TF			>	

Wale N, Watson IA, Karypis G., *Comparison of Descriptor Spaces for Chemical Compound Retrieval and Classification*, Knowledge and Information Systems (KAIS) Journal, Vol 14, No. 3, pp. 347–375, 2008.

- How do you encode domain information?

- Atom types, bond types, ring structure, scaffolds, synthesis reactions, etc.



- How to find them efficiently?

- Fragment-based descriptors require the efficient determination of the fragments and their number of distinct embeddings in the graphs.

Efficient discovery of graph fragments

- More challenging than the most commonly solved problem of frequent (closed/maximal) subgraph mining in sets of graphs.
 - There should be no (or extremely small) frequency cut-off
 - There is a need to determine the number of embeddings of each fragment
- These requirements lead to a problem formulation that is more similar to that of finding graph-based patterns in a single large graph (or network), as this problem also requires the enumeration of all of the embeddings
 - This is not a very-well studied problem and there is a need to develop better methods
- Also, any improvements here will have broader applications related to graph searching and retrieval as fragments have been shown to provide a powerful indexing scheme
 - i.e., inverted indexes for graphs

An example of potential gains and future directions

AFGEN 1.0

GF statistics for MLSMR.

Fragment Size	t[sec]	#f
4	255	2991
5	481	12494
6	1051	48917

MLSMR library contained 224,278 compounds; t is the running time in seconds on a 64-bit Intel Xeon 2.33GHz; $\#f$ is the total number of fragments that occur in more than one compound.

AFGEN 2.0

GF statistics for MLSMR.

Fragment Size	t[sec]	#f
4	56	3510
5	66	15147
6	84	61326
7	120	233928

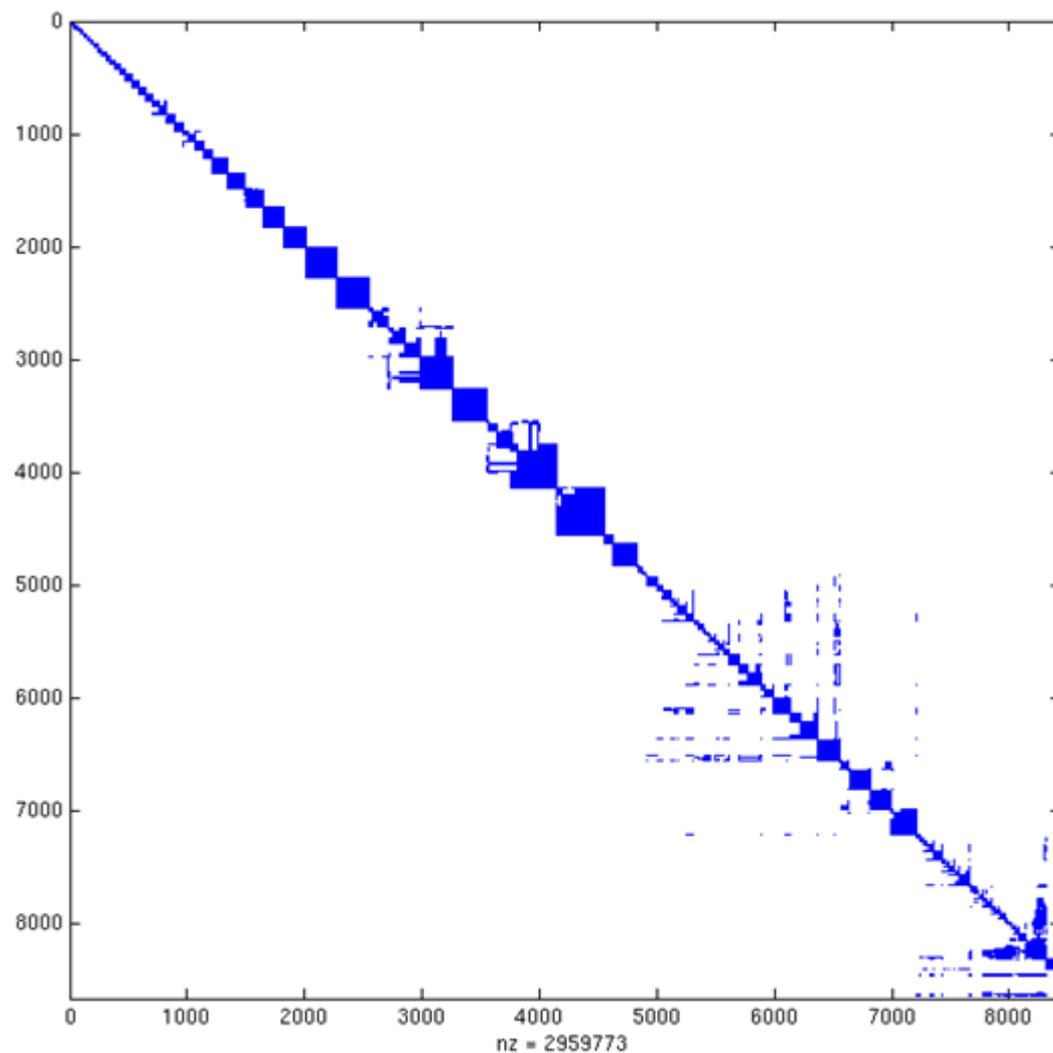
MLSMR library contained 245,870 compounds; t is the running time in seconds on a 64-bit Intel Xeon 2.33GHz; $\#f$ is the total number of fragments that occur in more than one compound.

<http://glaros.dtc.umn.edu/gkhome/afgen/overview>

• Future improvements?

- Smarter algorithms for traversing and putting together the fragments
 - E.g., construct/generate/grow the fragments by using paths instead of edges
- Incorporate a deeper graph-theoretic understanding of the problem
 - E.g., use of orbits of graphs to take advantage of symmetry in order to reduce the fragments that need to be explicitly enumerated

Compound co-activity network



Uses of the compound co-activity network

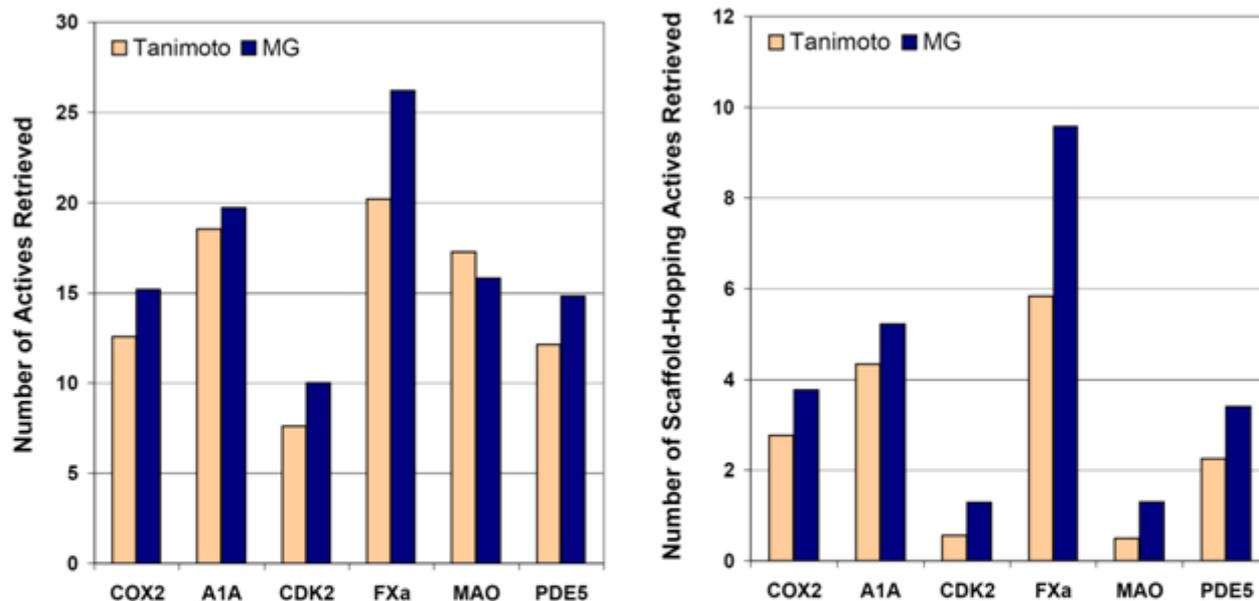
- As a planning tool to generate a diverse screening library that covers the currently covered portions of the chemical space
 - E.g., something like a maximal independent set...
- It can help identify the areas of the sparsely covered areas of the chemical space and thus focus future library developed towards poorly covered areas
 - E.g., low-degree compounds can form the seeds for such rational chemical space expansion
- Identify *frequent hitters* and eliminate them from consideration
- Identify compounds with different scaffolds that can potentially share the same biological activity (scaffold hopping). These compounds can potentially be better for optimizing ADME+tox properties
 - E.g., compounds connected via multiple short paths, manifold distance, etc.

Challenges associated with the co-activity network

- Current network is extremely sparse and incomplete
 - Incompleteness is due to the limited amount of available experimental screening data
 - This unfortunately will not change as we move forward
- Methods need to be developed that can populate the network
 - This is a huge *missing edge prediction* problem
 - Possible solution approaches:
 - *In silico* screening via supervised learning methods or molecular docking for each of the current targets
 - Computationally demanding and limited to only existing targets
 - Development of robust methods to determine the statistical significance of molecular graph similarity scores (like existing approaches for protein sequence comparisons)
 - This will require the development of new methods that account for how the similarity is computed, how compounds are created, and potentially for their 3D conformations
 - There is very little work in this area, but it will have wide applications well beyond drug and probe discovery
- Methods to analyze these *hybrid* networks (real+predicted) that take into account the errors associated with the predictions
 - Relation to graph-based semi-supervised learning methods

Co-activity networks – Benefits from even simple-minded solutions

Figure 1: Performance of indirect similarity measures.

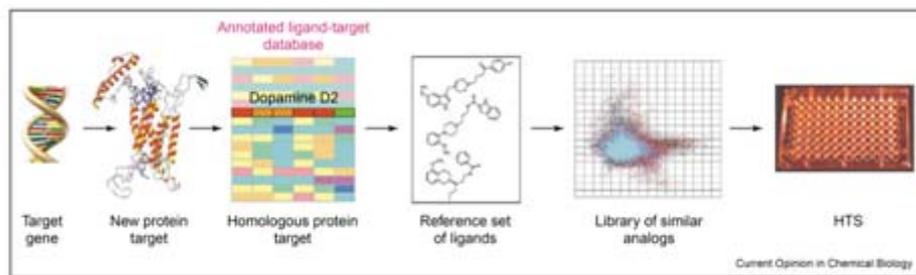


Tanimoto indicates the performance of similarity searching using the Tanimoto coefficient with extended connectivity descriptors; MG indicates the performance of similarity searching using the indirect similarity approach on the mutual neighbors graph.

Wale N, Watson IA, Karypis G., *Indirect similarity based methods for effective scaffold-hopping in chemical compounds*, J Chem Inf Model. 2008 Apr;48(4):730-41.

Use of the target co-ligand network

- The well-connected clusters of the network indicate sets of targets whose binding sites share key characteristics that are relevant to ligand binding
- Create screening libraries focused to a particular target by utilizing the ligands of other targets in the network
 - This class of methods are referred to as *chemogenomics*

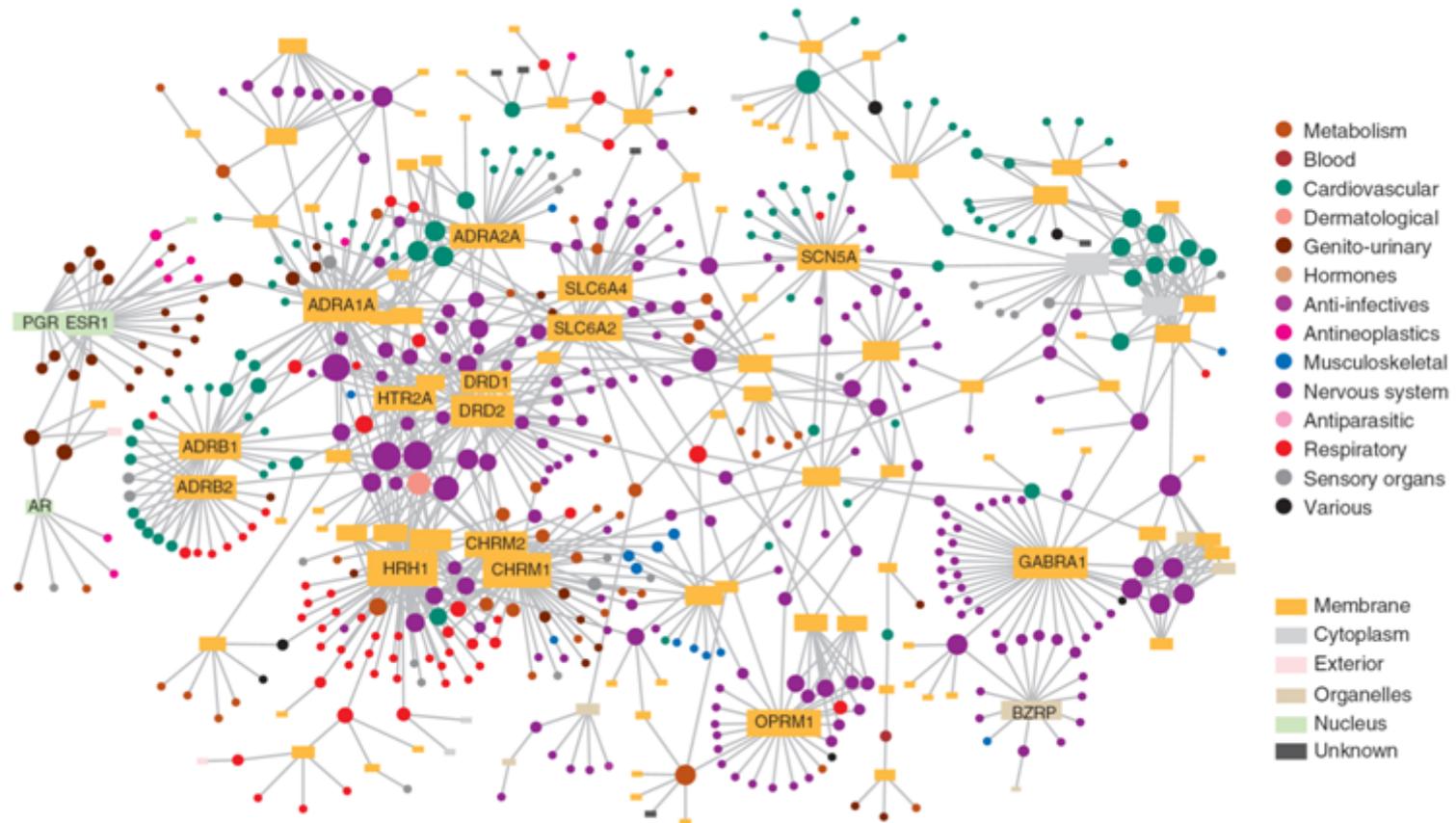


- Improve the quality of SAR models by using ligands of related targets
 - E.g., there is a high probability that the *unscreened* ligands of related targets will also be ligands for the target in question
- Analyze how ligand-binding related connections relate to other networks involving the proteins
 - evolutionary, structural, functional, etc.

Challenges associated with target co-ligand networks

- Identifying which connections are reliable for information transfer
 - Intrinsic properties of the proteins
 - E.g., sequence, structure, function, etc.
 - Properties of the actual shared ligands
 - E.g., molecular structure, fragments, number, etc.
- Determining the best way to utilize this information, especially when are used to improve the quality of SAR models
 - This can be thought of as an instance of semi-supervised learning in which the ligands of the related targets become the pool of the unlabeled instances
- The reliability determination problem can be viewed as an instance of learning which parts of the graph should be used to semi-supervised learning

The target-ligand network



NATURE BIOTECHNOLOGY

Uses and challenges

- Uses
 - Identify new ligands for targets and new targets for ligands
 - Identify relations between the chemical and biological spaces in order to build *target-hopping* models so that given a novel protein target be able to
 - Predict which compound will bind to it
 - Use it to guide synthesis of *de novo* ligands
- Challenges
 - Analyzing the network in order to build the target hopping models
 - Edge-prediction problem that needs to take into account
 - The structure of compounds and the structure of the proteins' binding sites
 - In many cases information about the protein will be computationally determined
 - The model needs to generalize to novel proteins targets

Summary of mathematical challenges

- Methods to find patterns in graphs and identify their complete set of embeddings
- Methods to determine the similarity between graphs and assess their statistical significance
 - Both global (are the graphs similar) and local (is one graph contained in the other) type of similarities
- Methods to predict edges (relations) in the networks that take into account the network structure itself and intrinsic information associated with the nodes
- Methods to find well-connected sub-networks that are tolerant to errors associated with the predicted nature of the links
- Methods to integrate information from different networks overlaid on the same set of nodes

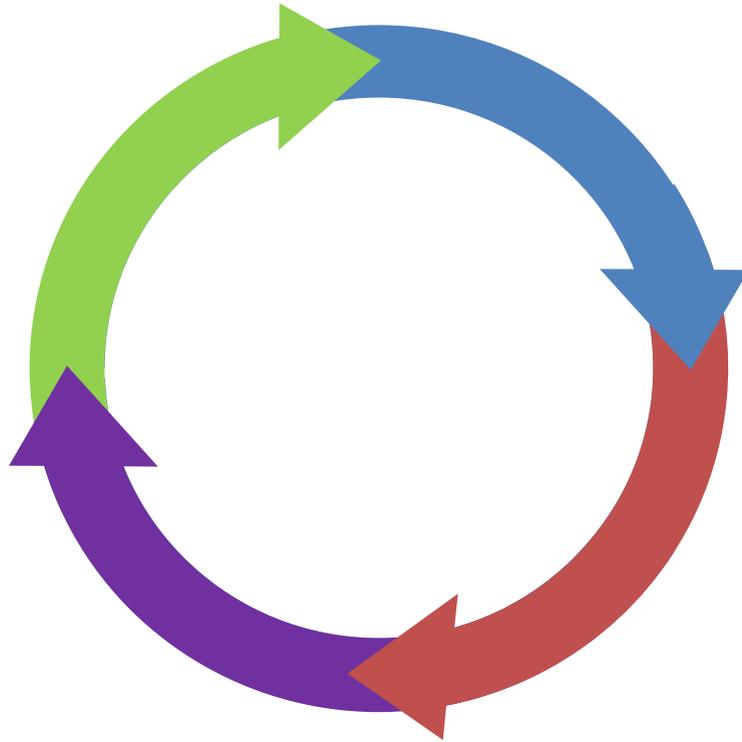
THANK YOU!

BACKUP SLIDES

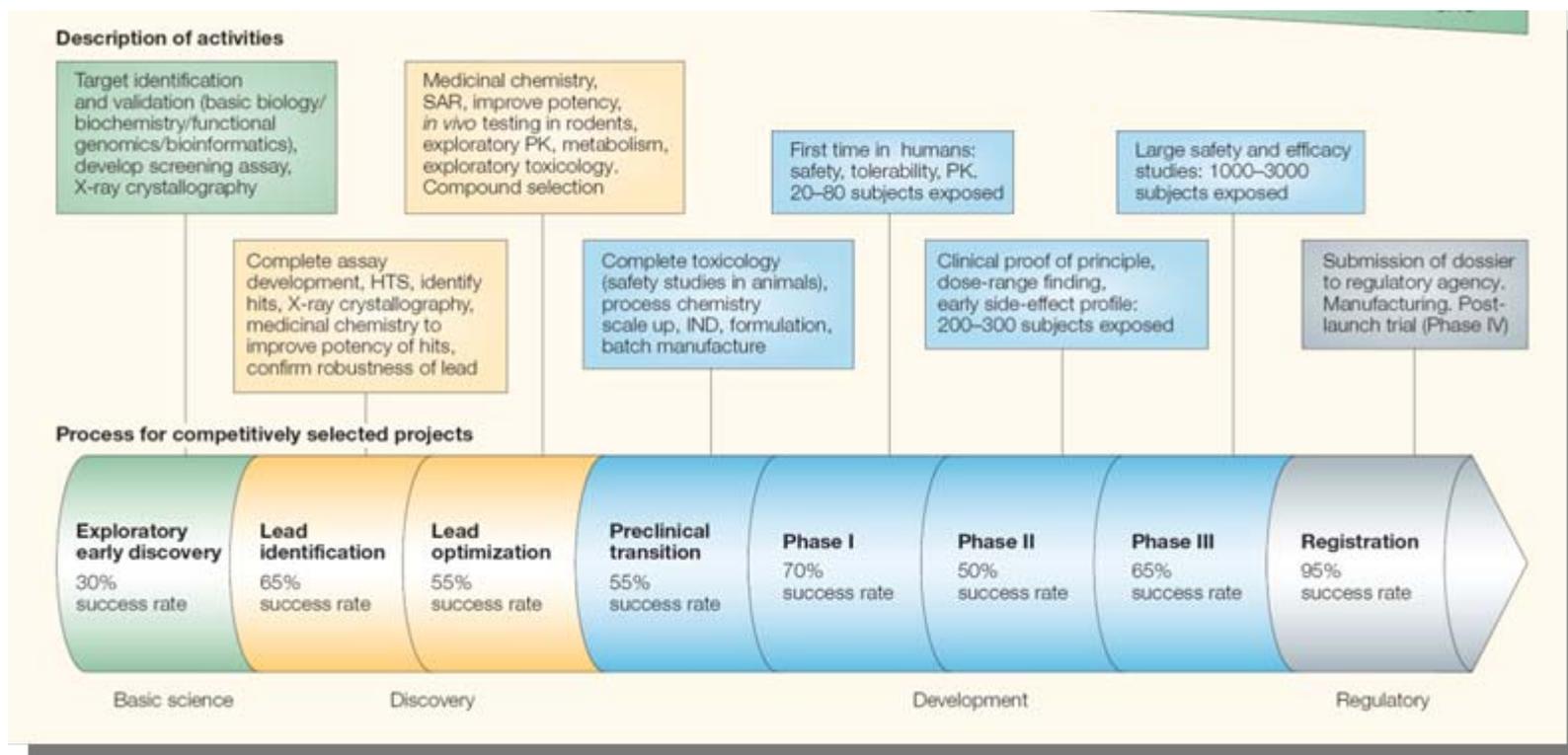
Problem Definition

Given one or more chemical compounds that have been experimentally determined to possess a desired biological activity, the goal is to find other compounds in a database that have similar bioactivity.

- Depending on the underlying hit/lead discovery process, there are two virtual screening approaches.
 - Classification (supervised learning problem)
 - Retrieval (unsupervised learning problem)



Failure rates



Where are the graphs/networks?

- Molecular graphs
 - The function of a compound is largely determined by its structure
 - Even though different approaches have been developed for capturing/modeling the structure of a compound, the most effective approaches focus on the compound's molecular graph
 - Atoms become vertices and bonds become edges
 - Thus, from a data representation and computational stand-point, a library of compounds is nothing more than a set of (small) graphs
- Compound co-activity network
 - Formed by connecting the compounds that share targets
- Target co-ligand network
 - Formed by connecting via an edge the targets that share ligands
- Target-ligand network
 - This is a bipartite network formed by viewing the target-ligand activity matrix in the form of a bipartite graphs
 - The targets and the ligands from the two set of vertices and for each target-ligand pair there is an edge between the corresponding target and ligand

