

The Proximal Lilly Collection

Bridging Chemical Synthesis Potential with Discovery Chemistry

C. A. Nicolaou*, I. Watson, J. Wang

Computational Chemistry & Chemoinformatics – C3

Lilly Research Laboratories, Eli Lilly & Company,
Indianapolis, IN 46285, USA

Overview

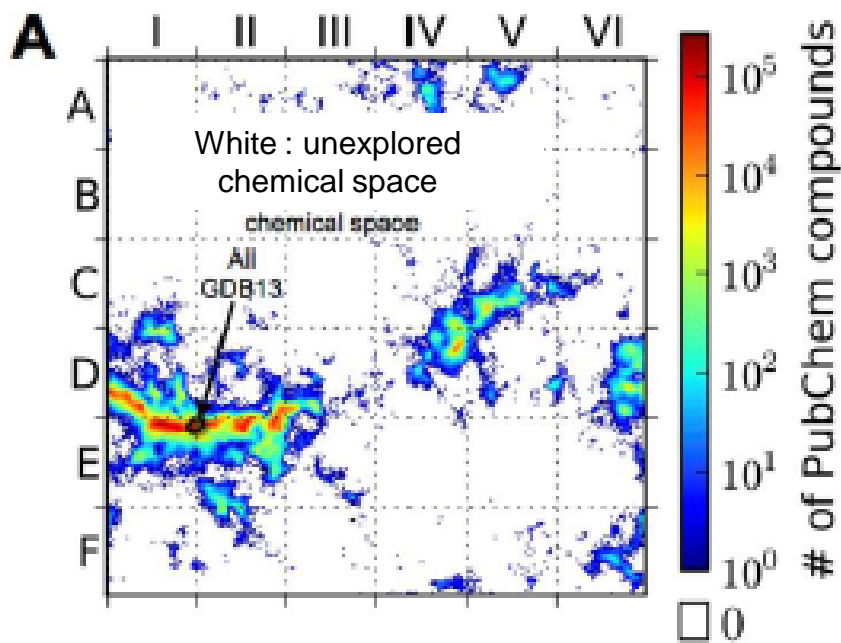
- Chemical Space - Background
 - Lilly Motivation
 - Automated Synthesis Lab
 - Proximal Lilly Collection Initiative
 - Design consideration
 - Implementation
 - Interface
 - PLC Idea to Data pilot project
-

Discovery Paradigm



Small Molecule Universe Overview

Lilly compound collection = 10^6



- Current compound collections represent a minute fraction of the small molecule universe both in size and diversity
- Opportunities for novelty
- What can we do about it?

Drug-like molecule space = 10^{63}
(estimates)

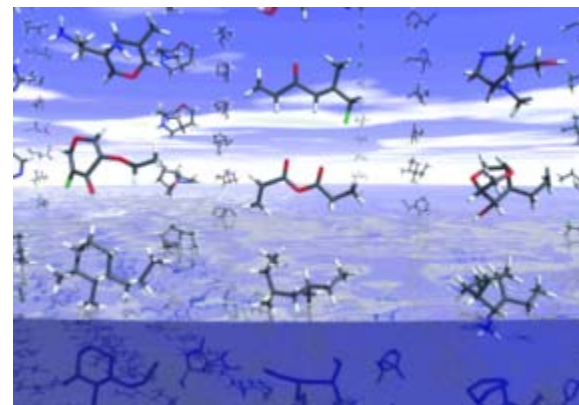
Chemical Space Ventures

- Initiatives

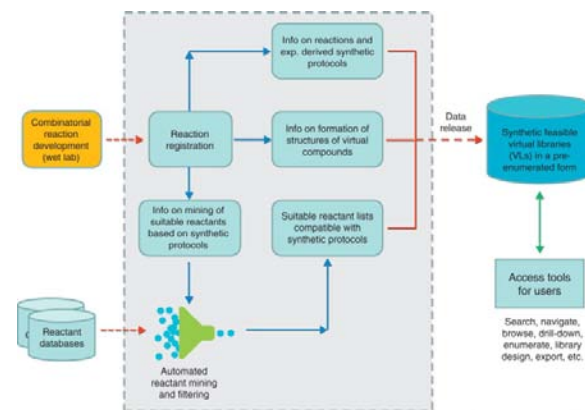
- Chemical space mapping
 - e.g. GDB-13
- De novo design
- Structure enumeration via chemical reactions
- ...

- Considerations

- Exploration Vs exploitation
- Diversity/novelty Vs synthetic feasibility
- Usefulness



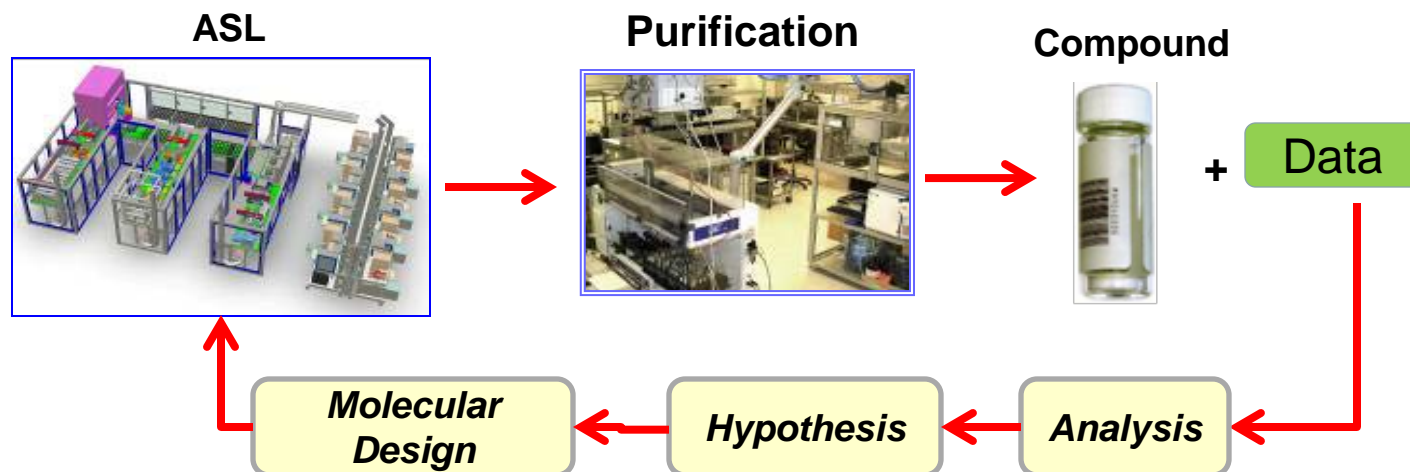
*Reymond, 2009



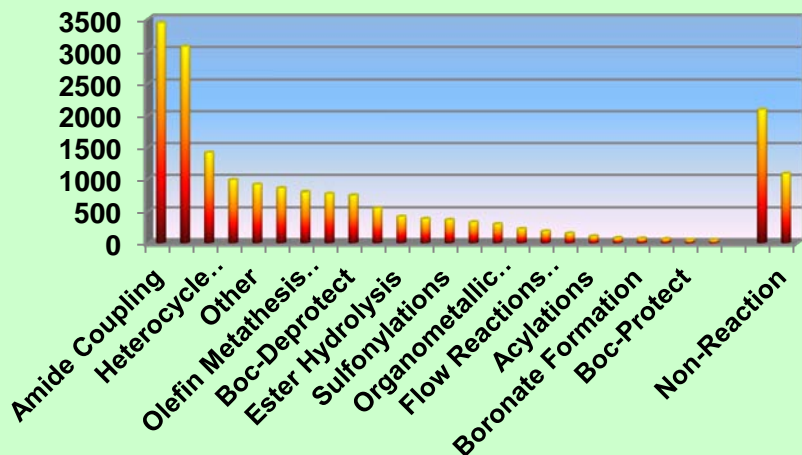
*Peng, 2013

Drug Discovery Today: Technologies

Automated Synthesis & Purification Lab



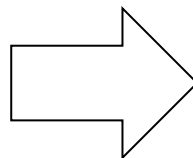
Sampling of Reactions



Lilly Opportunity & Motivation

Proximal Lilly Collection Vision

- ASL top 8 validated chemistries (>18,000 reactions)
- Combined virtually with reagents already physically existing at LCC



Results in >150 billion compounds with MW<500 that could be prepared in one step on the ASL - **today**

Lets conservatively assume only 10% of these compounds are feasible.* This leaves 15 billion compounds that are proximal to being prepared and can be created on demand. *This is the Proximal Lilly Collection*

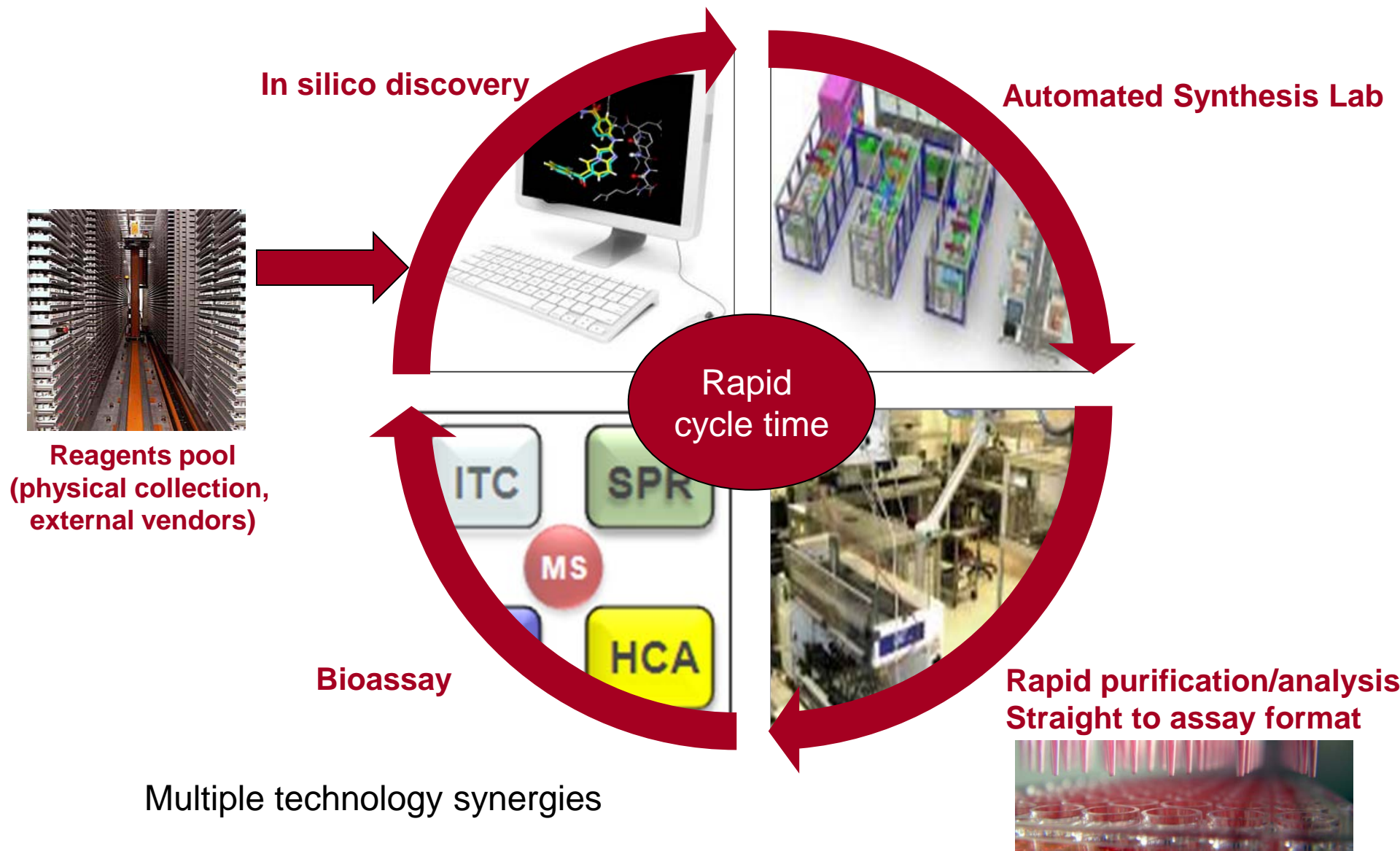


**As chemical cross-reactivity knowledge is added to the virtual combination, it is assumed that some reagent combinations will have functional group incompatibilities*

Proximal Lilly Collection - PLC

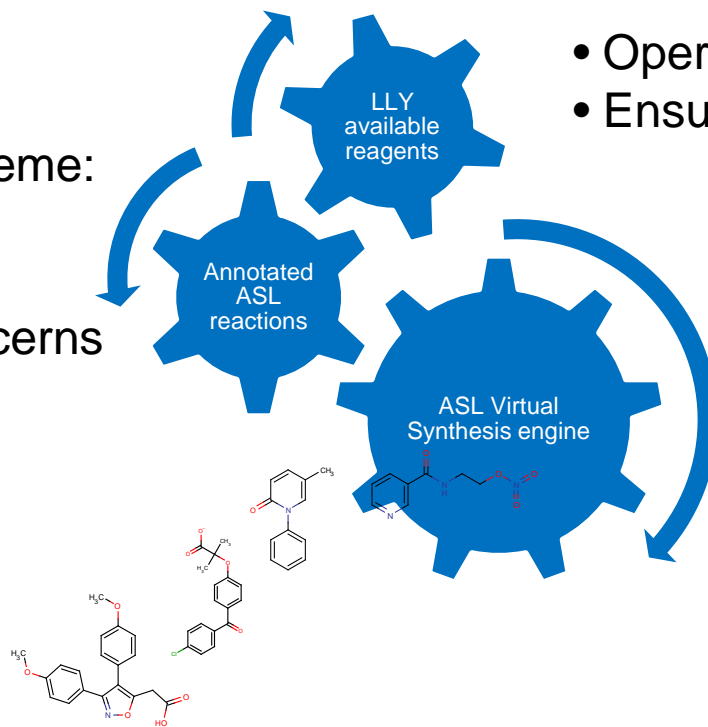
- Goal
 - Expand small molecule space available to Lilly discovery by accessing an “on demand” compound library
 - Provide 3rd compound source for discovery
 - LLY collection, external vendors
 - Facilitate exploitation of ASL for quickly discovering/selecting/synthesizing project related structures
- Specific Objectives
 - Define representative subset for Virtual Screening
 - Enable PLC-based hit expansion
 - Map synthetic route for seed compounds
 - ...

PLC Initiative – Idea2Data



Mapping PLC – Computational Strategy

- Lilly Annotated Reaction scheme:
(reaction description,
reagent rules)
- Address synthesizability concerns



- Operate in reagent space
- Ensure reagent availability

- Virtual Synthesis Engine
- Able to enumerate compounds feasible using annotated reactions and available reagents

**Proximal Lilly Collection
(PLC)**

- Intelligent sampling methods
 - Operating in reagent/reaction space
 - Avoid need for full PLC enumeration

PLC Reagents

- Readily available collections
 - ASL reagents
 - Lilly collection
 - Vendor catalogs (select)
- Optional pruning by:
 - Physicochemical properties
 - e.g size, ...
 - Inventory amount



Lilly Annotated Reaction Repository - LARR

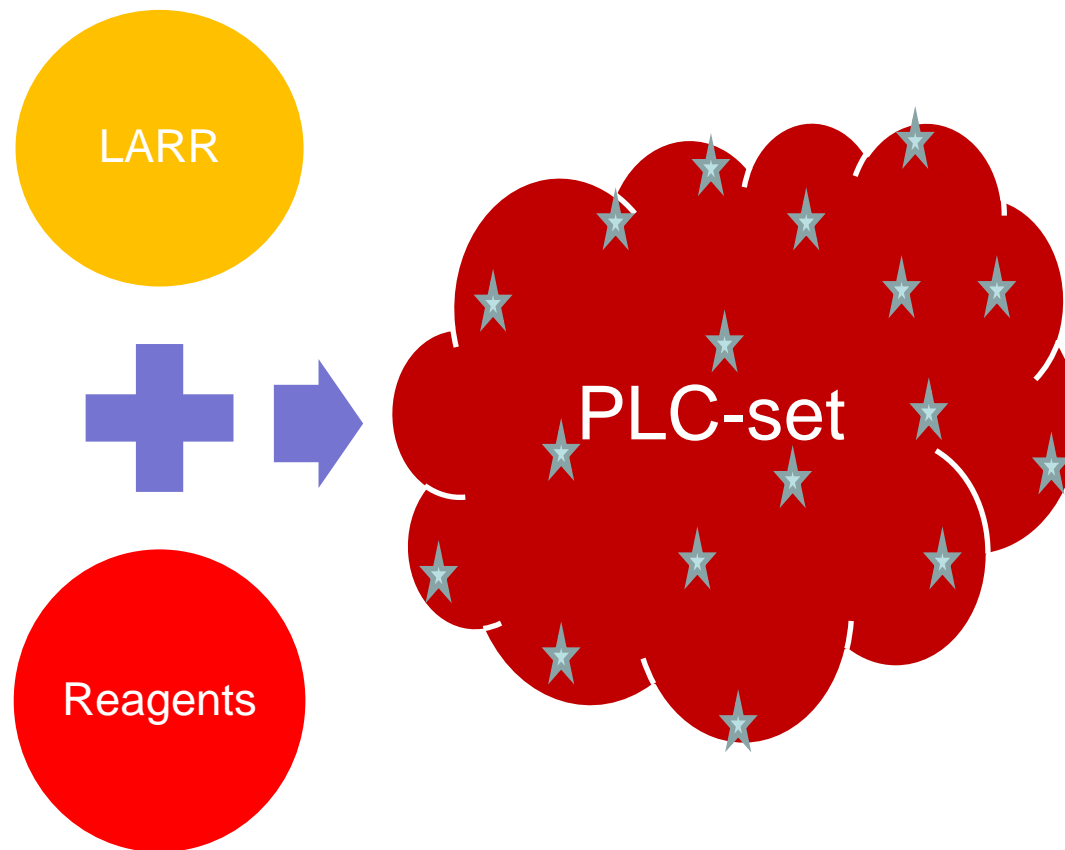
- Includes reactions performed on the ASL
 - Characterized by:
 - Reaction description
 - Reagent filter rules
 - Up to 50 rules describing in detail reagent structures that may/may not be used by the specific reaction
 - Reaction logistics
 - E.g. name, reagent order, etc.
 - Annotation: joint work by C3 and ASL chemists
 - Continuously updated
-

PLC-Virtual Synthesis Engine

- Virtually synthesize compounds using
 - LARR reaction(s)
 - Reagent pool(s)
 - Apply all available reaction annotations for reagent selection
 - Apply Lilly medicinal chemistry rules
 - Bruns & Watson, 2012
 - Optional:
 - Property/structure related filters
 - Sample representative subset
-

PLC Representative Subsets

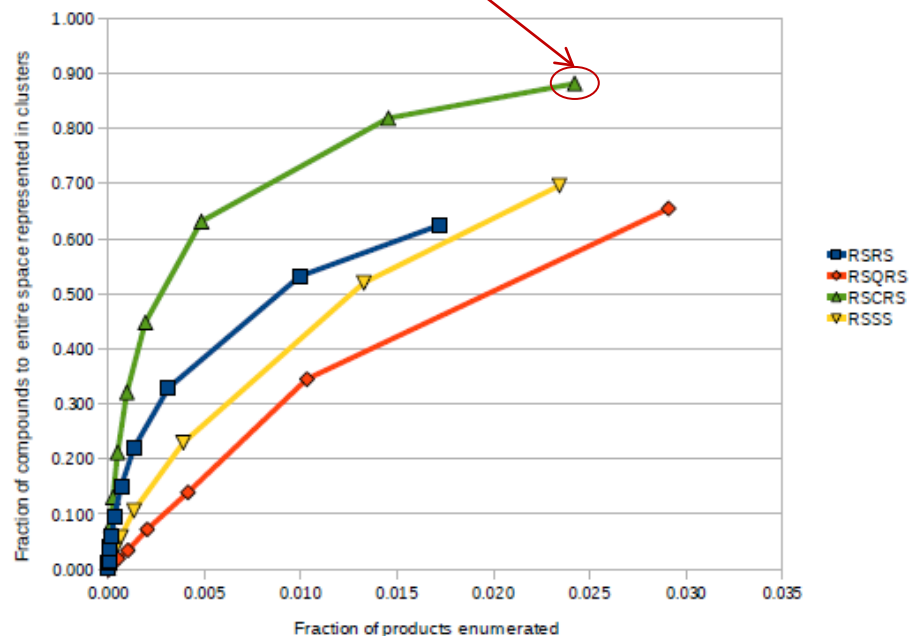
- PLC is very large
 - Troublesome to operate directly on it
- Approach:
 - Use a representative set
 - Operate in reagent space
- Compromise:
 - Efficiently mapping the entire space
 - Algorithmic complexity; map >> approximation



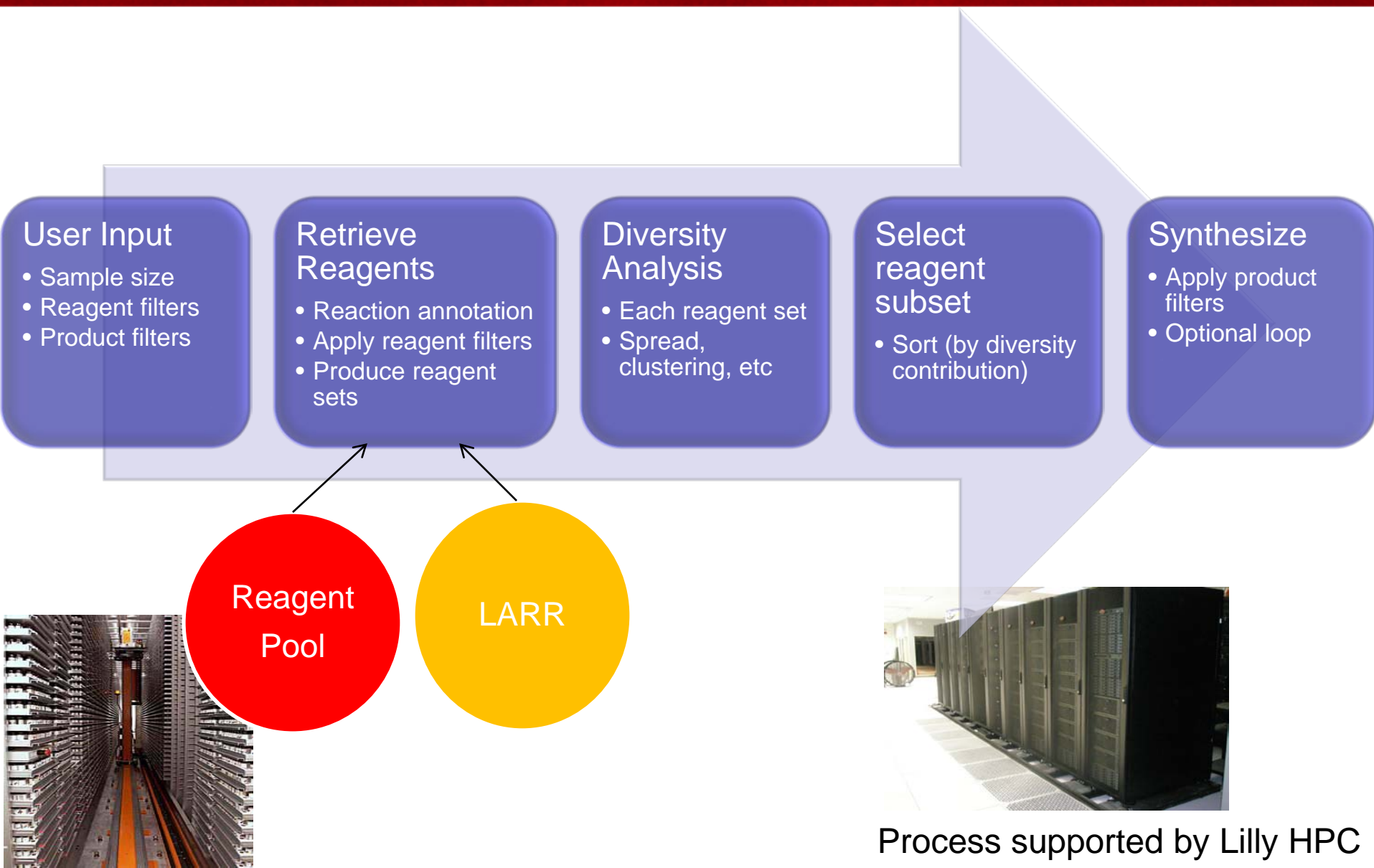
PLC Space Coverage

- Intelligent sampling
 - Operate in reagent space
 - Multiple methods implemented
 - Random
 - Cluster reagents + sample
 - Spread analysis + select
 - Spread analysis + cherry-pick
 - Iterative search-assess-synthesize
- Representation
 - Compound-based
 - Cluster-based
- Exploiting Brainiac
 - Lilly High Performance Computing system

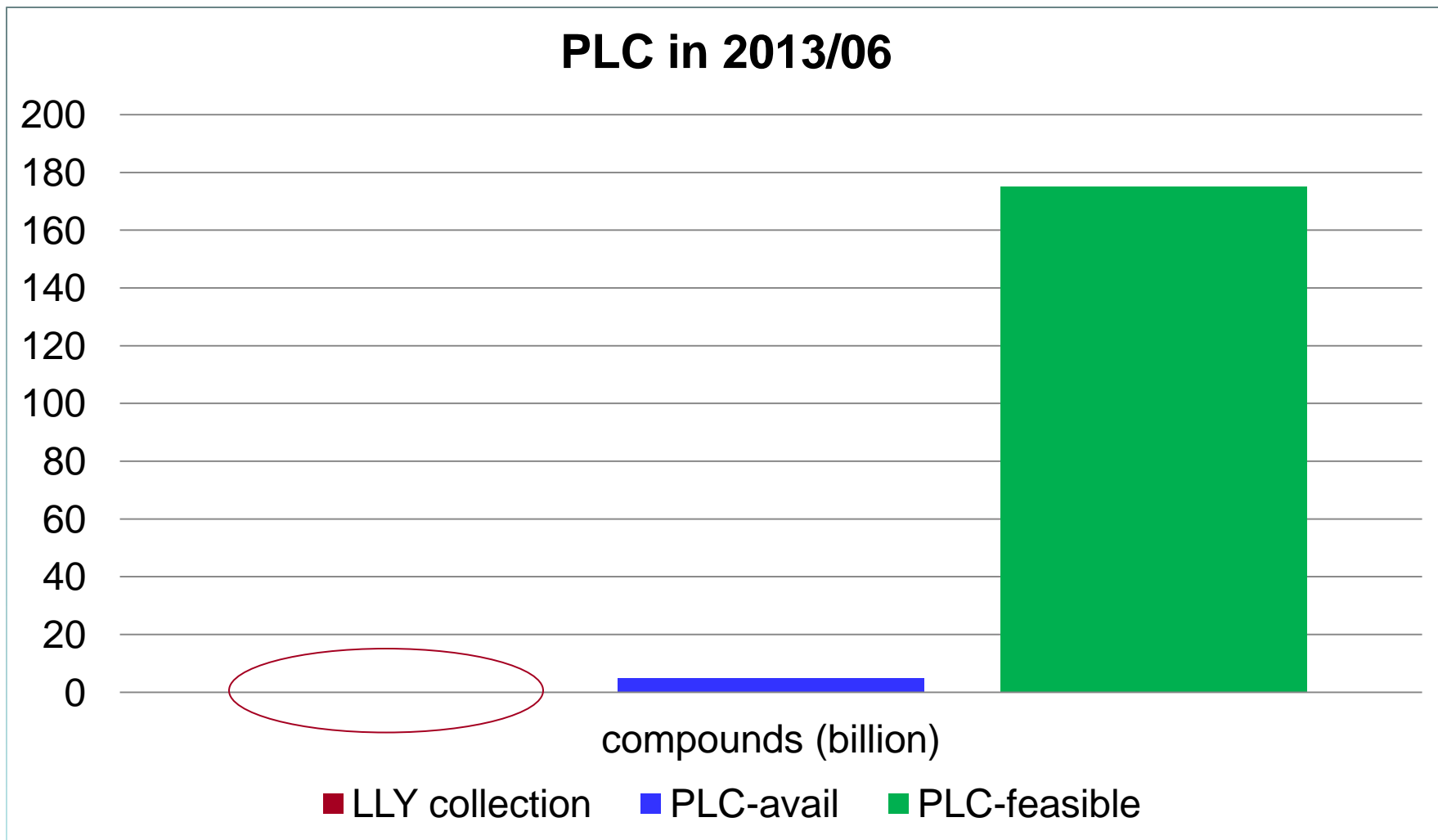
>90% coverage of PLC diversity by enumeration of 2.5%



Sampling PLC - Example



PLC Current Snapshot

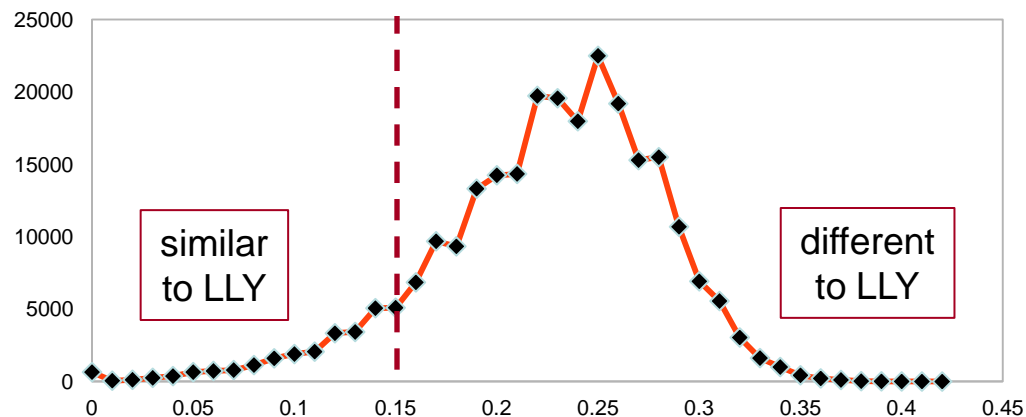


*PLC-avail: readily synthesizable

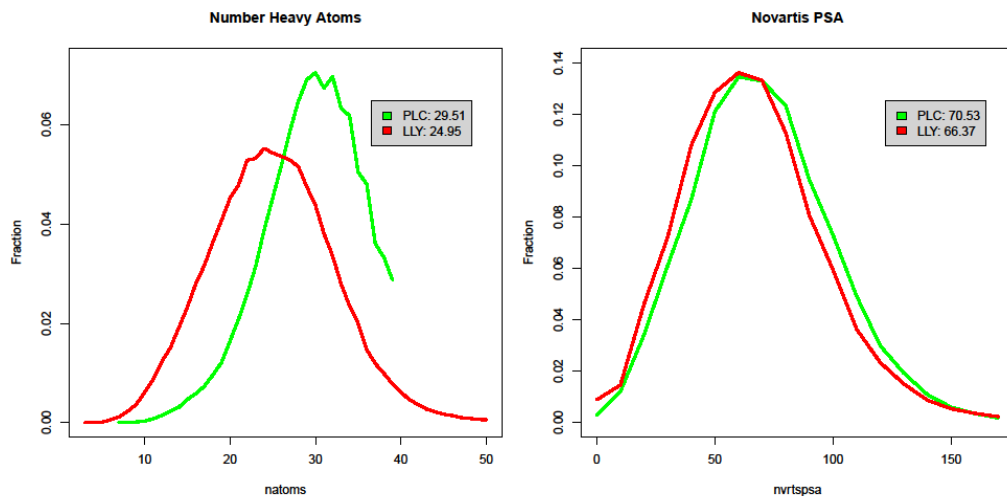
PLC Chemical Structure Profile

- Diverse
- Contains both structurally similar and novel
- Drug-like
 - Property profile similar to LLY collection
- Exploiting existing Lilly resources; emphasizing practical concerns (including synthesizability)

PLC chemical structure novelty

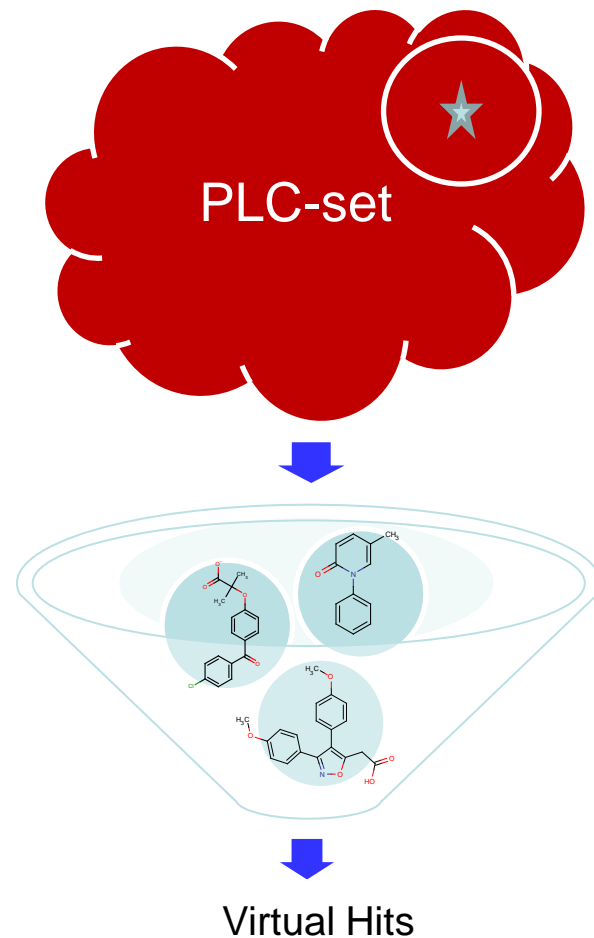


PLC chemical properties



PLC Usage Scenarios

- Original
 - Virtual Screening
 - Hit Expansion
 - Original PLC hit; any structure
 - Synthetic Route
- Emerging scenarios
 - Scaffold exploration
 - De Novo Design
 - ...

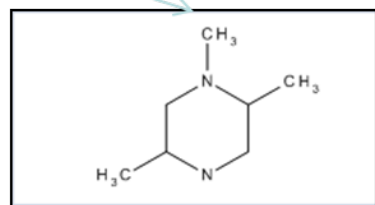
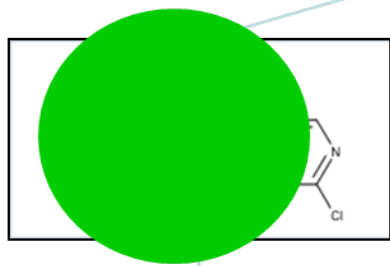
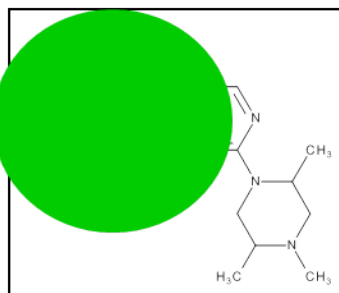


PLC-Expand

- Hit Expansion in PLC space
 - Via near neighbor search on representative set
 - Fingerprints, descriptors, hit scaffold, etc.
 - Via PLC-hit reagent investigation
 - Decompose hit/seed/needle into constituting reagents
 - Search for similar reagents in reagent pool
 - Virtually synthesize all similar reagents
 - Assess similarity to seed compound
 - Via scaffold-based analysis
 - Retrieval of compounds from original scaffold-based clusters of hits identified by the Virtual Screening process
 - Combination of the above methods
 - E.g. given any seed structure, find near neighbor in representative set and apply PLC-hit reagent investigation
-

Example

3169606_2358848_XXX



Similar reagent list A

Similar reagent list B

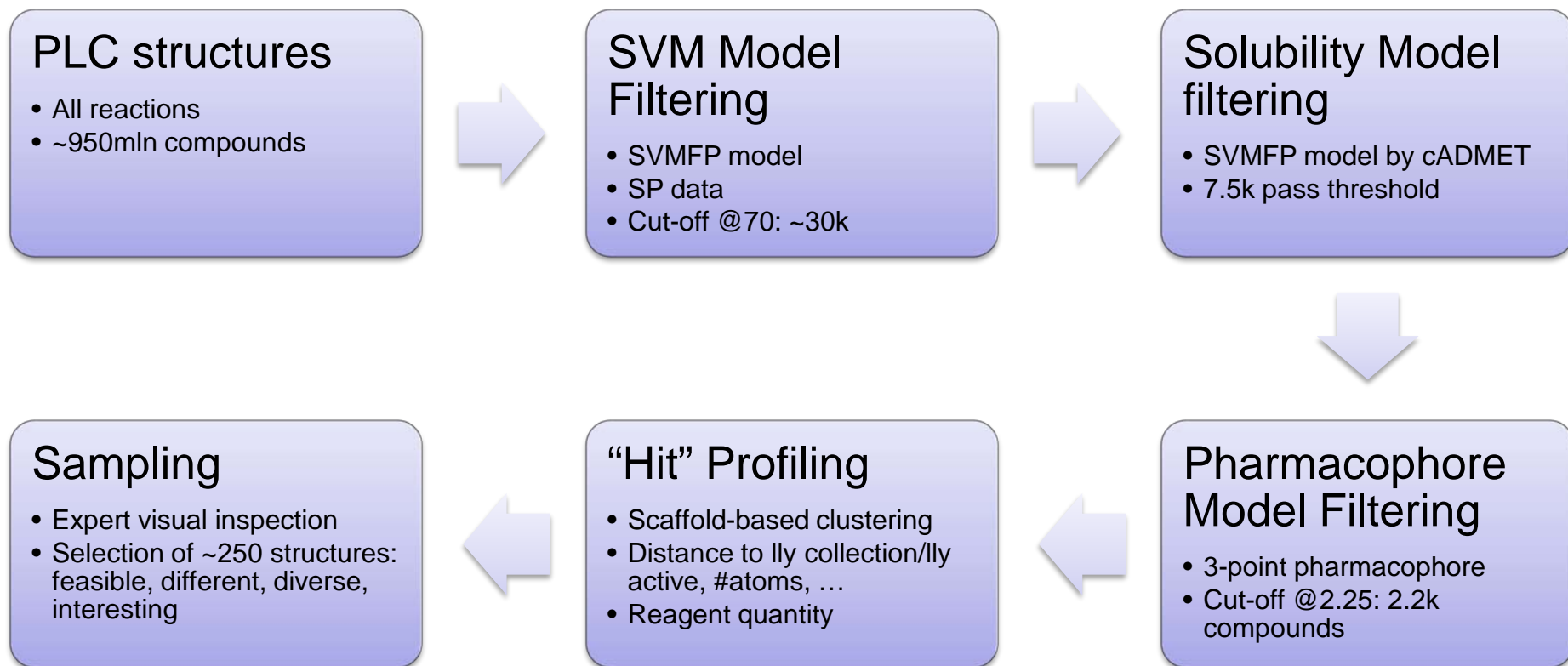
ASL Virtual
Synthesis
engine

	<chem>CN1CCN(C)CC1Cl</chem>	3169606_21090	_XXX	0.077
	<chem>CN1CCN(C)CC1Cl</chem>	3169606_81699	_yyy	0.084
	<chem>CN1CCN(C)CC1Cl</chem>	3169606_21090	_XXX	0.086
	<chem>CN1CCN(C)CC1Cl</chem>	3169606_21090	_yyy	0.086
	<chem>CN1CCN(C)CC1Cl</chem>	3169606_21072	_ZZZ	0.094
	<chem>CN1CCN(C)CC1Cl</chem>	3169555_23588	ZZZ	0.101

PLC Pilot

- Pilot study on realistic showcase
 - Challenge posed
 - Apply Virtual Screening on PLC to identify interesting new structures for ongoing project
 - Objectives
 - Test PLC-based Virtual Screening; refine process as needed
 - Support Idea2Data initiative
 - Well-known kinase target
 - Ongoing project
 - Active assay
 - Screening data availability
 - Target structure availability
-

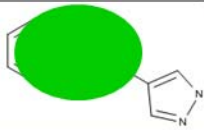
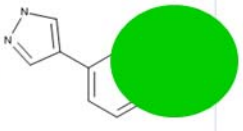
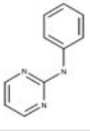

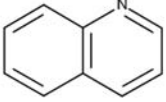
Pilot Process - 1

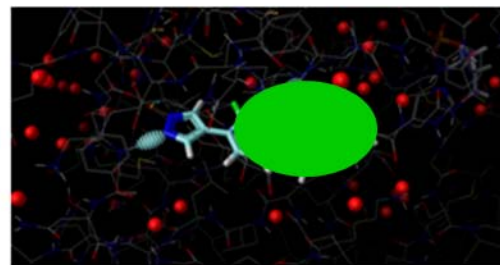


- Input 950Mn
- Processing via SVM and PP models
- Profiling via clustering, quantity, novelty, ...
- Output: ~250 structures

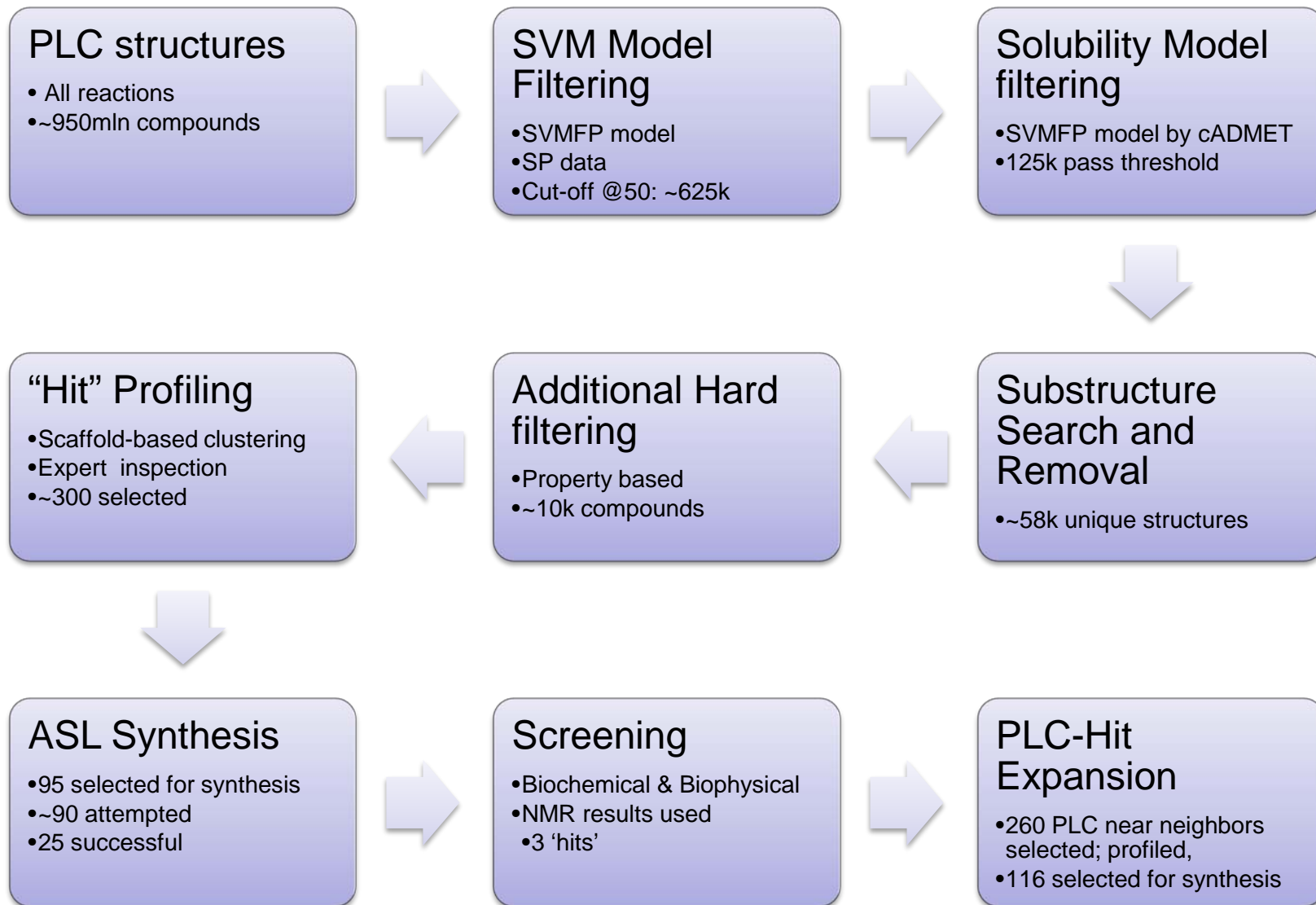
Process Verification

- Known structures retrieved
 - structural motifs frequently found in Kinase binders
- Reassuring result

Smiles	A	B	C	D	E	F	G
	Cluster#	#inCluster	#wIC50	#wIC50<1uM	#wSP	#wSP>70%	
	1	935	122	59	216	122	
	2	699	330	128	67	45	
	3	622	1126	611	591	270	
	4	287	480	341	210	148	
	5	11	180	73	550	160	



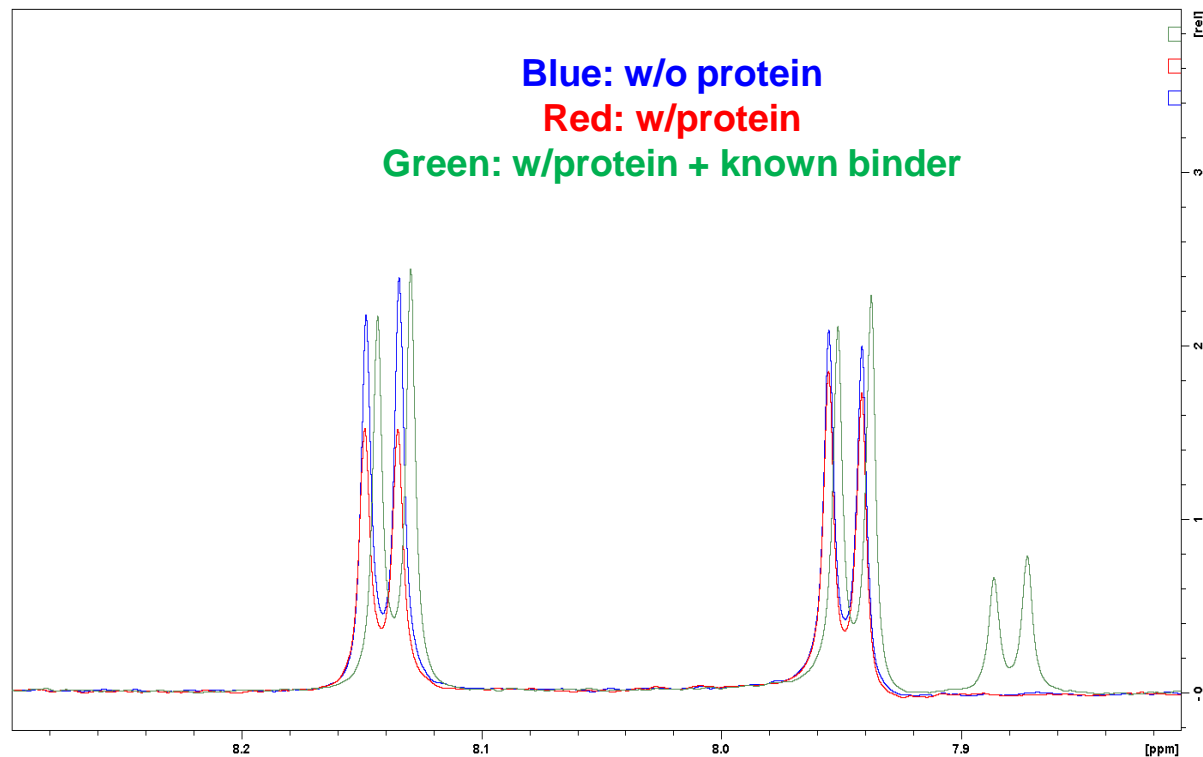
Pilot Process – 2 (to date)



Ongoing!

Results

- Known structures retrieved
- Novel structures meeting VS criteria identified
- 3 hits discovered
- Further investigation ongoing
 - Hit expansion–260 structures
 - 116 undergoing synthesis, ...



- Signal reduction upon addition of protein indicates binding
- Signal recovery upon addition of known strong binder suggests binding is competitive

Lessons learned

- PLC can serve as a 3rd source of structures for discovery
 - Includes known actives; known structures retrieved
 - Suggests new, interesting hits
- PLC diversity limited by:
 - Kind of reagents available
 - Type of reactions currently encoded
 - Requires continuous refinement, reaction encoding, select acquisitions,...
- Distinct application scenarios emerging
 - Early stage projects: identify promising structures
 - Mature projects: emphasis on structural novelty
 - avoid known structural motifs
 - Address selectivity issues
 - Optimize ADME property(ies)
 - ...

Acknowledgements

- Management
 - Scott Sheehan
- Project Team
 - Maria Carmen Fernandez
 - Eva Maria Martin
- C3
 - Jon Erickson
 - Christine Humblet
 - Christos A Nicolaou
 - Hong Wang (ImClone)
 - Jibo Wang
 - Hongzhou Zhang
- cADMET
 - Prashant Desai
- ASL
 - James Beck
 - Alexander Godfrey
 - Hong Hu
 - Angela Marquart
- AT
 - Bob Boyer
 - Keith Burton
 - Gary Sharman
 - Paul Tan
 - Ken Visscher
 - Beth Wright
- Statistics
 - Suntara Cahya
 - Ian Watson
- GSB
 - Jorg Hendle

Abstract

- Efforts for mapping chemical space and reaching out to less explored, but potentially promising regions for pharmaceutical development have been hampered by the sheer number of theoretically feasible compounds and the practical concern on the synthesizability of the chemical structures proposed. In a typical setting, such virtual compounds are conceived through the enumeration of the products of chemical reactions when supplied reagent sets appropriate to the specific reaction. Alternatively, virtual compounds may be proposed simply through the permutation of a set of atoms abiding to some rudimentary chemical structure rules. In either case the result is a large virtual collection of chemical structures of questionable synthesizability and, therefore, practical use.
- This presentation provides a description of LiRCS, the Lilly Reachable Chemicals Space system, designed to bridge the chemical synthesis knowhow and potential at Eli Lilly with the needs of ongoing discovery chemistry projects. LiRCS can be thought of as the computational counterpart to our Automated Synthesis Lab (ASL) system which served as the main motivation for this work. In its current incarnation, the system focuses almost exclusively on reactions validated on the ASL.
- LiRCS consists of the Lilly Annotated Reaction Repository (LLARR), a flexible virtual synthesis engine (VSE) supported by a high-performance computing system, and, a collection of search and retrieve utilities dedicated to a number of user-defined usage scenarios. In order to ensure synthesizability of the proposed compounds, LLARR relies on the usage of an annotated reaction scheme which captures a wealth of information on a reaction, including the detailed profile of the reagents that may (or may not) be used. Reactions in LLARR can be used by the VSE to either enumerate the full matrix of virtual compounds possible with a set of reagents or to simply generate structures by combining specific reagents subject to user imposed restrictions on the final products, related to e.g. compound size or chemical structure. VSE is tightly coupled with the Lilly chemical sample management system to ensure that the reagent sets used in the process can readily be accessible for chemical synthesis.
- The system is designed to support a number of usage scenarios by Lilly scientists. Among them are the generation of virtual screening sets based on user defined objectives; the retrieval of reachable compounds satisfying structural similarity restrictions to a query compound (LiRCS-Grow); the retrieval of the synthetic route of a specific query compound (LiRCS-Retrieve). The presentation will provide an account of the algorithmic development that led to the system stressing the techniques implemented to manage the size of the reachable chemical space. Specific examples showcasing results produced by the system and demonstrating its capabilities will be presented. A discussion on lessons learned, issues to be resolved, and future development directions will conclude the presentation.