



Data Pipelining: A new paradigm for chemical data processing and analysis in high throughput discovery

Robert D Brown, PhD,
Senior Director, Cheminformatics, SciTegic Inc

8th Solid Phase Synthesis and Combinatorial Libraries,
London 2003

Outline

- Informatics in the high throughput discovery age
- Data pipelining: a new paradigm for high throughput informatics
- Application of data pipelining to combinatorial chemistry
- A practical example at a top ten pharma
- Summary

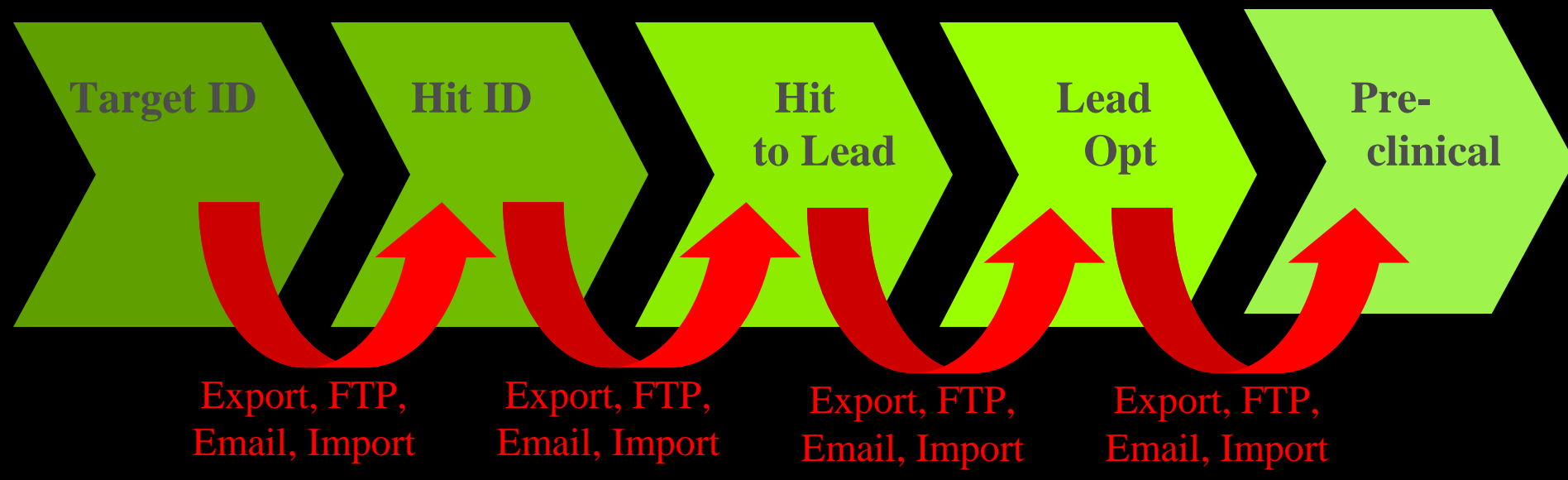
High Throughput Discovery

- Industrial discovery has undergone a revolution to become automated high throughput process, e.g.
 - 1990 – Pharma screened ½ million compounds
 - 2000 – Pharma screened 1500 million compounds



High Throughput Informatics

- Informatics solutions have struggled to keep pace
 - Disparate data without standards
 - Disparate point solution applications on diverse hardware
 - Sheer volume of data is overwhelming
 - Database centric solutions are slow
 - Data processing and communication requires manual intervention

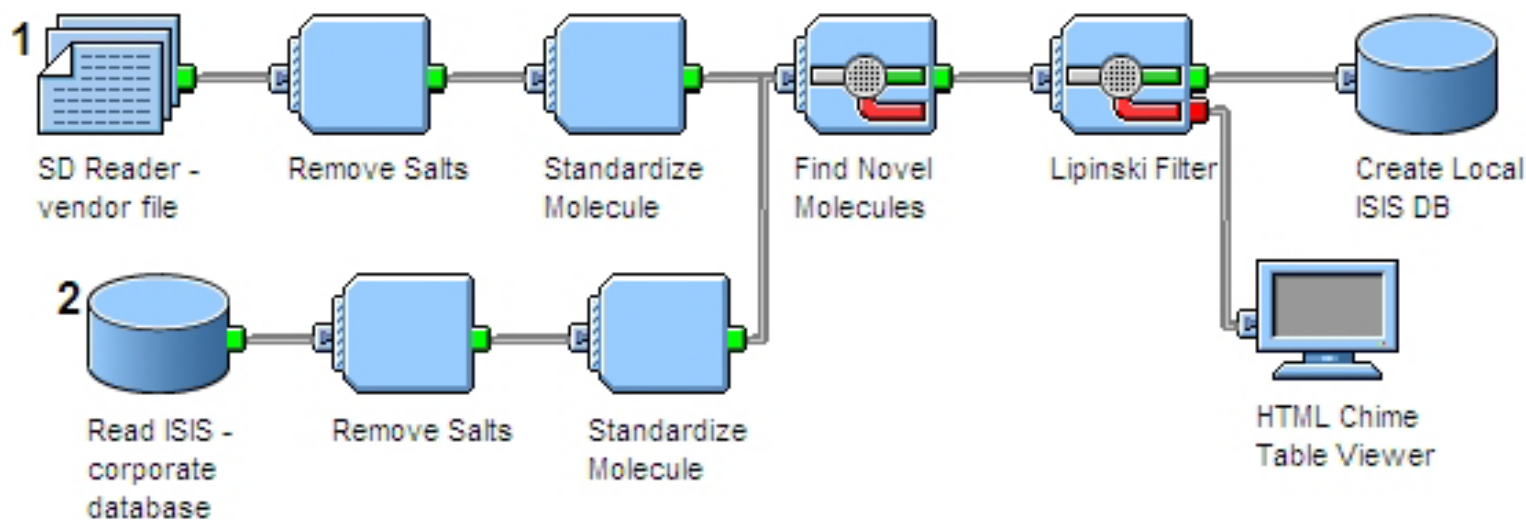


Informatics Must Evolve

- Integrate disparate data sources
- Integrate disparate applications
- Process and analyze data in real time
- Automate processes, removing manual intervention
- Capture and document best-practice processes
 - For reuse
 - For regulatory and auditing purposes
- Deploy informatics across the organization
 - Computational expertise is limited

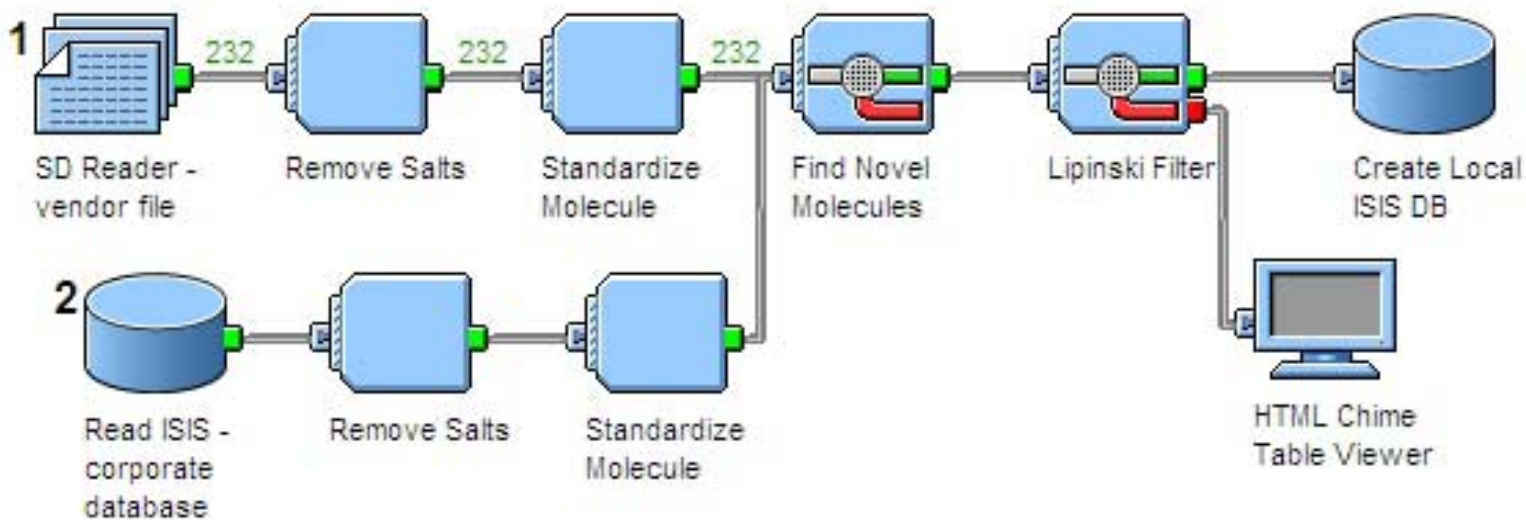
Data Pipelining

- A powerful new paradigm for data processing
- Pipelines guide the flow of data through a network of modular computational components



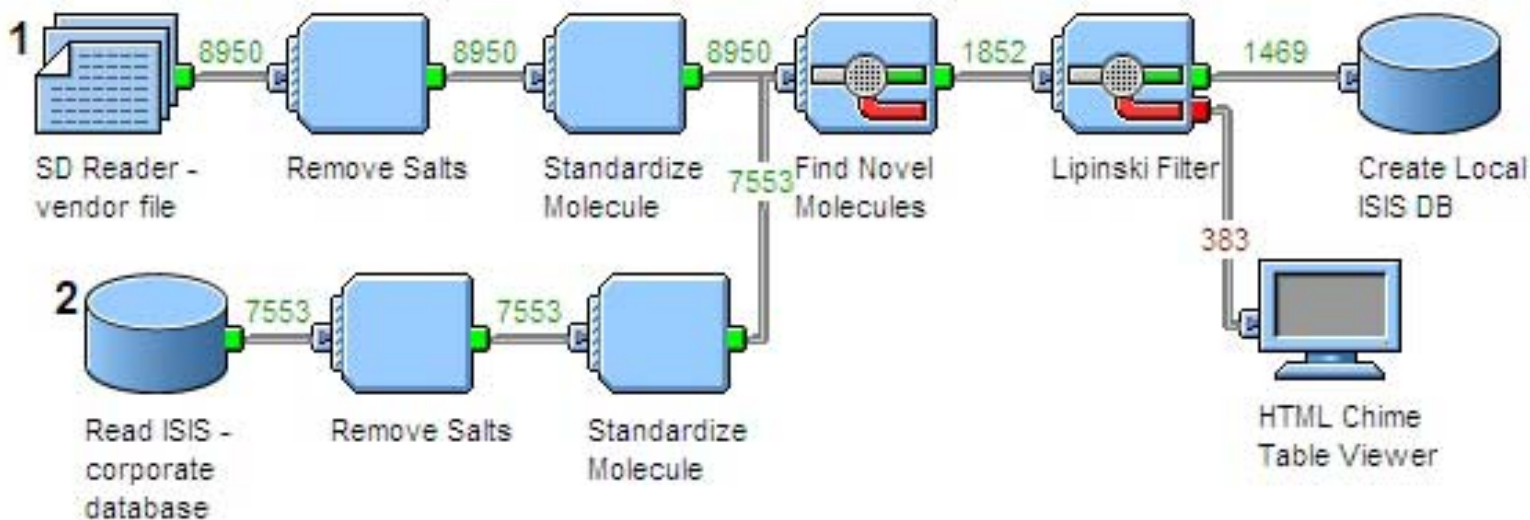
Data Pipelining

- A powerful new paradigm for data processing
- Pipelines guide the flow of data through a network of modular computational components



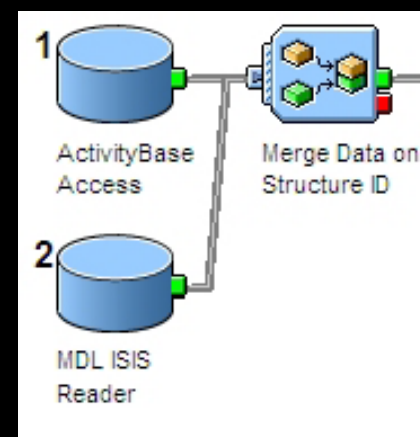
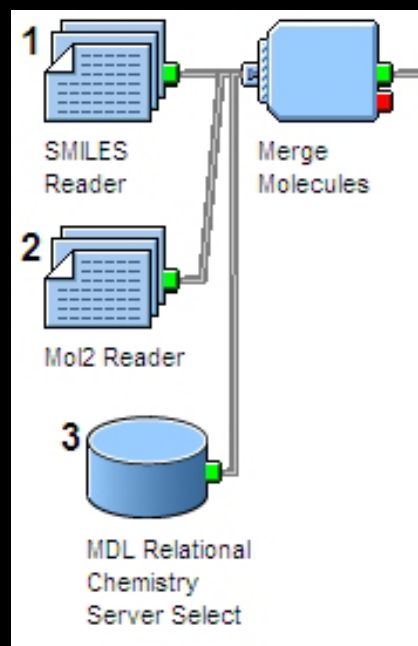
Data Pipelining

- A powerful new paradigm for data processing
- Pipelines guide the flow of data through a network of modular computational components



Data pipelining enables

- Processing of data from multiple disparate data sources
- Integration of disparate applications
- Rapid processing of large amounts of data
- Automated execution of routine processes
- Capture of best practice



Data pipelining enables

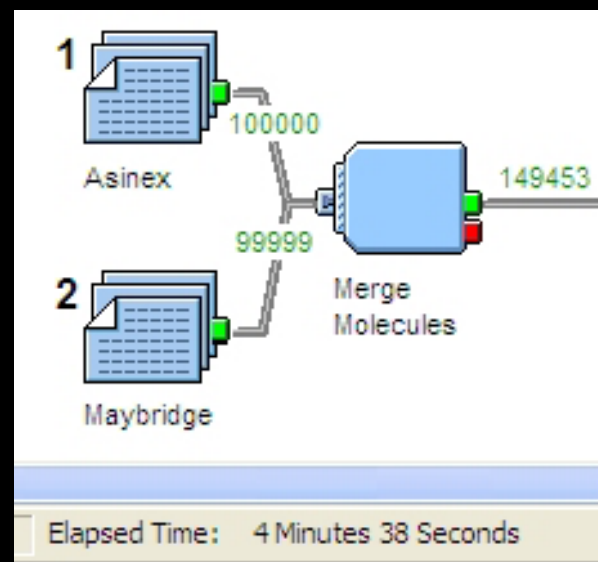
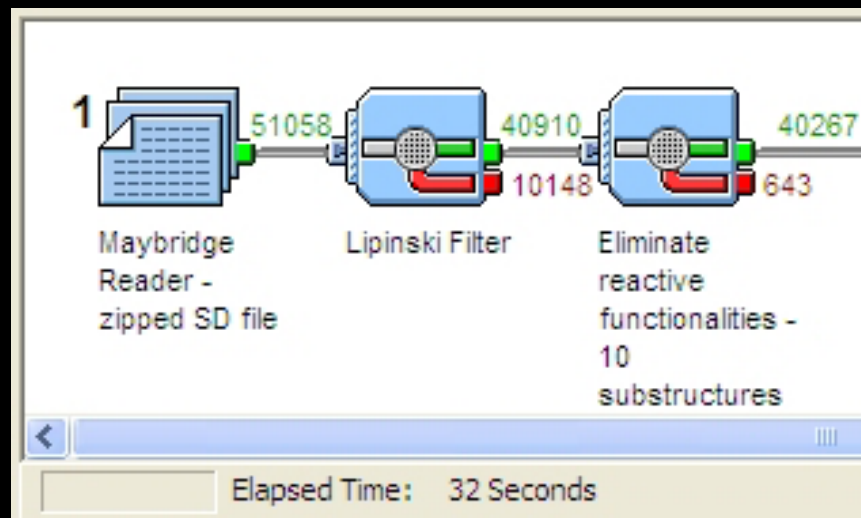
- Processing of data from multiple disparate data sources
- Integration of disparate applications
- Rapid processing of large amounts of data
- Automated execution of routine processes
- Capture of best practice



The Structure	CLogP	CLogP_MR	MDL_SS_Keys_166_N
	2.002	5.160	32
	1.372	3.591	28
	0.721	3.651	32

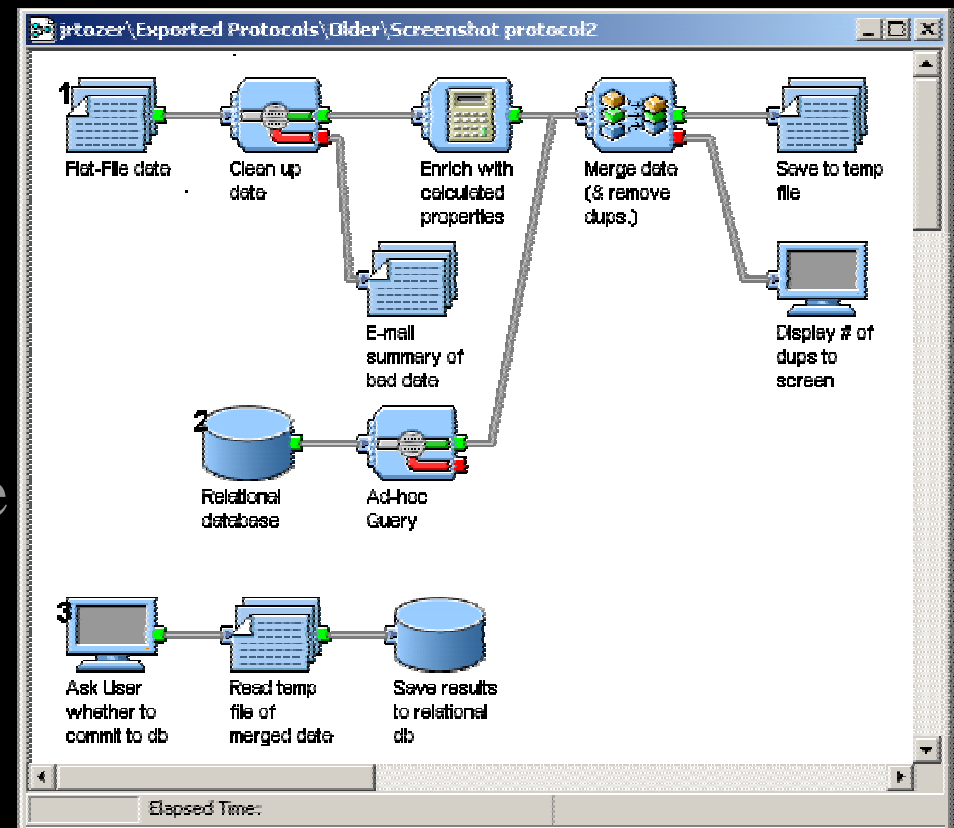
Data pipelining enables

- Processing of data from multiple disparate data sources
- Integration of disparate applications
- Rapid processing of large amounts of data
- Automated execution of routine processes
- Capture of best practice



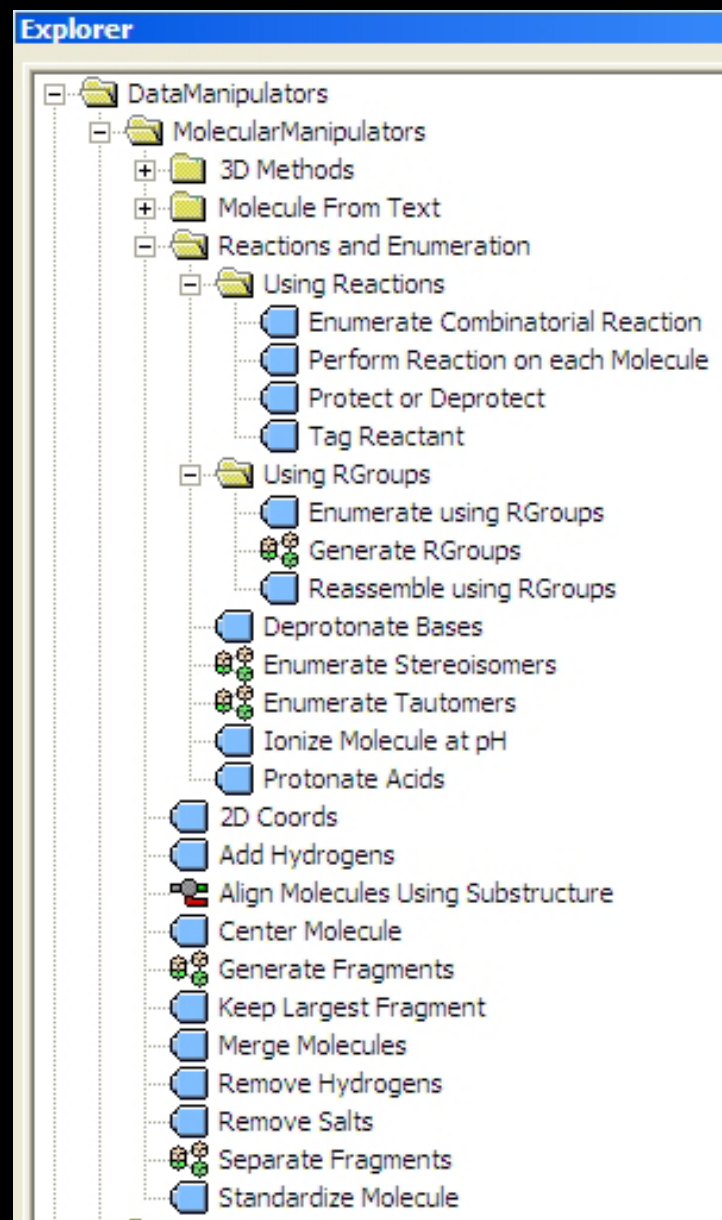
Data pipelining enables

- Processing of data from multiple disparate data sources
- Integration of disparate applications
- Rapid processing of large amounts of data
- Automated execution of routine processes
- Capture of best practice



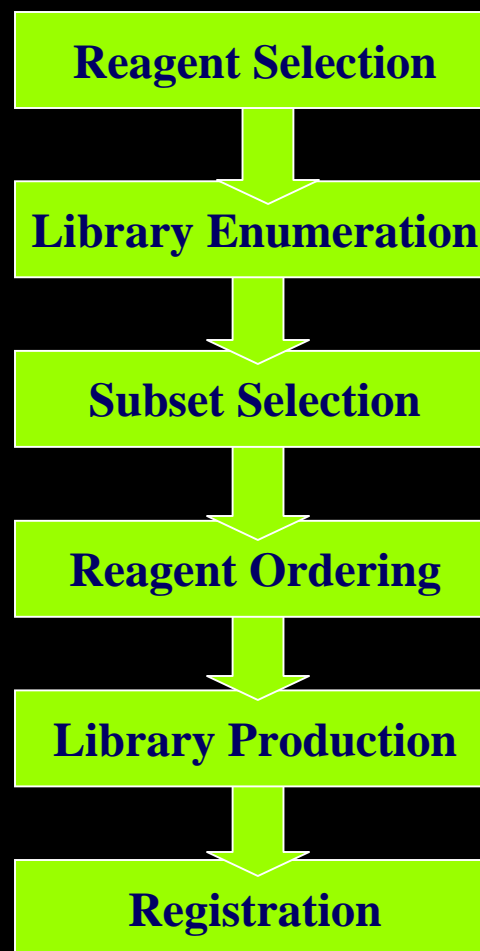
Creating Data Pipelines

- Build information handling protocols graphically
 - From a palette of available components
 - Experts can build new components as required
- Easy-to-use (e.g., drag and drop)
 - No code writing necessary
 - Yet highly configurable for broad applicability

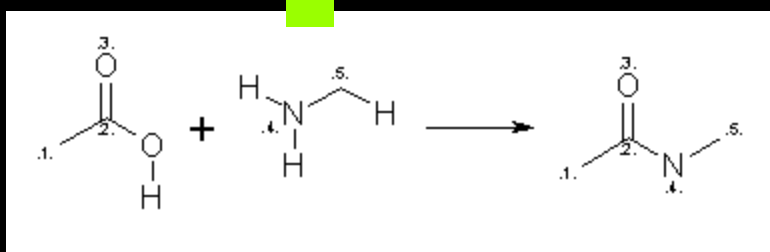
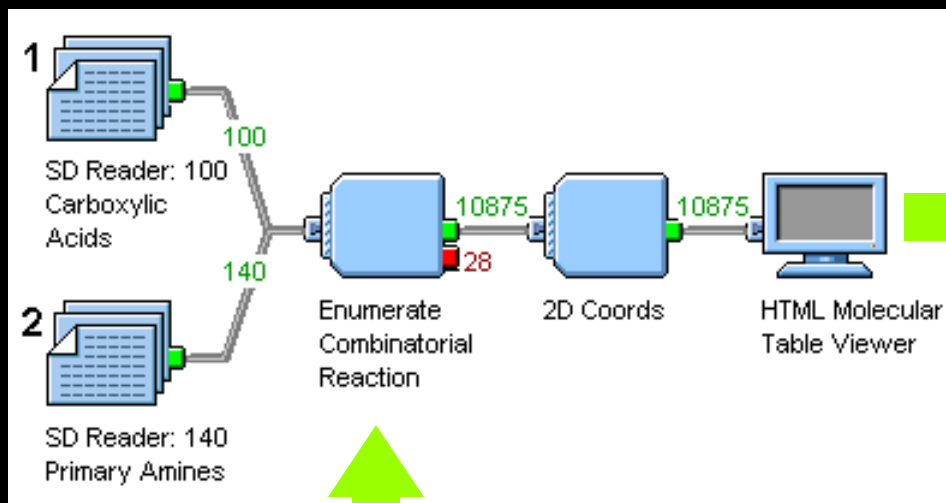


Application to Combinatorial Chemistry

- Pipelining provides high speed, automated processing infrastructure
- Scientific components required for
 - Reagent selection
 - property and substructure filters
 - Library enumeration
 - by reactions or by scaffold
 - Library analysis & subset selection
 - for diversity, lead follow up, “lead-likeness”
 - Ordering and tracking reagents
 - Registering final structures

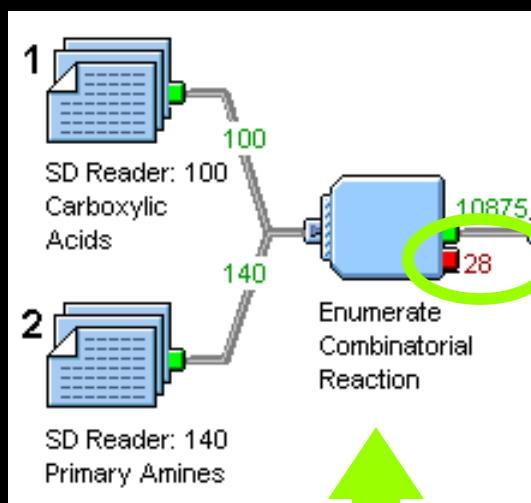


Reaction-based enumeration

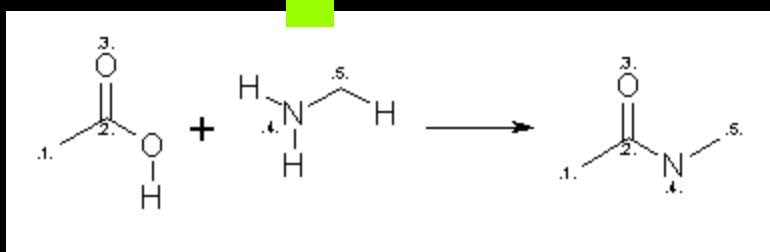


Molecule	SourceTag	IDNUMBER
	PrimaryAmines CarboxylicAcids	BAS 0009781 BAS 0004192
	PrimaryAmines CarboxylicAcids	BAS 0040482 BAS 0004192
	PrimaryAmines CarboxylicAcids	BAS 0043450 BAS 0004192

Reaction-based enumeration

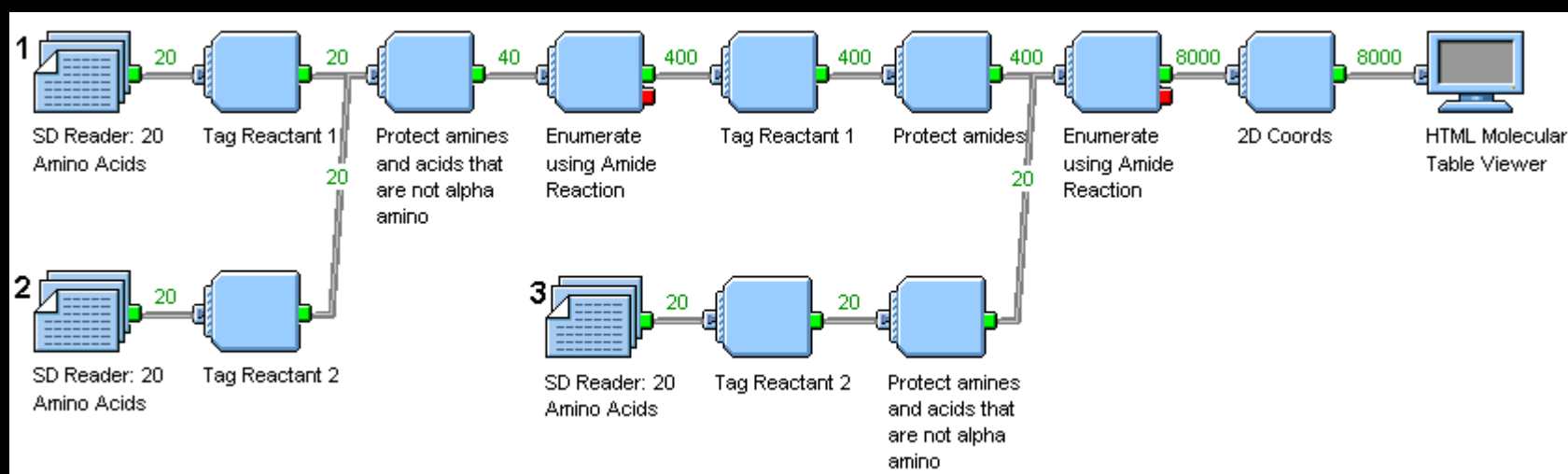


	CarboxylicAcids	BAS 0094487	No reactant mapped onto starting material
	CarboxylicAcids	BAS 0096741	Multiple reactants mapped onto starting material: 1, and 2
	CarboxylicAcids	BAS 0110375	Reactant 1 mapped more than one way onto starting material

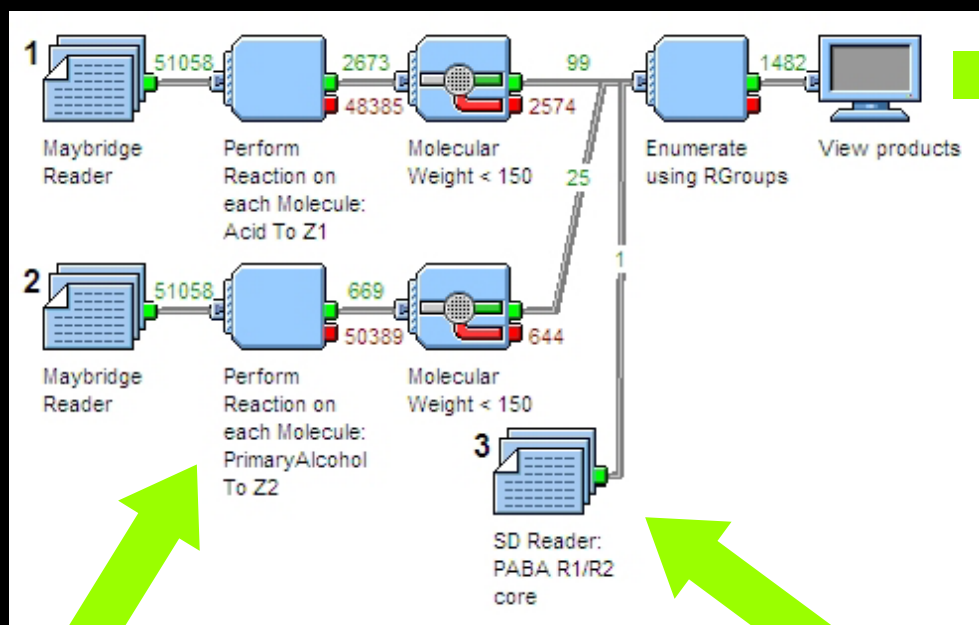


Reaction based enumeration

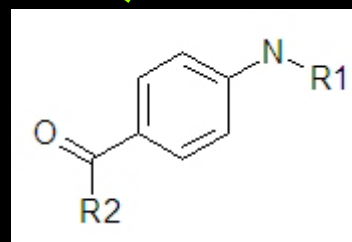
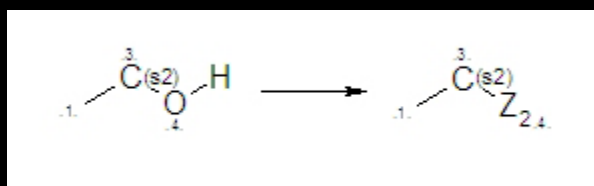
- Protection and deprotection
- Stereochemistry: create and change stereocenters
- Multistep reactions
- Multicomponent steps
- Speed – up to 3000 mols/second (10M/hour)



Scaffold-based enumeration

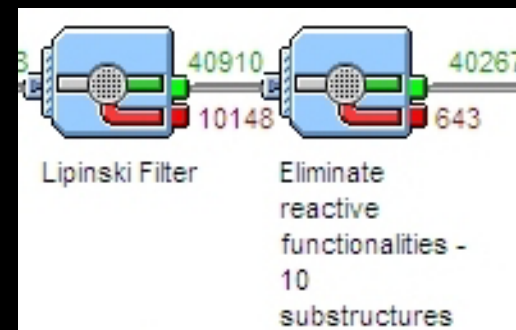


Molecule	CODE
	TL 00169 NBX 00002
	TL 00169 BTB 01548
	TL 00169 BTB 13819



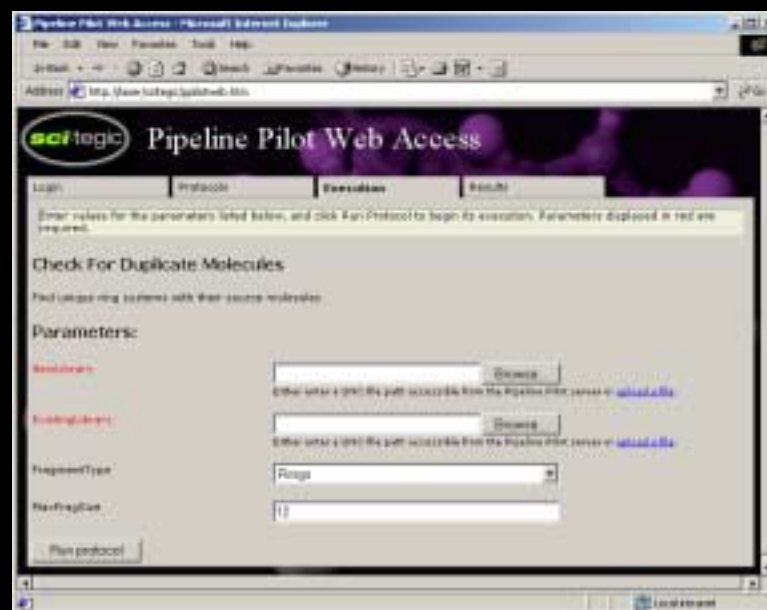
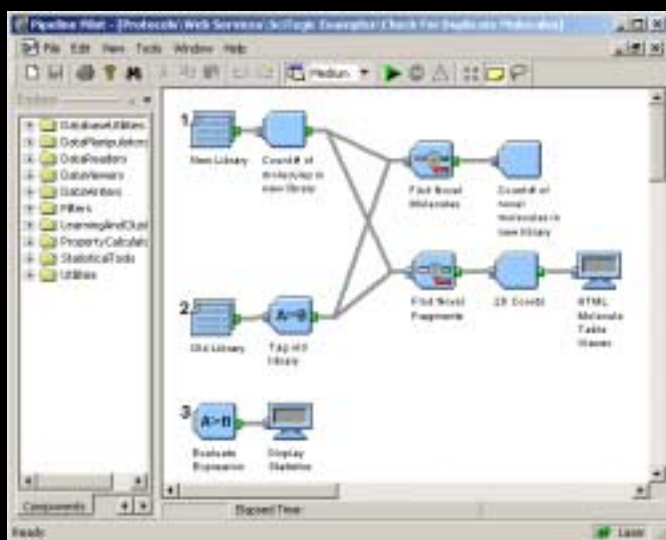
Library Analysis

- Lead-likeness or drug-likeness
 - Filter for undesirable substructures
 - Filter for undesirable properties
- Lead generation by diversity
 - Select by maximum dissimilarity
- Lead optimization
 - Select by similarity radius or cluster
- Many 3rd party programs typically used – these are easily integrated as components

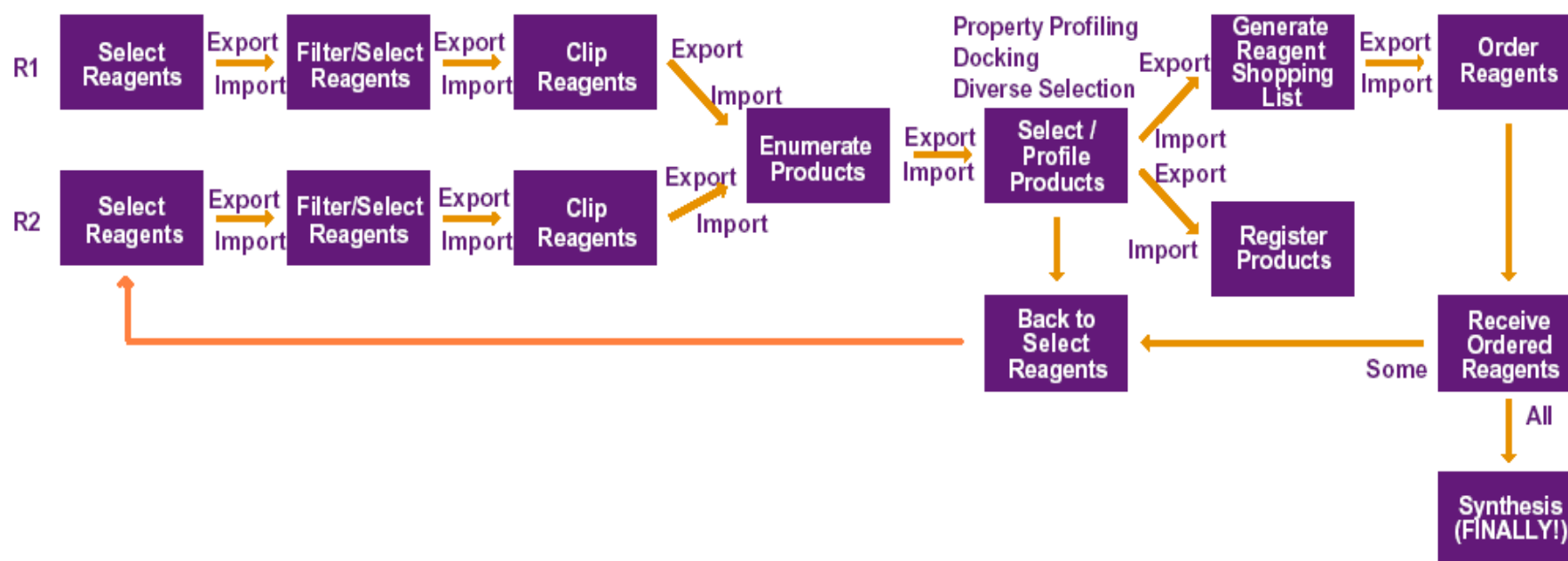


A Practical Example

- Data pipelining is used at a top ten pharma to provide library design services directly to chemists
- Takes advantage of the web deployment of protocols to provide a familiar look-and-feel to end users



Typical Library Design Workflow Before Data Pipelining



Most of the effort is spent “running the gauntlet”

Issues With This Workflow

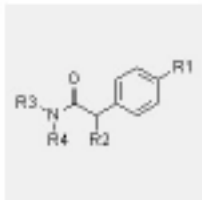
- Too many steps, many steps too complex for the average chemist
- Multiple computing platforms add to the complexity
- Tracking of compounds through the process
 - Export/Import steps makes this especially difficult
- Existing process does not allow for refinement of library design during synthesis
 - When a reagent replacement is required or desired
- Chemists should be spending “Intellectual Effort”, not “File Management Effort”

Product-Based Enumeration

Current User: Select Scaffold + RGroups Enumerator Help

Virtual Library Name: Load Library Data:

Number of RGroups:

Core Structure: 

NOTE: The enumerator will fail and generate errors when:

- The number of RGroups selected exceeds the number of RGroups on the Core Structure. If the opposite is the case, then any additional RGroups on the Core Structure will be ignored.
- The RGroups are not numbered sequentially beginning with R1.

RGroup 1

Source SD File: RGroup Core

[UPLOAD a FILE] to your home directory on the SciTeGic server
Or select a file already located on the SciTeGic server
File locations must be specified with a UNC path, such as: \\SciTeGic\Users\backes\Documents\Backes-Suzuki\R1\Boronic Acids.sdf

Number of Reagents to Use (131):

Leave Markers on all reagents:

Skip Reagent Transformation:

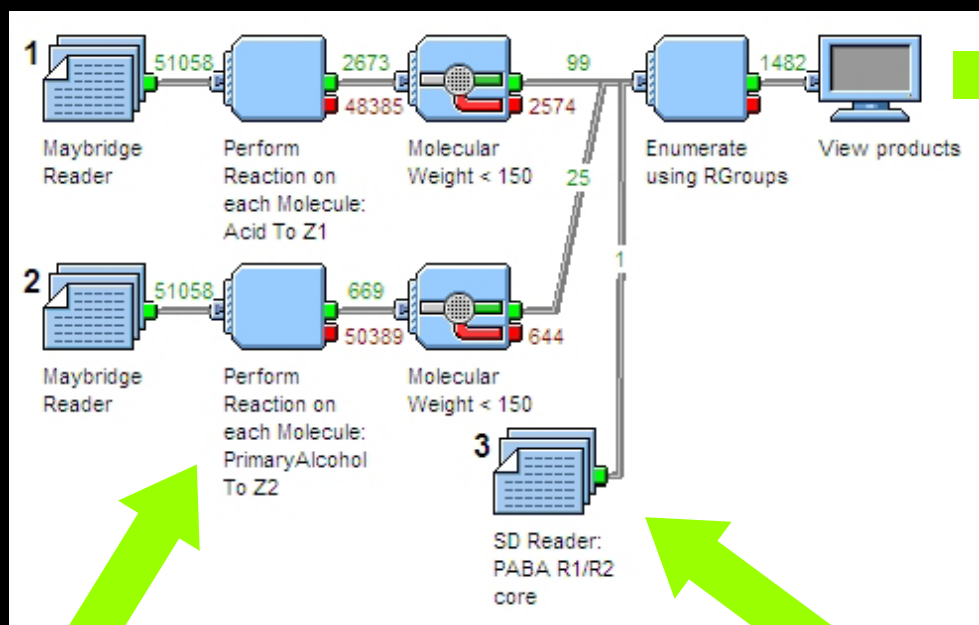
SD File contains clipped reagents using single Z attachment points
RGroup Core and Transformation are not required

Compound ID Field:

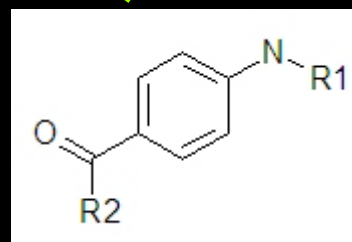
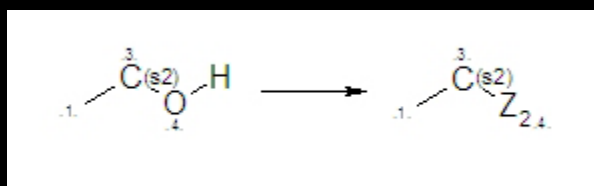
Reagent Transformation

[UPLOAD a FILE] to your home directory on the SciTeGic server
Or select a file already located on the SciTeGic server
File locations must be specified with a UNC path, such as: \\SciTeGic\Users\backes\Documents\Backes-Suzuki\R1\R1 Transform.rxn

Scaffold-based enumeration



Molecule	CODE
	TL 00169 NBX 00002
	TL 00169 BTB 01548
	TL 00169 BTB 13819

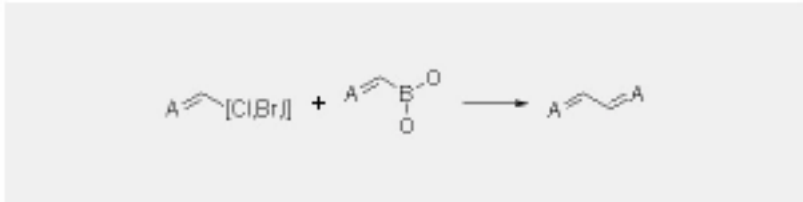


Reaction-Based Enumeration

Current User: Select Reaction + Reagents Enumerator Help

Virtual Library Name: Load Library Data: [Clear Data]

Number of Reagents: Load Reaction:

Reaction Structure 

Multiple Mapping Options:

Reagent 1

Source SD File

[UPLOAD a FILE] to your home directory on the SciTegic server
Or select a file already located on the SciTegic server
File locations must be specified with a UNC path, such as

Number of Reagents to Use (109)

Leave Blank to use all reagents

Enter Compound ID Field:

Reagent 2

Source SD File

[UPLOAD a FILE] to your home directory on the SciTegic server
Or select a file already located on the SciTegic server
File locations must be specified with a UNC path, such as

Number of Reagents to Use (131)

Leave Blank to use all reagents

Enter Compound ID Field:

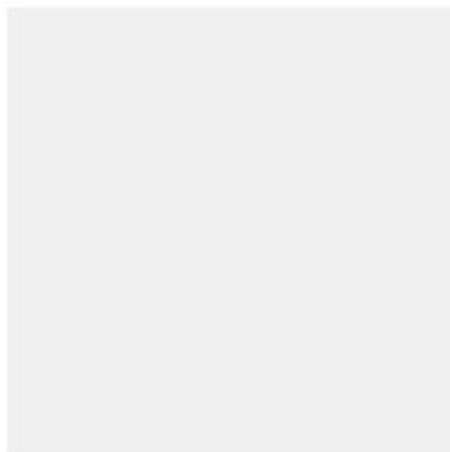
Structure Profiler

Current User: Select Compound Source to Profile Select Profile Help

Select Enumerated Library: Product Count: 384

Select Reagent Selector List: List Size: 124

Input Single Structure:

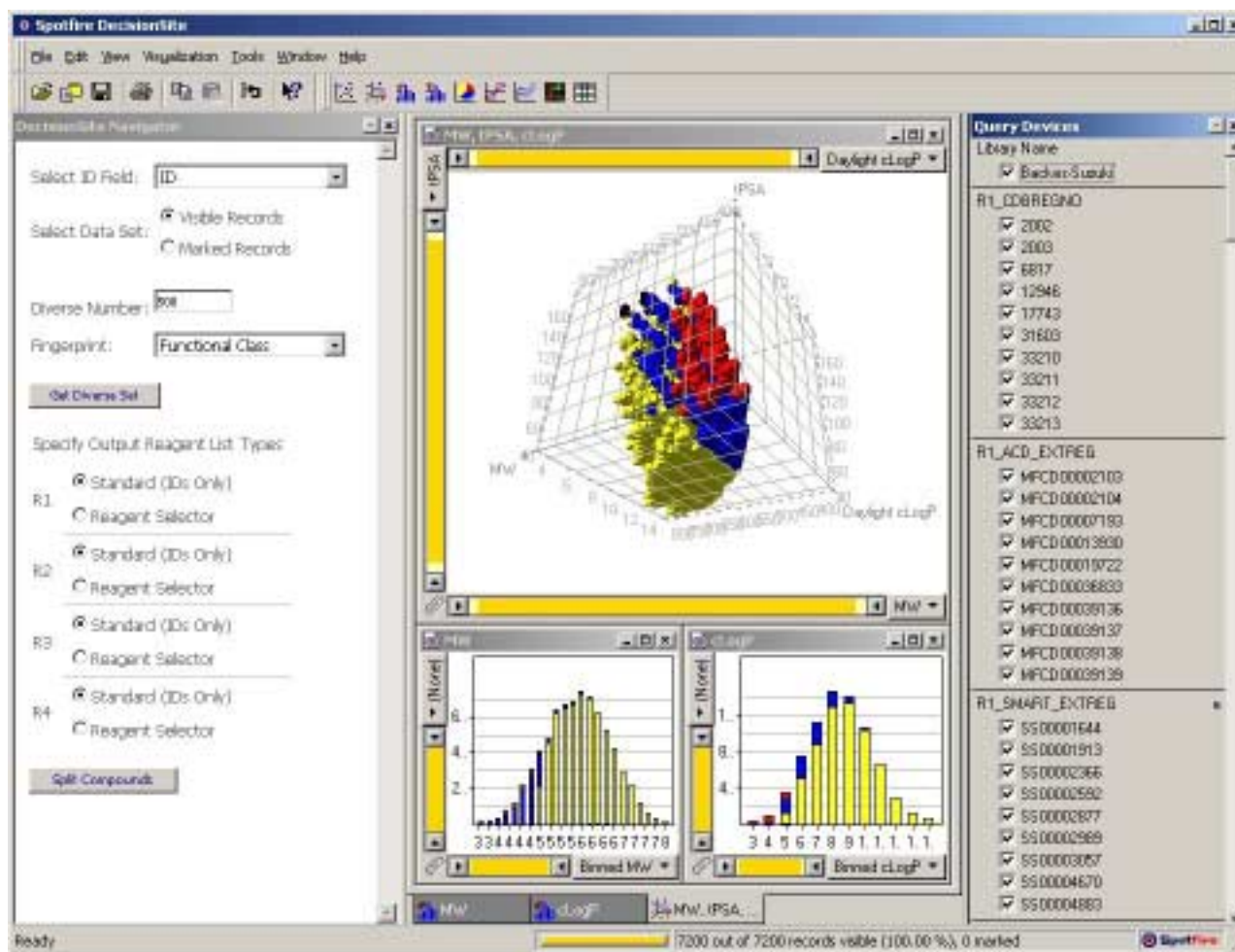


[UPLOAD a FILE] to your home directory on the SciTegic server
Or browse for a file already located on the SciTegic server
File locations must be specified with a UNC path, such as \\SciTegic\Users\Data.sdf

Select SD File:

Enter Compound ID Field: [\[Display Fields\]](#)

Spotfire and Selection



Summary

- Data pipelining addresses the informatics challenges of high throughput chemistry
 - Integrate disparate data sources and applications
 - Process and analyze data in real time
 - Automate processes, removing manual intervention
 - Capture and document best-practice workflows
 - Deploy informatics across the organization

Acknowledgements

- SciTegic
 - Matt Hahn
 - David Rogers
 - Moises Hassan