ChemBank Technical Platform and Data Pipeline
Chemical Biology Workflow

- **Library Conceptualization**
  - modeling synthetic pathways
  - reagent selection
  - library enumeration
  - diversity analysis

- **Physical Library Creation**
  - pathway & reagent testing
  - library setup and execution
  - encoding

- **Library Validation**
  - decoding
  - analytical chemistry

- **Assay Development**
  - projects
  - protocols and methods
  - experimental design

- **Assay Execution & Data Analysis**
  - error modeling
  - results ("hit") analysis

- **Biological Interpretation**
  - capturing and tracking biological activity
Data Sources

- Seed data
  - admin users, controlled vocabularies, imported ontology
- Compound registration
  - vendor libraries, DOS libraries, FDA approved drugs
- Project data & assay metadata
- Assay measurements and analytical data
  - readout values, computed values, “hit”-lists
- Biological interpretation ‘findings’
  - linkages between ontology elements and experimental objects
  - supporting evidence pulled from journals and/or from screens in ChemBank
ChemBank Data Import Pipeline

Users
- imported controlled vocabularies & ontology
- experimental objects

Integrate seed data
Import files

ChemBank

Compound registration

Screening

Biological interpretation
ChemBank Data Import Pipeline

seed data

compound registration

vendor libraries
DOS libraries
other sources

salt-stripping, canonicalization
import files

ChemBank

screening

register findings
ChemBank Data Import Pipeline

- seed data
- compound registration
- screening
  - project & assay metadata
  - raw plate-reader data
  - error correction, "hit" analysis
  - import files
- register findings
- ChemBank
ChemBank Data Import Pipeline

- Seed data
- Compound registration
- Screening
- Register findings
  - External publications
  - Internal “hits”
  - Curate supporting evidence
  - Import (findings)

ChemBank
ChemBank Technology Platform

- J2EE (language/SDK)
  - **JUnit** (unit testing)
  - JMeter (stress testing)
  - JFreeChart (visualization)
- **Eclipse**
- **Ant**
- **Struts**
- **JBoss**
- **Oracle 10g**
- Daylight (cheminformatics cartridge and toolkit)
- **Subversion** (source code control)
- **Bugzilla** (bug tracking)
- **dotProject** (project planning)
- **RequestTracker** (error, help request, and suggestion tracking)
- **TWiki** (requirements/design/implementation documentation)
ChemBank Application Architecture

- Application
- Management Objects, Reusable Classes
- Java Services
- Databeans
- Oracle
Extensible Architecture

ChemBank v2.0

cell line sample
RNAi sample
protein sample

ChemBank v2.0

finding

phenotype
biochemical interaction
therapeutic use

ChemBank v2.1

...
ChemBank Schema Domains

- Users, organizations, departments
- Library conceptualization (synthetic pathways, reagents, molecular descriptors)
- Library creation (compounds, vendor/chemist, plates)
- Library testing (analytical chemistry)
- Project tracking
- Plate management
- Sample tracking
- Assay metadata, protocols
- Assay measurements and computed values
- Biological findings and related supporting evidence
**ChemBank Data Architecture**

**common database architectures**

- **transactional database**
  - Designed for high performance across many transactions.
  - Transactions are atomic units of work.

- **data warehouse**
  - Output only
  - Optimized for data analysis

ChemBank is a hybrid (transactional + data warehouse)

- data extract

- **transactional & analysis data**
“Extended” molecular fingerprint

- A fingerprint is a vector of properties for a molecule.
- ChemBank search engine spans entire fingerprint.