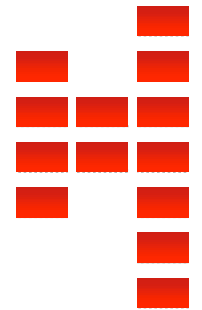
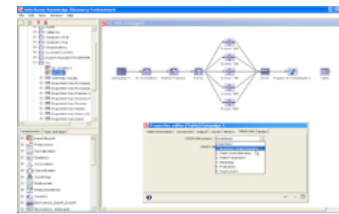
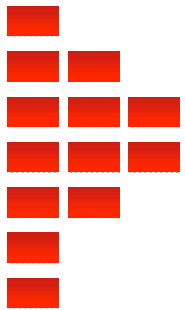


***Rapid Application Development
using InforSense Open Workflow
and Daylight Technologies***



Anthony Arvanites
Daylight User Group Meeting
March 10, 2005

- 1. Company Introduction**
- 2. What Does InforSense Do?**
- 3. Daylight DayCart Integration**
- 4. Demos**
 - **Chemical Warehouse**
 - **Modeling Application**

1. Company Introduction



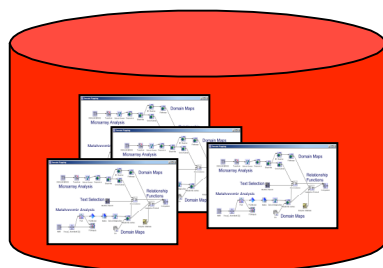
- ▶ Company founded Nov 1999 to commercialize Imperial College Super Computing & Data Mining IP
- ▶ 45 staff plus 20 developers in Shanghai
- ▶ 4 out of top 15 pharma are customers

- ▶ Workflow centric integrative analytics infrastructure
- ▶ Integrate cross-domain data, applications, components into process workflow without programming
- ▶ Capture, manage, deploy & audit analytic workflow processes



2. What does InforSense do?

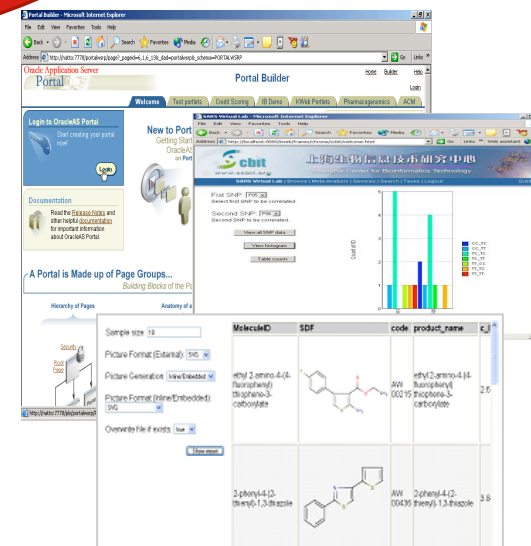
Workflow Warehouse



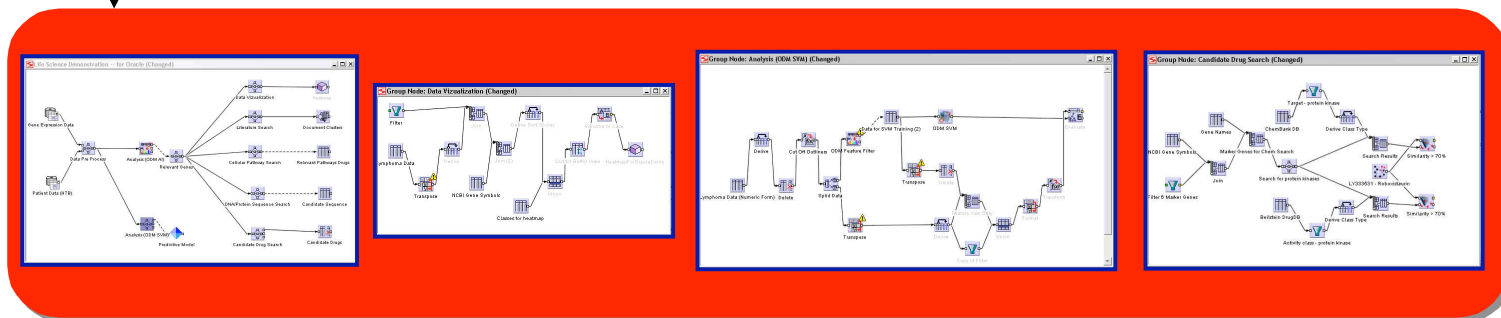
Expert User / Admin

Deployed Web App

Portal / Web















Integrative Analytics Workflow Environment



Data Applications Components



DayLight Functions Integrated into KDE

 Smiles to Mol Conversion Toolkit	 Mol to Smiles Conversion Toolkit	 Structure Standardization DayCart
 Generate Fingerprints DayCart	 Fold Fingerprints DayCart	 Fingerprint Statistics DayCart
 Molecular Properties DayCart	 Substructure Filter DayCart	 Substructure Match DayCart
 Similarity Search DayCart	 Reaction Search DayCart	 Index Creation DayCart

- ▶ **Daylight 'Contrib' Code**
 - Smiles2Mol and Mol2Smiles
- ▶ **DayCart Components**
 - Exact Match, Substructure, Similarity, Structure Standardization, Molecular Properties, Fingerprint, Fingerprint Folding and Statistics, Reaction Handling, and Index Creation
- Installed & created prototype within 8 hours with existing oracle components
- 1 week development time based upon specs



- ▶ **Daylight 'Contrib code' conversion utilities**
 - MOL2SMI and SMI2MOL performs conversion to and from SD and MOL formats.
- ▶ **Structure Standardization**
 - Operators smi2cansm, vcs_desalt and vcs_normalize
 - Remove molecular fragments found in c\$dcischem.salts
 - SMIRKS-based structure normalization on input Smiles based on c\$dcischem.transform
- ▶ **Molecular Properties**
 - Operators smi2netch, smi2hcount, smi2mf, smi2amw
 - Access Dayprop via Dayproptalk from DayCart
- ▶ **Index Creation**
 - Creates indexes ddexact, ddblob, ddgraph, or ddrole



► **Generate Fingerprints**

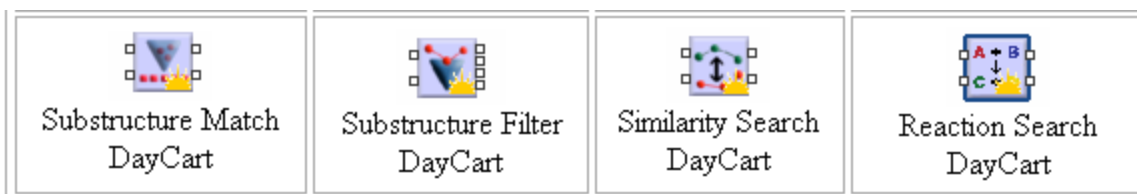
- Operator smi2fp – User able to define min, max and number of bits

► **Fold Fingerprints**

- Operator foldfp – User able to define number of bits and density

► **Fingerprint Statistics**

- Operators bitcount, nbits and isfp – Returns the number of bits and total size of the fingerprint. Also, performs a fingerbit check so that the syntax of a fingerprint can never be confused with a valid SMILES string.



▶ **Substructure Match / Filter**

- Operators exact, contains
- Match searches entire database and returns 1 or 0
- Filter restricts searches for 1 or 0 taking advantage of DayCart chemical indexing

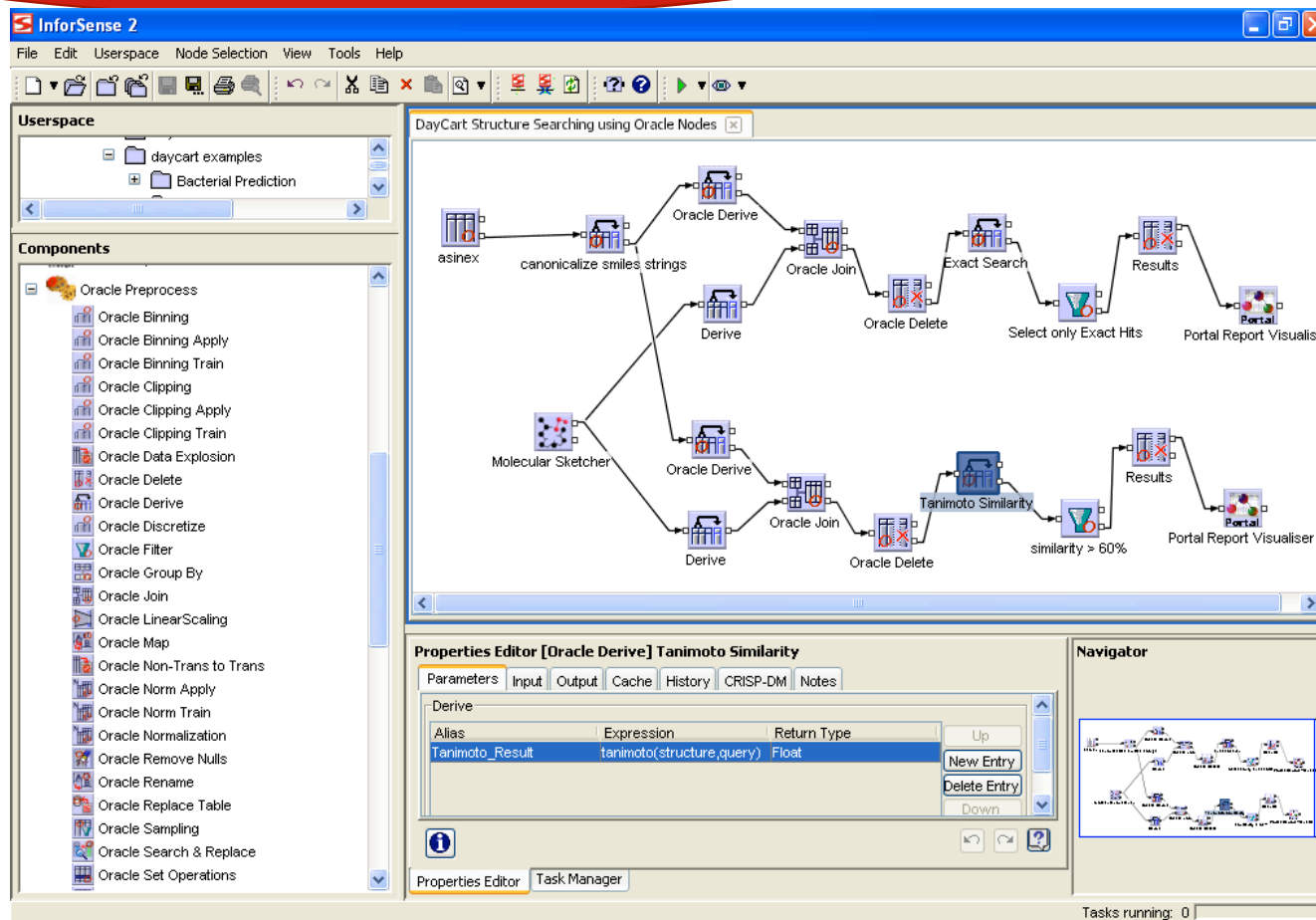
▶ **Similarity Search**

- Operators tanimoto, euclid, Tversky

▶ **Reaction Search**

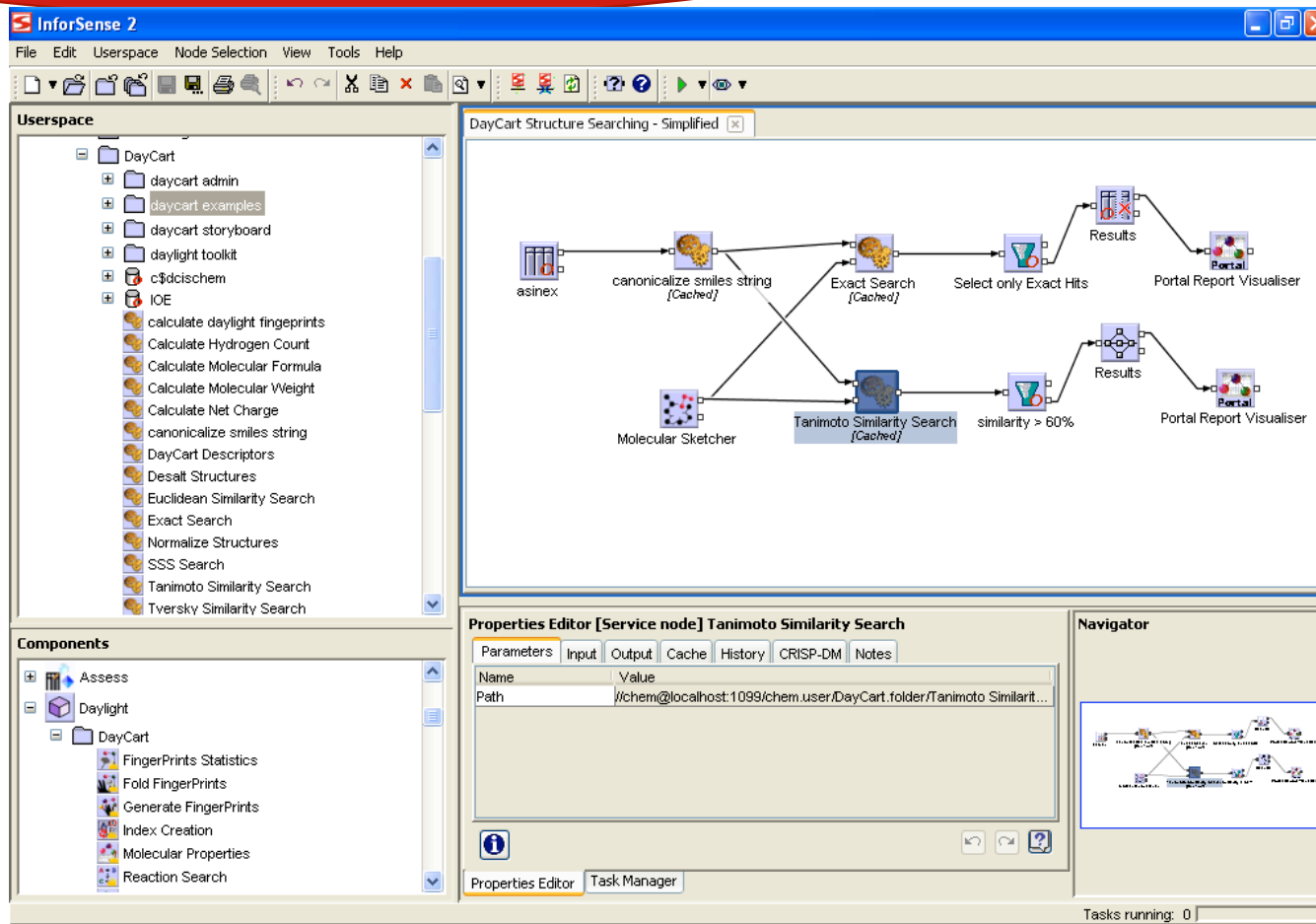
- Operators reactant, agent, product, component.

Working DayCart Prototype Nodes Using Existing IOE nodes



- ▶ Existing IOE derive or search & replace nodes can be quickly used to integrate other DayCart functions (e.g. graph, tautomer, asmiles)

Prototype of Reusable DayCart Service Nodes



- ▶ **Encapsulated and Reusable Group or Service nodes**
- ▶ **Installed & created prototype within 8 hours using existing oracle components**

Chemical Warehouse / Modeling Application



- ▶ **Purpose:** Create a large scale chemical warehouse consisting of commercial vendor and proprietary screening databases that is easily searched and models deployed to bench scientists architected so that all processes are performed in the Oracle databases .
- ▶ **Six major vendor databases (Chembridge, ChemDiv, Enamine, Maybridge, Specs, Interbioscreen) were downloaded in smiles format from the Zinc web site.**
- ▶ **DayCart & Oracle Data Mining (ODM) running in Oracle 10G environment**
- ▶ **Chemical databases downloaded from <http://blaster.docking.org/zinc/>**
 - Ref: Irwin and Shoichet, *J. Chem. Inf. Comput. Sci.* 2005;45(1):177-82



► Requirements - warehouse

- Capability of integrating with other in-house chemical databases (e.g. compound repository, subscription databases, virtual collections)
- Easily create / modify using a visual programming environment
- Capability of performing external or in-database data-mining analytics
- Support chemical registrations
- Capability of integrating with other databases (e.g. MDL, Activity Base)

► Requirements - model

- ADME predictions
- Deployment of application to benefit bench scientists
- Oracle Data Mining Support Vector Machine (SVM) model
- Decision Tree Classification models
- MOE descriptors and/or Daylight Fingerprints
- Blood Brain Barrier Permeation(1670 compounds classified BBB+ or BBB-)
- P-glycoprotein (PgP), Human Intestinal Absorption (HIA), Torsades de Pointes (TdP)

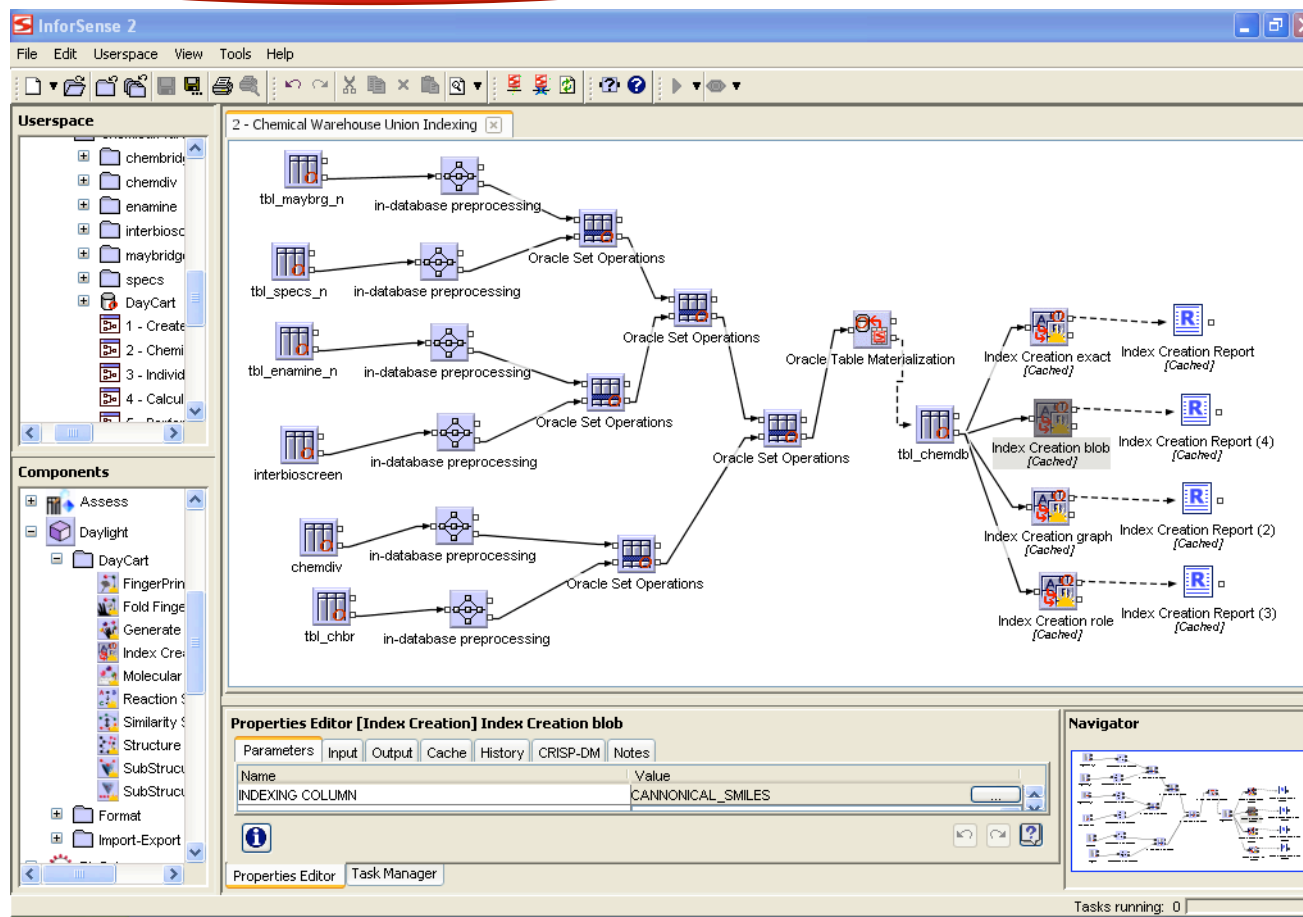
Data Entry and Processing of Commercial Databases using DayCart

The screenshot displays the InforSense 2 software interface. On the left, the 'Userspace' pane shows a hierarchical tree of folders including 'MUG 2005', 'ADME', 'Chemical/Warehouse', and 'DayCart'. Below it, the 'Components' pane lists various tools like 'FingerPrints Statistics', 'Fold FingerPrints', and 'Structure Standardisation'. The main workspace shows a workflow diagram with three parallel paths: one for 'maybridge' data, one for 'specs' data, and one for 'enamine' data. Each path starts with an 'Oracle Table Materialization' step, followed by 'Structure Standardisation', 'Molecular Properties', and 'Generate FingerPrints'. The 'maybridge' path ends with 'Oracle Delete' and 'Oracle Table Materialization' to create 'tbl_maybridge'. The 'specs' path ends with 'Oracle Table Materialization' to create 'tbl_specs'. The 'enamine' path ends with 'Oracle Table Materialization' to create 'tbl_enamine'. A 'Properties Editor' for 'Structure Standardisation' is open, showing a table of parameters and their values.

Name	Value
SMILES COLUMN	SMILES
NEW COLUMN NAME	Canonical_Smiles
UNIQUE	<input checked="" type="checkbox"/>
REMOVE SALTS	<input checked="" type="checkbox"/>
SALT CLASS	1999
NORMALISE STRUCTURE	<input checked="" type="checkbox"/>
NORMALISE CLASS	1999

Description: Individual commercial compound vendor database (e.g. Maybridge) are entered into an Oracle database. Structure Standardization, molecular properties and fingerprint operations are being performed using Daycart to create a table all within the Oracle Database

Creating a chemical warehouse via in-database processing



Description: Six commercial ‘cleaned’ compound vendor databases are being preprocessed to unionize matching data columns (e.g. structure, CAS number, name, availability) creating a master chemical data warehouse within Oracle. DayCart structure Indexing (exact, role, graph, blob) was performed to enable fast structure searching on over over 1 million compounds.

DayCart Chemical Substructure and Similarity Searching

The screenshot shows the InforSense 2 software interface. The main window displays a workflow for chemical searching. The workflow starts with 'Sketch Molecule', followed by 'To Smiles [Cached]', 'Similarity Search [Cached]', and 'Pivot - Create Report [Cached]'. A 'Molecule Sketch' window shows a query molecule (2-phenylpyridine). A 'Table Editor' window shows search results with columns for CANNONICA, Chembridge, ChemDiv, Eranite, Interbioscreen, Maybridge, Specs, and Query. The results table is highlighted with the word 'results'.

CANNONICA	Chembridge	ChemDiv	Eranite	Interbioscreen	Maybridge	Specs	Query
		ZINC00040046				ZINC00040046	
	ZINC00036301	ZINC00036301				ZINC00036301	
		ZINC00035154				ZINC00035154	
						ZINC00043898	

Description: A similarity or substructure search is being performed using DayCart. Results are filtered and pivoted to produce a report of structure, query, and vendor with catalog numbers for ordering.

Calculate Molecular Descriptors using MOE

The screenshot shows the InforSense 2 software interface. The main workspace displays a workflow diagram with the following steps: Sketch Molecule, To Smiles, SubStructure Filter sss, MOEDESC, and Oracle Table Materialization. The MOEDESC node is highlighted, and its properties are shown in the Properties Editor. The Properties Editor for MOEDESC shows the following parameters:

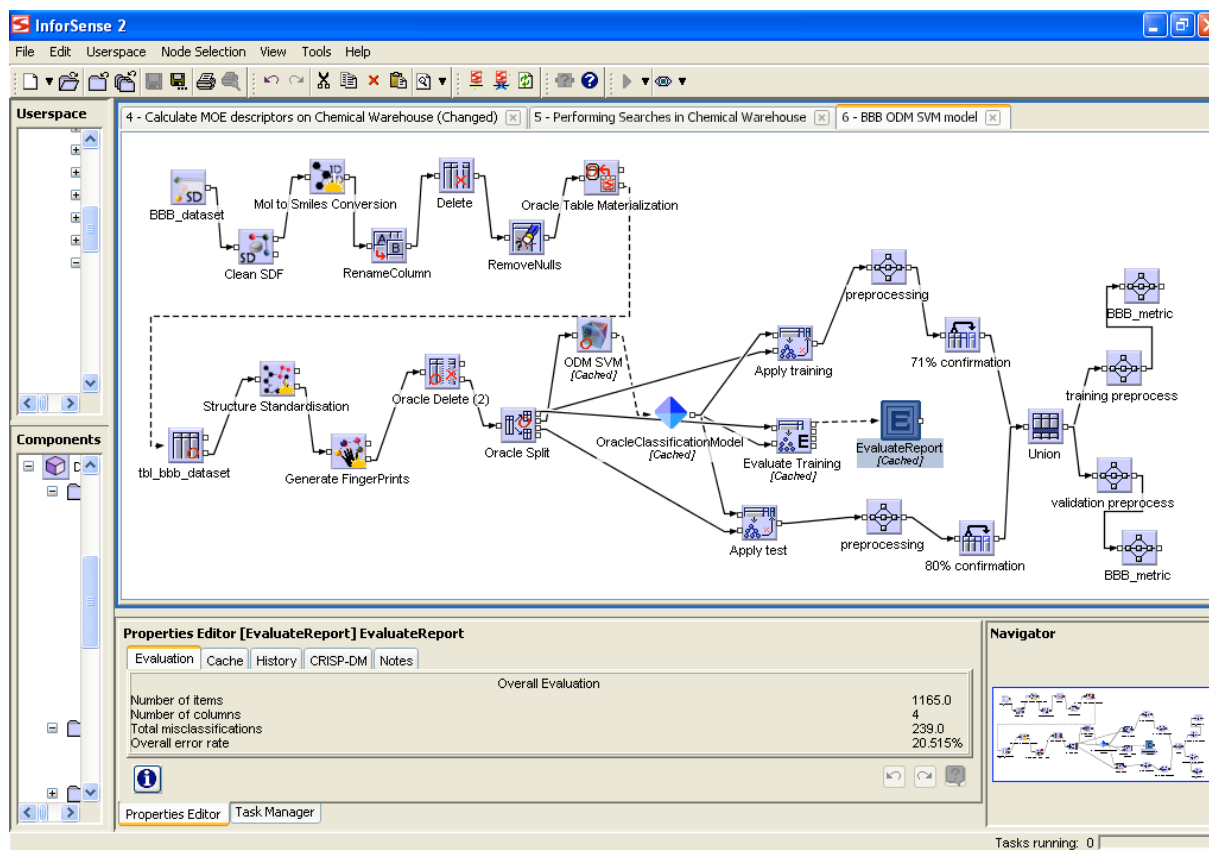
Parameter	Value
Name	CANNONICAL_SMILES
molecule column	
select descriptors	[2D:151]
descriptors	'apol','a_acc','a_acid','a_aro','a_base','a_cou
2D -> 3D	<input type="checkbox"/>
optimize	<input type="checkbox"/>
wash	<input type="checkbox"/>

The Interactive Browser window displays two bar charts and a table of chemical structures. The first bar chart shows the Count of H_COUNT, and the second bar chart shows the Count of AVG_MOL_WEIGHT. The table below shows the results of the MOE descriptor calculation:

ZINCID	SUPPLIER	CANNONICAL...	NET_CHA...
ZINC00041745	Interbioscreen	E1cc1#H)tc2c0	
ZINC00040046	enamine	Dctccn#H)X0	
ZINC00040046	specs	Cctccp#H)X0	
ZINC00031414	enamine	NcClcn#H)tc1c0	
ZINC00031414	specs	NcClcn#H)tc1c0	
ZINC00032686	chemdiv	Dctccn#H)X0	
ZINC00032686	specs	Dctccn#H)X0	
ZINC00001757	maybridge	Dctccn#H)X0	
ZINC00032601	chemdiv	Dctccn#H)X0	
ZINC00032601	chemdiv	Dctccn#H)X0	
ZINC00035154	chemdiv	Dctccn#H)X0	
ZINC00035154	specs	Dctccn#H)X0	
ZINC00045272	maybridge	FcC(F)X)tc1cc0#H	

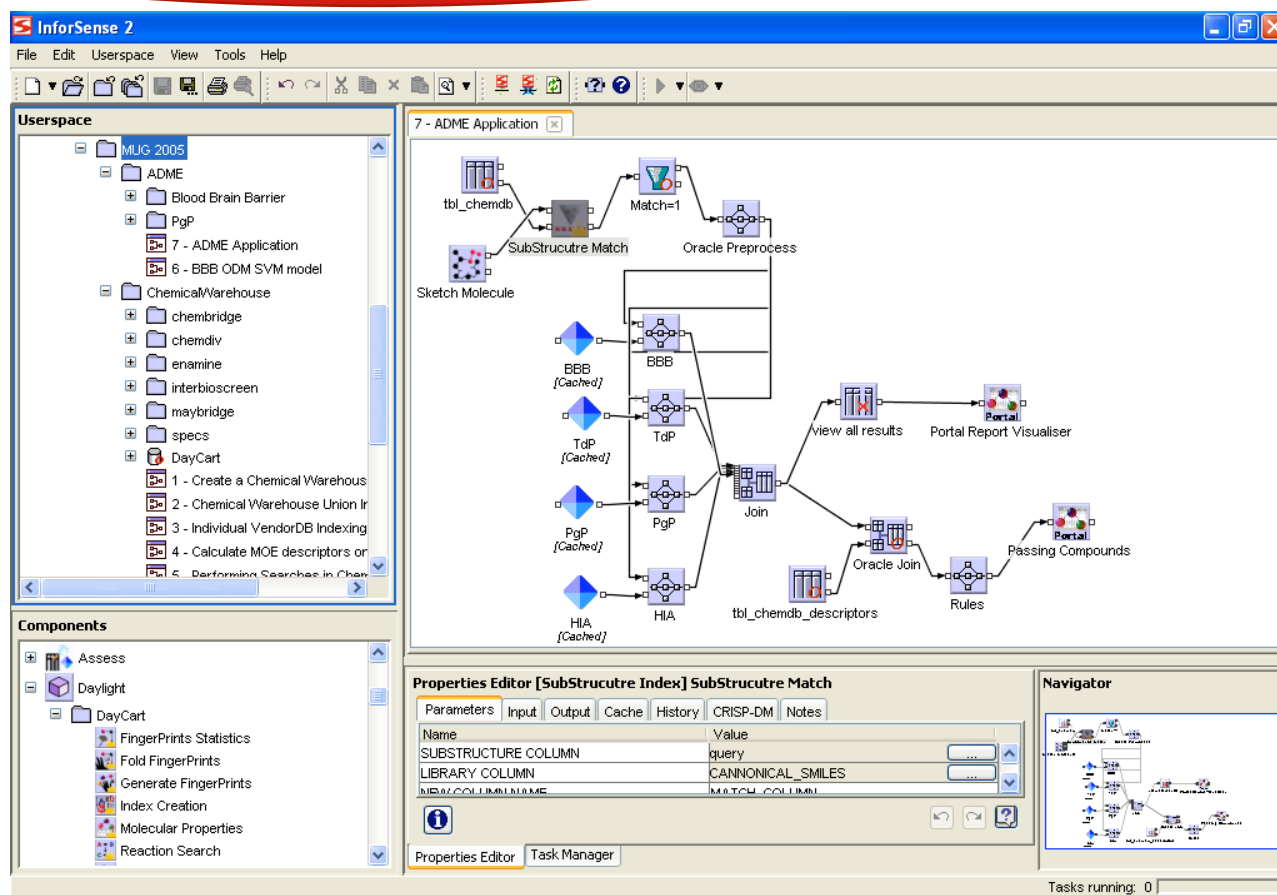
Description: MOE descriptors (2D) are being calculated on a selection from the chemical warehouse. Results can either be used to create a MOE database or materialized into an Oracle table. Calculated molecular descriptors can be graphically viewed in an interactive browser.

KDE / IOE Blood Brain Barrier Model



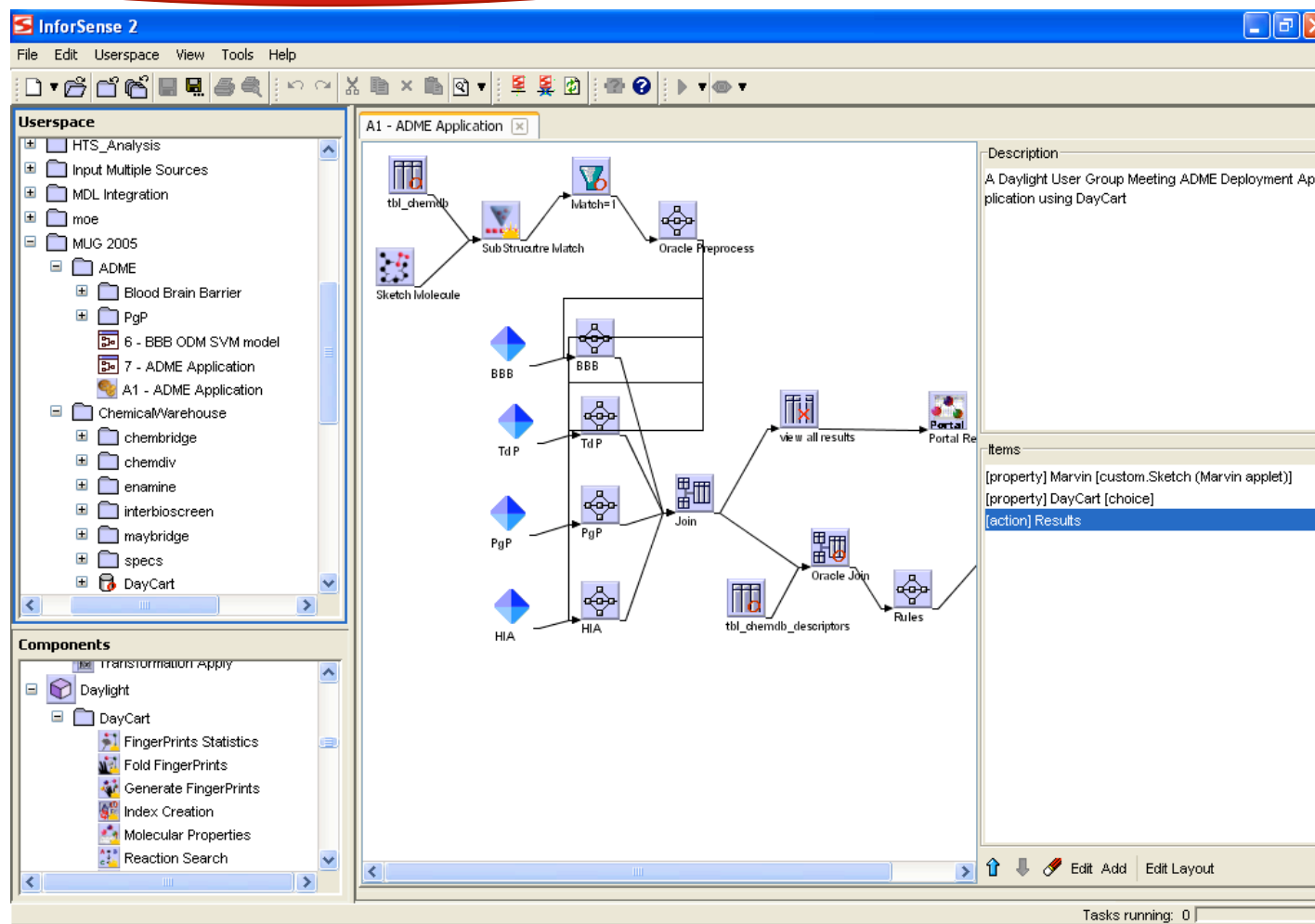
Description: A Blood Brain Barrier dataset in SD format is converted into Daylight Smiles via 'contrib' code and imported into an Oracle table. The dataset is standardized, daylight fingerprints calculated and used to produce an SVM model all within the Oracle database. Results are then processed to calculate evaluation metrics to determine the efficiency of the model.

Modeling Application - ADME



Description: Results of a chemical search, performed in DayCart, are used to predict Blood Brain Barrier (oracle SVM), P-glycoprotein (Decision Tree), human intestinal absorption (Weka j48 DT classifier), Torsades de Pointes (oracle SVM). Predicted results are deployed via a web portal.

Deployed Application to Bench Scientists



Description: The predictive data mining ADME suite is deployed to scientists within InforSense KDE.

Deployed Application to Bench Scientists

InforSense Discovery Portal - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address: http://localhost:8090/kweb/do/service/execute

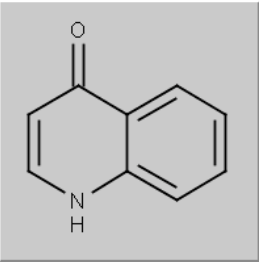
InforSense Discovery Portal

Admin Browse Services Tasks Help

A1 - ADME Application for Daylight meeting

A Daylight User Group Meeting ADME Deployment Application using DayCart

Marvin



Please sketch a molecule

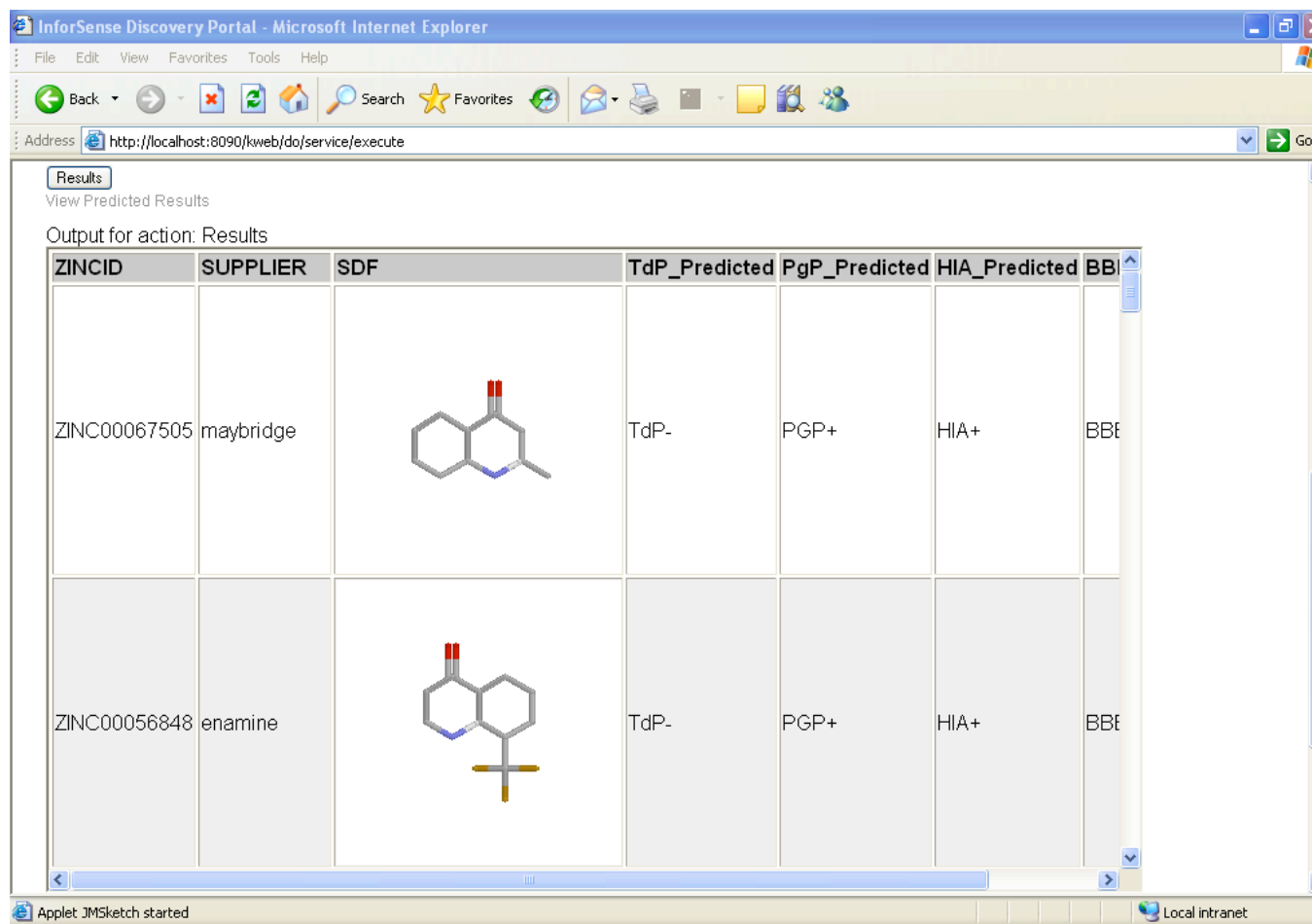
SearchType

Applet JMsketch started Local intranet


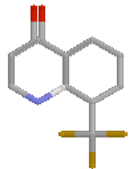
Description: The deployed predictive ADME application is presented via a web portal. Scientists can sketch a molecule using Marvin, select the type of search to be performed using DayCart, and results will be presented in tabular format.

Continued on next slide

Deployed Application to Bench Scientists



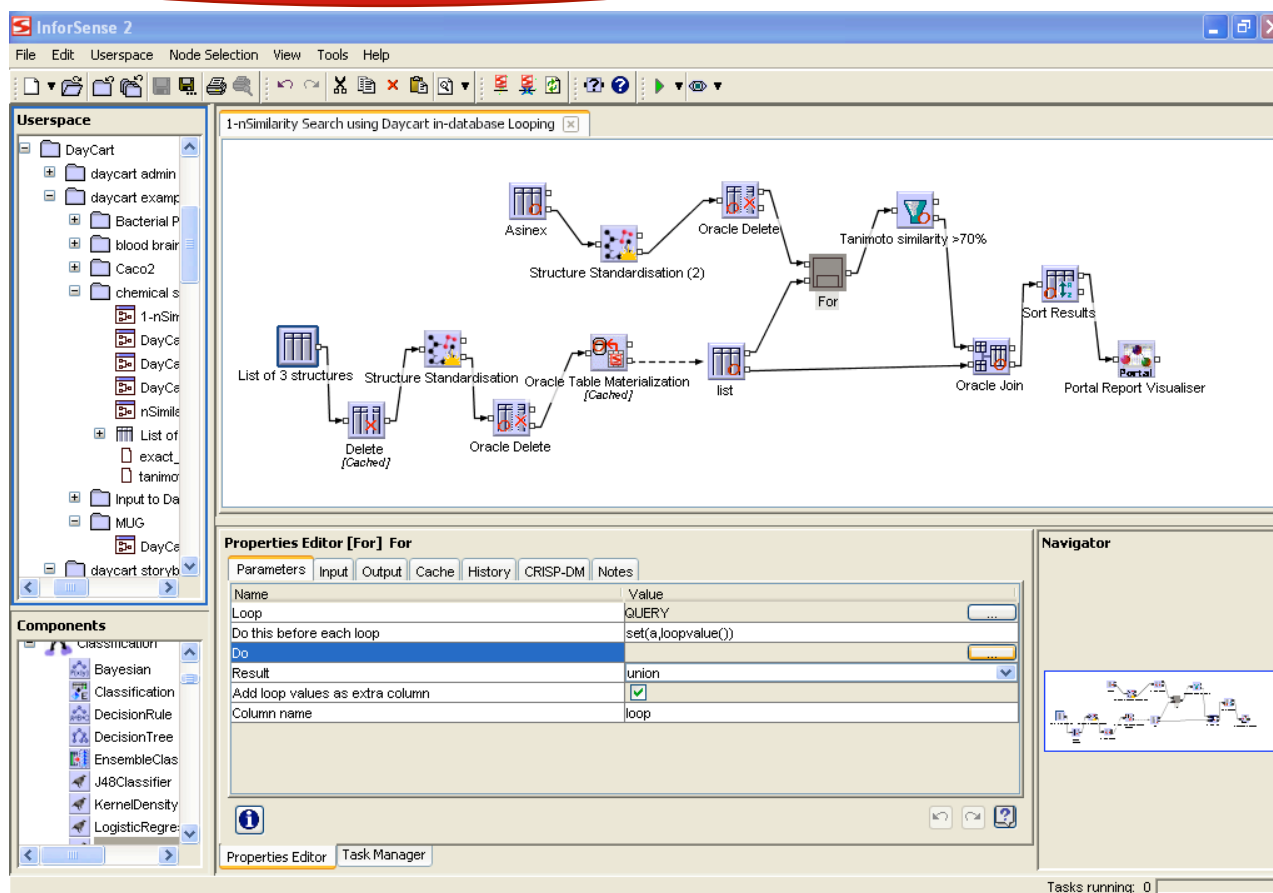
Results
View Predicted Results
Output for action: Results

ZINCID	SUPPLIER	SDF	TdP_Predicted	PgP_Predicted	HIA_Predicted	BB
ZINC00067505	maybridge		TdP-	PGP+	HIA+	BB
ZINC00056848	enamine		TdP-	PGP+	HIA+	BB

Applet: JMSketch started
Local intranet

Description: Chemical structures are shown with predicted ADME results (TdP, PgP, HIA and BBB). The chemical structures are visualized using MDL Chime.

Multiple in-database chemical searching using Control (Looping) Functionality



Description: A list of chemical structures are standardized and used as queries to be performed against a vendor database. The For control looping mechanism search each query individually and filters out all compounds with a Tanimoto similarity of less than 70 %. These processes occur all within the Oracle database.

- ▶ **J. Chem. Inf. Comput. Sci., Vol. 44, pg 1630-1638 (2004) Effect of Molecular Descriptor Feature Selection in Support Vector Machine Classification of Pharmacokinetic and Toxicological Properties of Chemical Agents.**
- ▶ **J. Chem. Inf. Comput. Sci., Vol. 44, No.4 (2004), Prediction of P-Glycoprotein Substrates by a Support Vector Machine Approach.**
- ▶ **J. Chem. Inf. Comput. Sci., Vol. 38, 726-235 (1998) Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure.**
- ▶ **J. Chem. Inf. Comput. Sci., Vol. 44, No.1 (2004) Blood-Brain Barrier Permeation Models: Discrimination between Potential CNS and Non-CNS Drugs Including P-Glycoprotein Substrates.**

Contact InforSense

InforSense Limited U.K

48 Princes Gardens
London SW7 2PE (UK)

Telephone: +44 (0) 20 7594 6817
Fax: +44 (0) 20 7594 6836

InforSense Limited U.S.A

25 Moulton Street
Cambridge, MA 02138 (USA)

Telephone: +1 617 547 2500

Ryoka Systems Inc.

1-28-38 Shinkawa, Chuo-Ku, Tokyo
104-0033 (Japan)

Telephone: 03-3553-6341

Additional information can be found at

www.inforsense.com