UMDNJ
ROBERT WOOD JOHNSON MEDICAL SCHOOL
University of Medicine & Dentistry of New Jersey

Next Generation Computational Chemistry Tools to Predict Toxicity of CWAs

William (Bill) Welsh

welshwj@umdnj.edu

Prospective Funding by DTRA/JSTO-CBD

ebCTC environmental bioinformatics and Computational Toxicology Center

A State-wide, Regional and National Resource

< www.ebCTC.org >

Funded with support from the U.S. EPA

Consortium Members

UMDNJ
Robert Wood Johnson Medical School
University of Medicine & Dentistry of New Jersey

The State University of New Jersey
Rutgers

FDA U.S. Food and Drug Administration
Center for Toxicoinformatics, NCTR

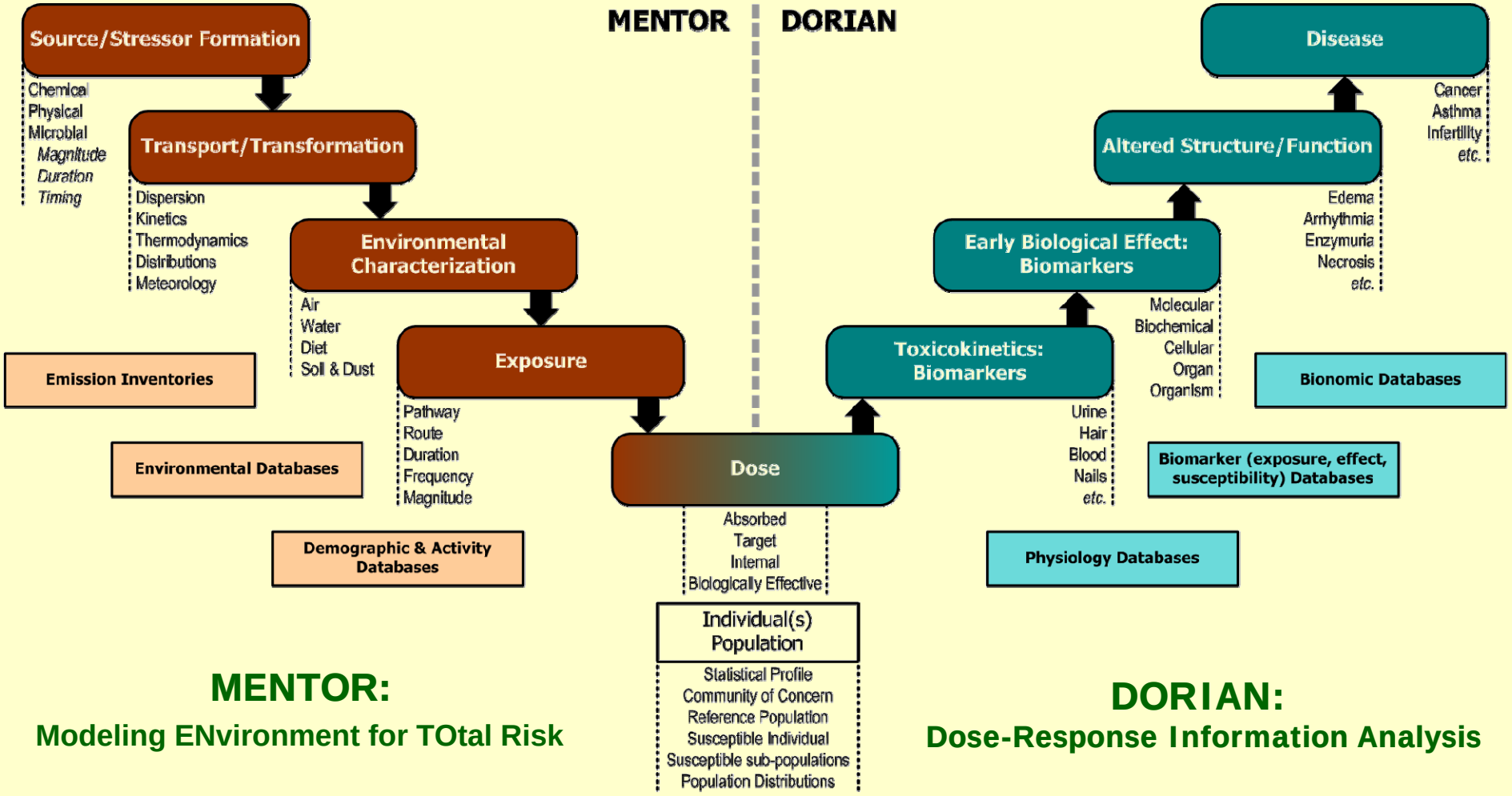
Princeton University

Major Research Thrusts

- MENTOR-DORIAN Computational Toxicology System that spans the Source->Dose->Outcome continuum
- The Environmental Bioinformatics Knowledge Base (ebKB: www.ebCTC.org)
- *ArrayTrack*: toxicological bioinformatics platform to process genomics, proteomics and metabonomics data
- Hepatocyte Metabolic Model for Xenobiotics
- **ChemTox, a suite of chem-informatics tools for toxicant identification & characterization**

MENTOR & DORIAN

Address the Source-to-Outcome Continuum



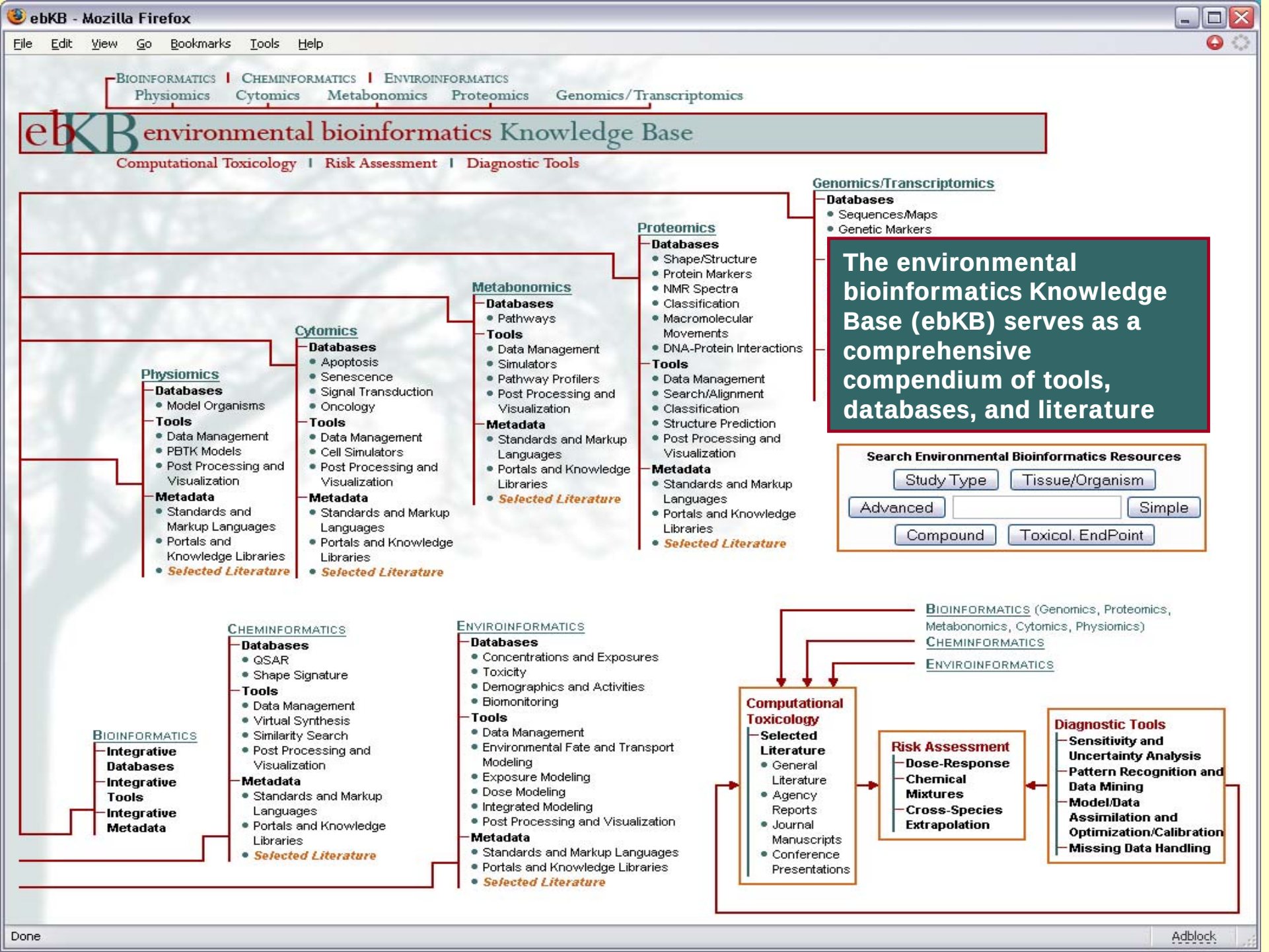
MENTOR:

Modeling ENvironment for TOtal Risk

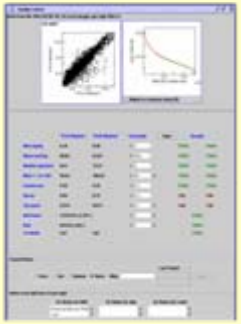
DORIAN:

Dose-Response Information Analysis

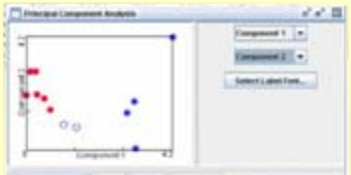
Adapted from chart by R. Calderon, USEPA/NHEERL, 2003


ebKB - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics | Cheminformatics | Enviroinformatics
Physiomics Cytomics Metabonomics Proteomics Genomics/Transcriptomics
ebKB environmental bioinformatics Knowledge Base
Computational Toxicology | Risk Assessment | Diagnostic Tools
Genomics/Transcriptomics
Databases
Sequences/Maps
Genetic Markers
Proteomics
Databases
Shape/Structure
Protein Markers
NMR Spectra
Classification
Macromolecular Movements
DNA-Protein Interactions
Tools
Data Management
Search/Alignment
Classification
Structure Prediction
Post Processing and Visualization
Metadata
Standards and Markup Languages
Portals and Knowledge Libraries
Selected Literature
Metabonomics
Databases
Pathways
Tools
Data Management
Simulators
Pathway Profilers
Post Processing and Visualization
Metadata
Standards and Markup Languages
Portals and Knowledge Libraries
Selected Literature
Cytomics
Databases
Apoptosis
Senescence
Signal Transduction
Oncology
Tools
Data Management
Cell Simulators
Post Processing and Visualization
Metadata
Standards and Markup Languages
Portals and Knowledge Libraries
Selected Literature
Physiomics
Databases
Model Organisms
Tools
Data Management
PBTK Models
Post Processing and Visualization
Metadata
Standards and Markup Languages
Portals and Knowledge Libraries
Selected Literature
The environmental bioinformatics Knowledge Base (ebKB) serves as a comprehensive compendium of tools, databases, and literature
Search Environmental Bioinformatics Resources
Study Type
Tissue/Organism
Advanced
Simple
Compound
Toxicol. EndPoint
Cheminformatics
Databases
QSAR
Shape Signature
Tools
Data Management
Virtual Synthesis
Similarity Search
Post Processing and Visualization
Metadata
Standards and Markup Languages
Portals and Knowledge Libraries
Selected Literature
Enviroinformatics
Databases
Concentrations and Exposures
Toxicity
Demographics and Activities
Biomonitoring
Tools
Data Management
Environmental Fate and Transport Modeling
Exposure Modeling
Dose Modeling
Integrated Modeling
Post Processing and Visualization
Metadata
Standards and Markup Languages
Portals and Knowledge Libraries
Selected Literature
Bioinformatics
Integrative Databases
Integrative Tools
Integrative Metadata
Bioinformatics (Genomics, Proteomics, Metabonomics, Cytomics, Physiomics)
Cheminformatics
Enviroinformatics
Computational Toxicology
Selected Literature
General Literature
Agency Reports
Journal Manuscripts
Conference Presentations
Risk Assessment
Dose-Response
Chemical Mixtures
Cross-Species Extrapolation
Diagnostic Tools
Sensitivity and Uncertainty Analysis
Pattern Recognition and Data Mining
Model/Data Assimilation and Optimization/Calibration
Missing Data Handling
Done
Adblock

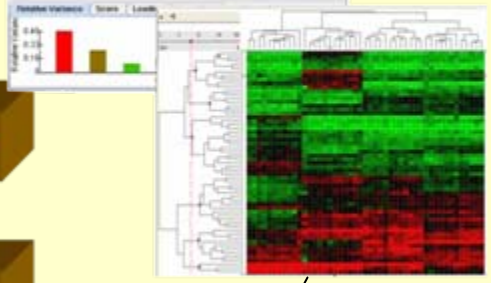
ArrayTrack Suite of Bioinformatics Tools



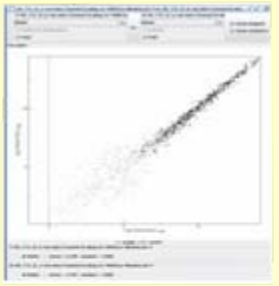
Data input and QC



PCA



2-way HCA



Scatter Plots

Normalization methods

Expression pattern bar charts

Data Import

Normalization

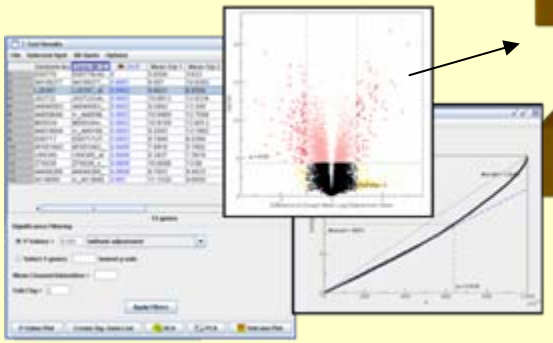
Apply to

Gene Selection

Data Exploration

Apply to

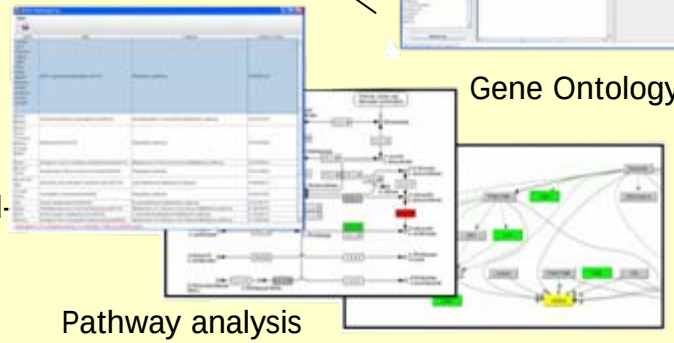
Interpretation



Identify significant genes

- **Volcano Plot** (considering both *p* and fold-change
- **P-value Plot** (considering false positives/negatives)

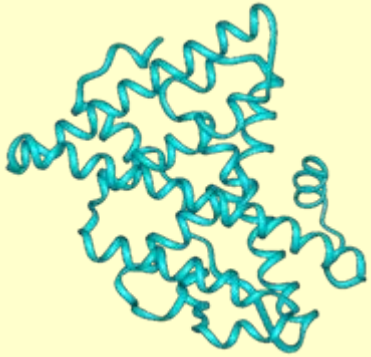
Gene Ontology



Pathway analysis

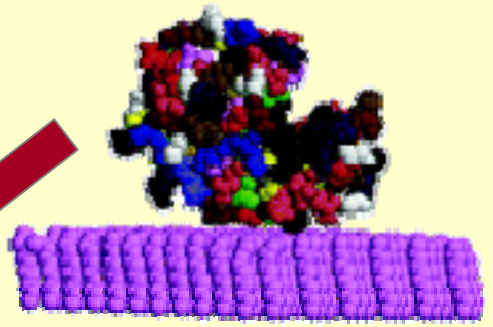
Individual gene analysis

ChemTox, an Integrated Suite of Cheminformatics Tools

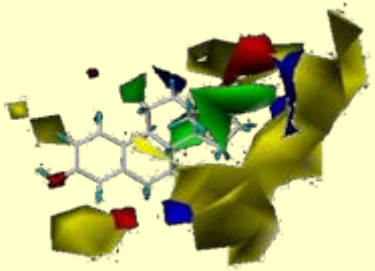


Protein Modeling

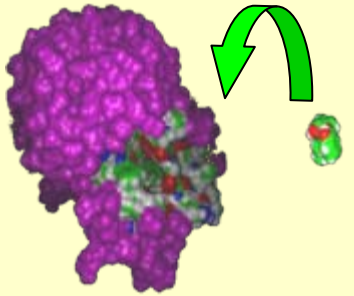
Predictive Molecular Toxicology



Molecule-Surface Interactions
< skin, water, polymer >



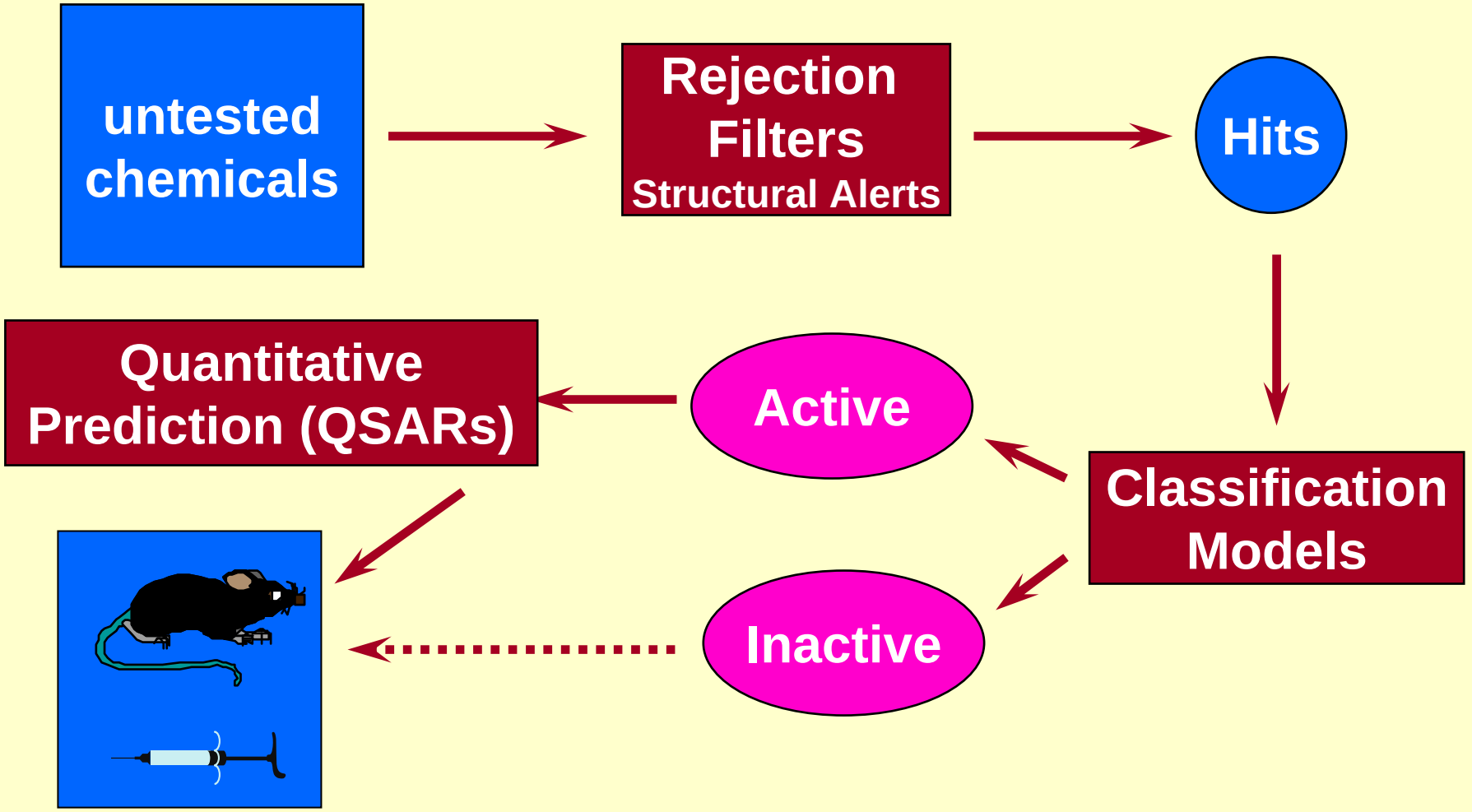
Chemical Modeling



Database Mining / Pattern Recognition

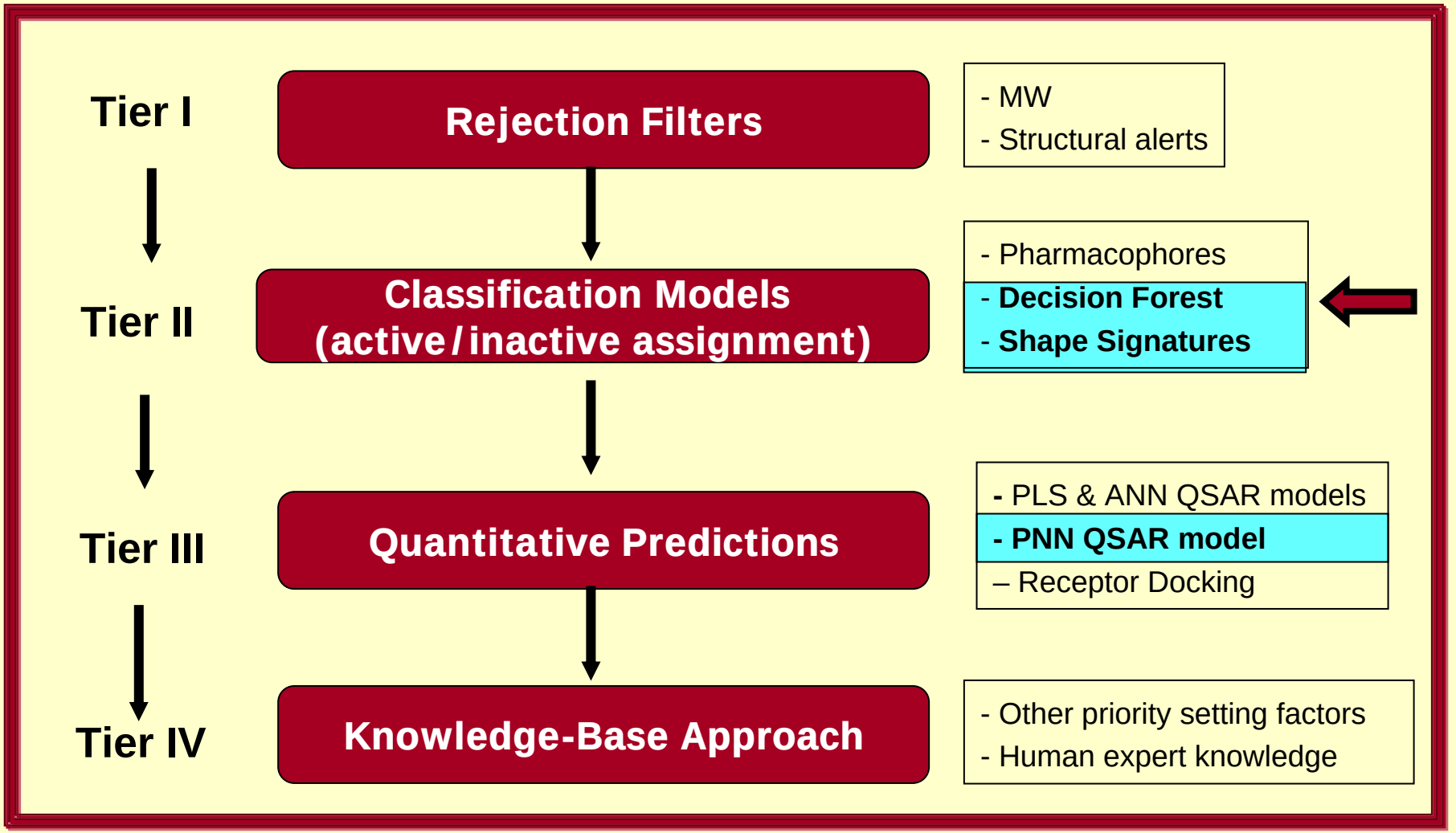
Computational Screening Paradigm

- Priority Setting -



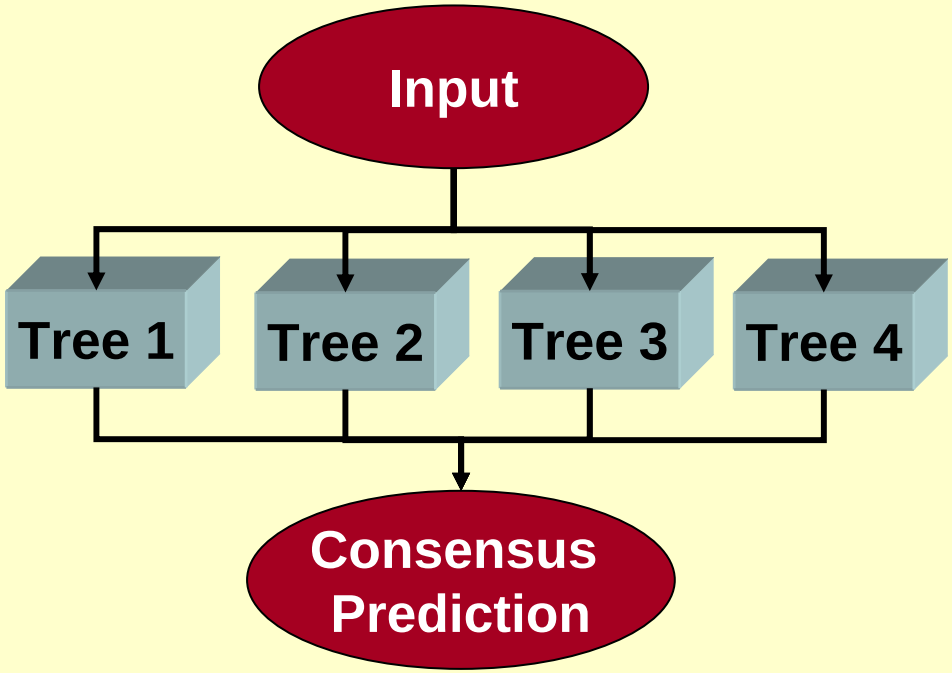
Hierarchical Screening Framework

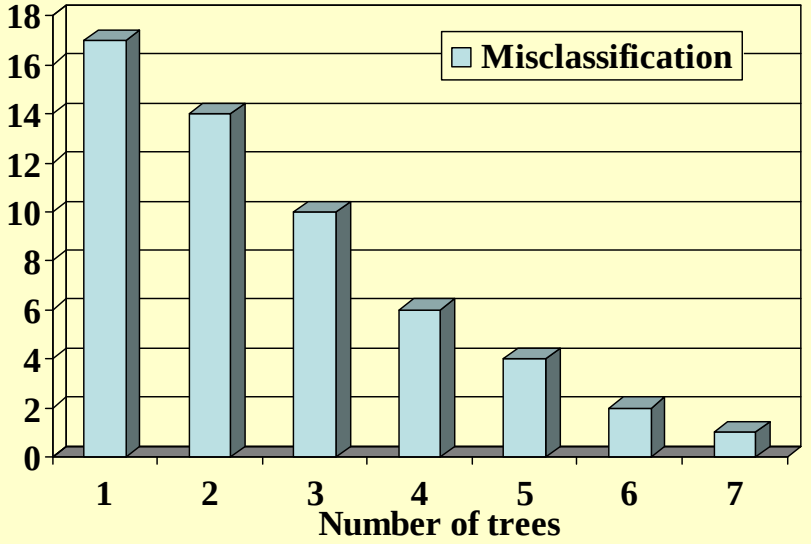
- addresses the need to minimize *false negatives* and *uncertainties*
- recognizes that no single computational model is adequate



Decision Forest

- Improved classification by combining independent Decision Tree models -



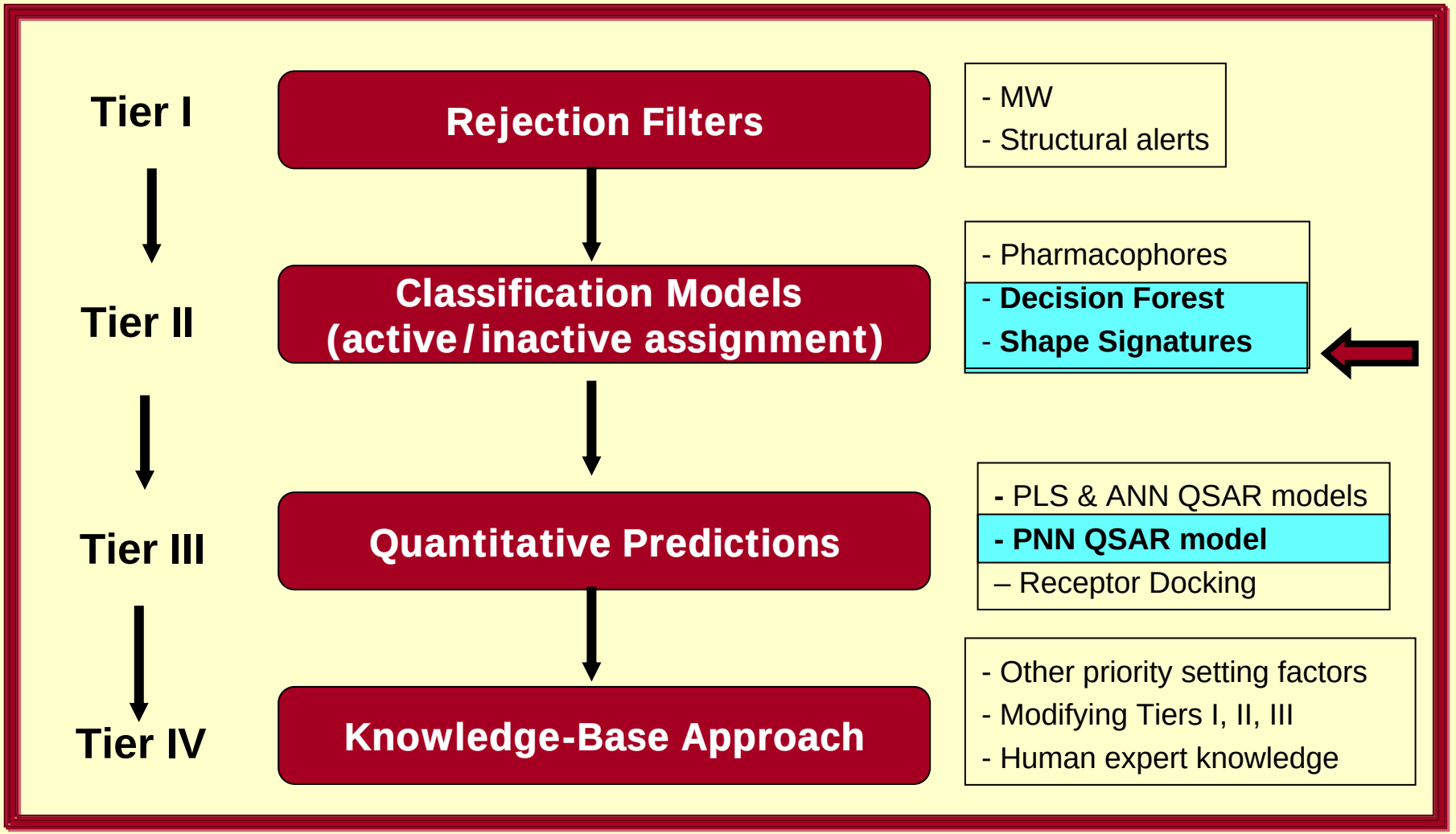


Key Features

- **Combining several independent yet predictive trees reduces misclassification**
- **DF structure permits assessment of prediction confidence**
- **Each tree consists of simple 'If-Then' branches, hence the DF is extremely fast**

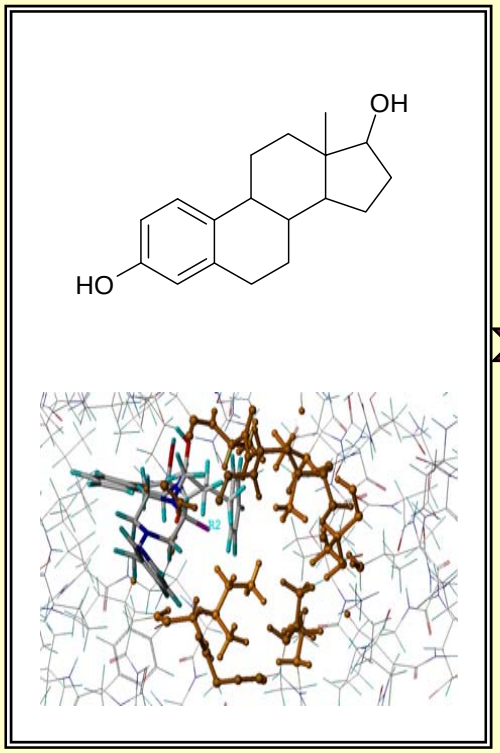
Schematic of Hierarchical Framework

- addresses the need to minimize *false negatives* and *uncertainties* -



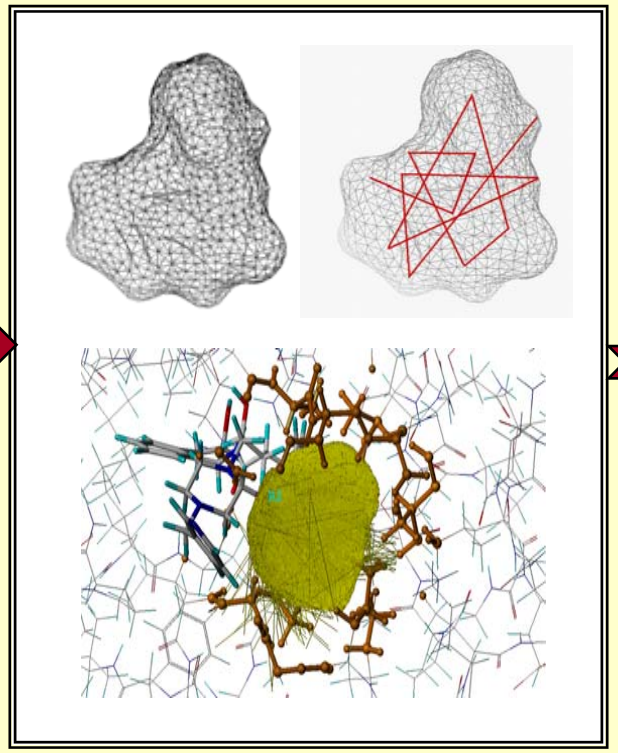
Shape Signatures Tool

START



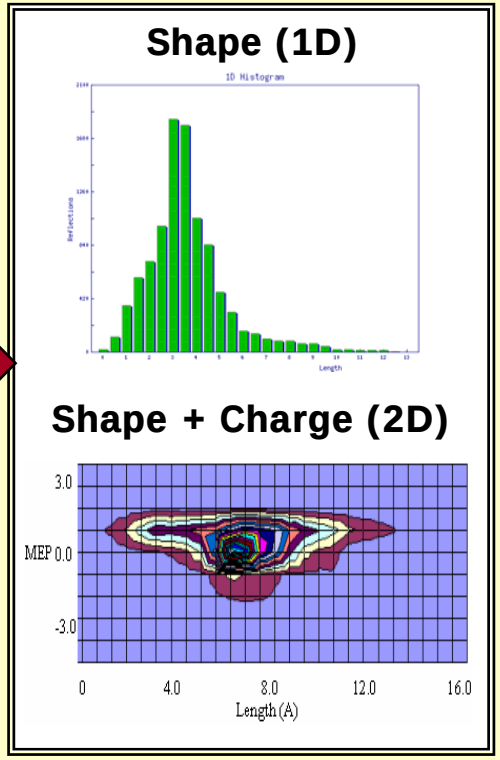
Small molecule or Protein binding pocket

PROCESSING



Ray tracing to generate the raw data

OUTPUT

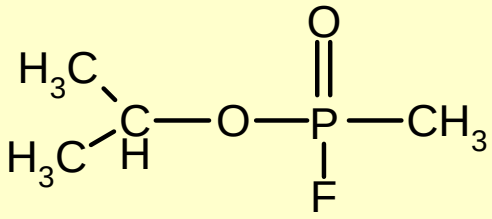


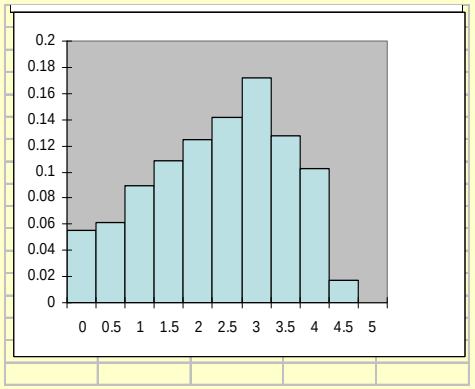
1D and 2D Shape Signatures

Shape Signatures Tool

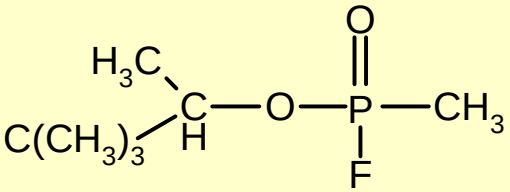
molecules are compared by subtracting their histograms

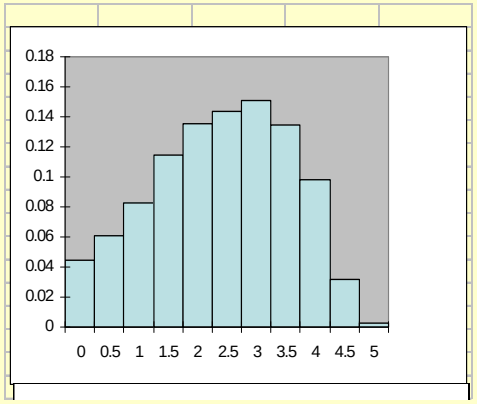
Sarin





Soman

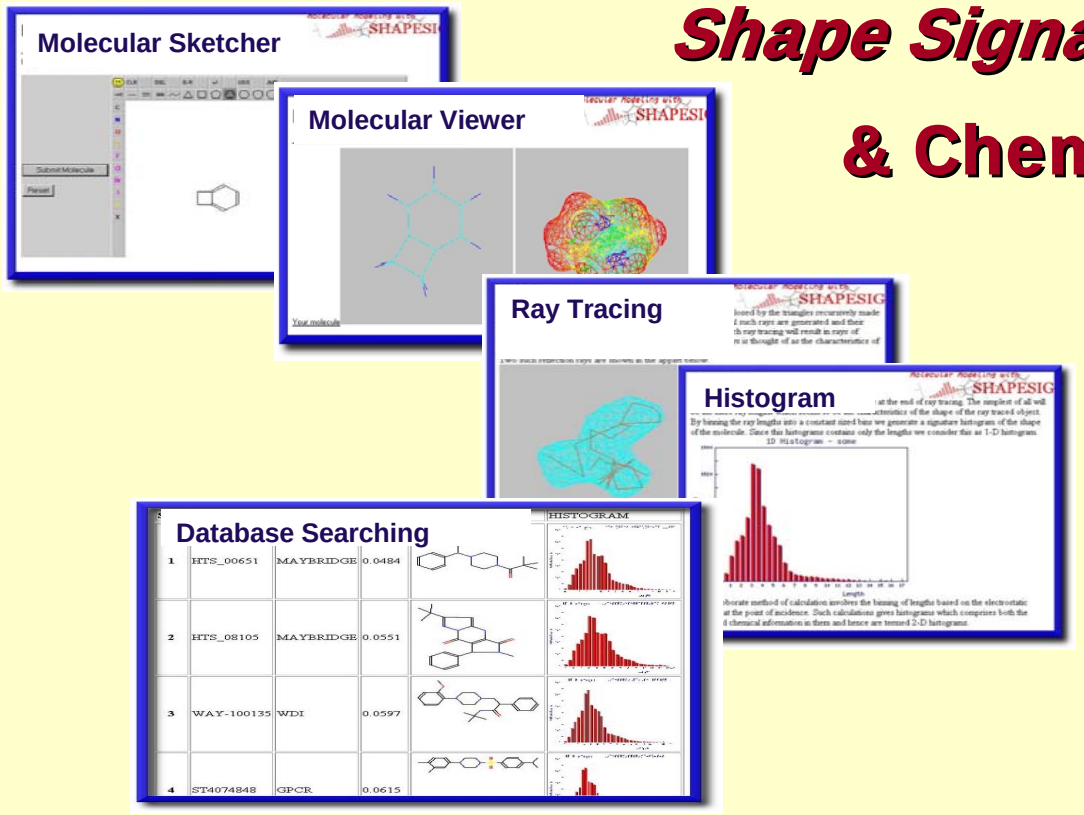




***Diff* = 0.022**

Small *Diff* value means that two molecules have similar shape and polarity

Shape Signatures Software Tool & Chemical Databases

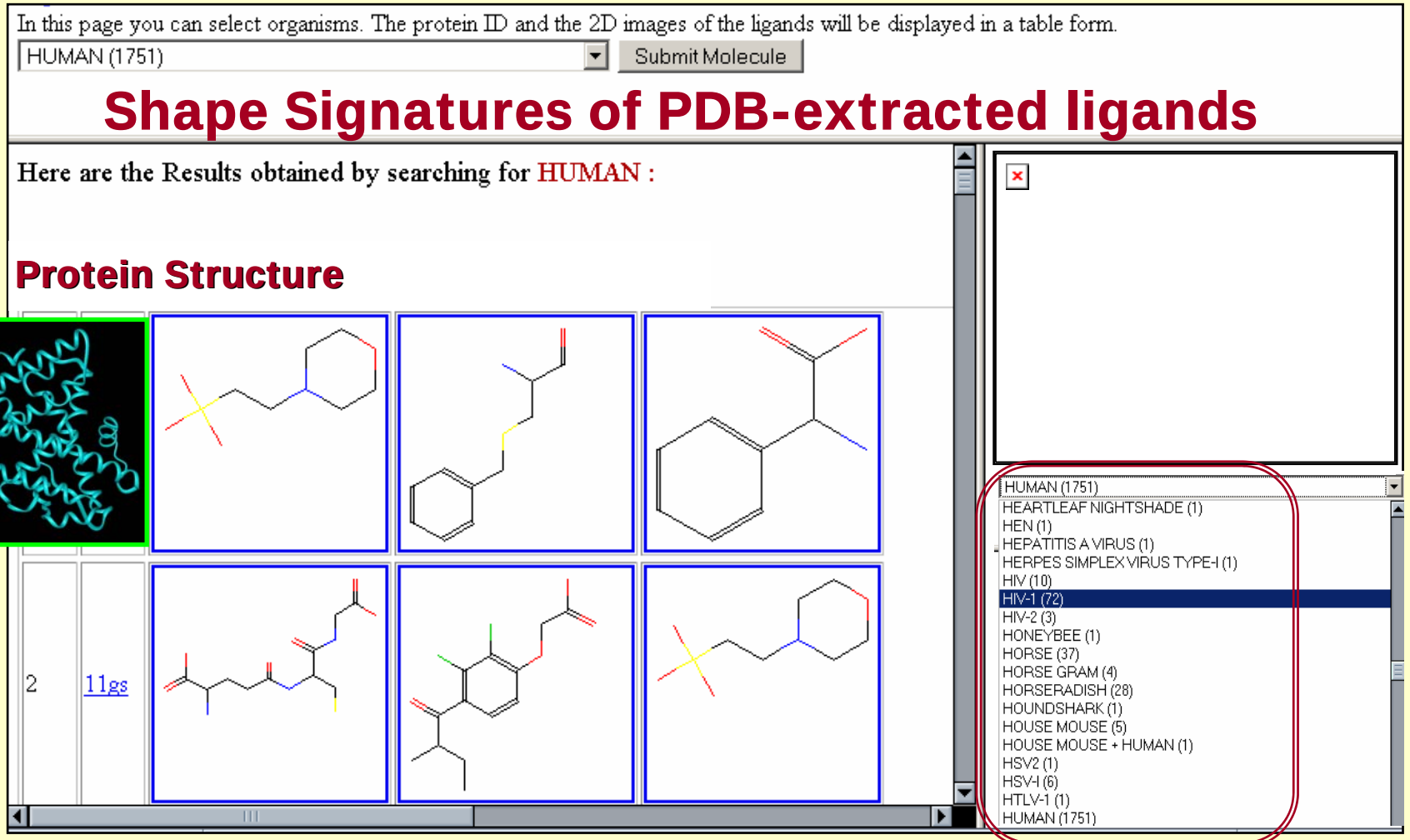


Searchable *Shape Signatures* Databases

- **3+ million commercially available organic compounds**
- **40,000 Natural Products**
- **Hazardous Chemicals (pesticides, nerve agents, mustards, psychotropic agents, other real or potential CWAs, TICs)**
- **PDB-extracted ligands**

Chemical → Target Protein → Mechanisms

Protein Data Bank (PDB): World Repository of ~35,000
Protein-Ligand Crystal Structures (http://www.rcsb.org/pdb/)



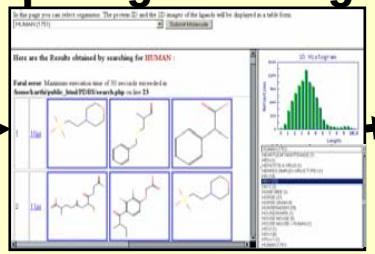
Species / Protein Family

Molecules → Target Protein → Mechanism

Shape Sigs PDB Ligands

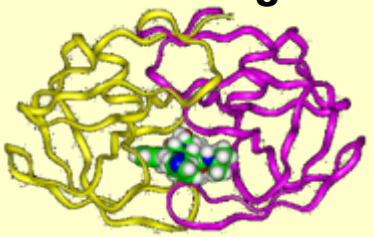
Protein's Binding Site

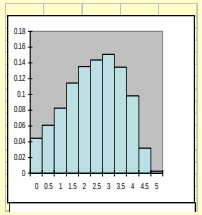
Query Molecule

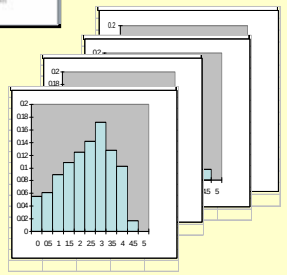


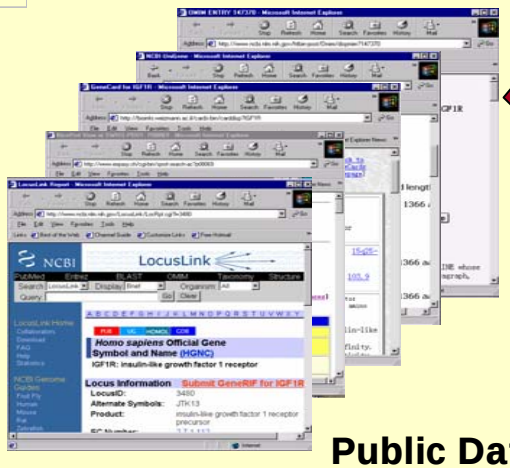
Matching PDB Ligands

Target Proteins





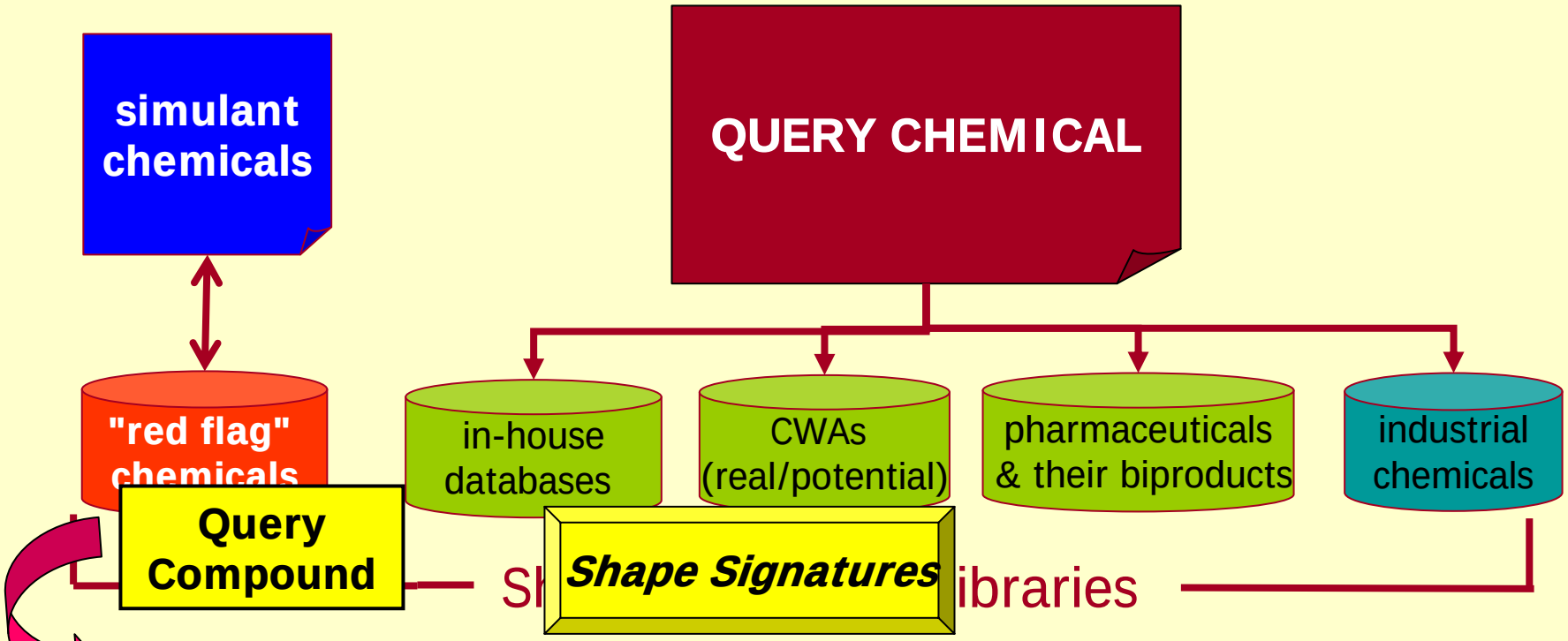




Links to Toxicity Pathways

Public Databases

Identifying Problem Chemicals

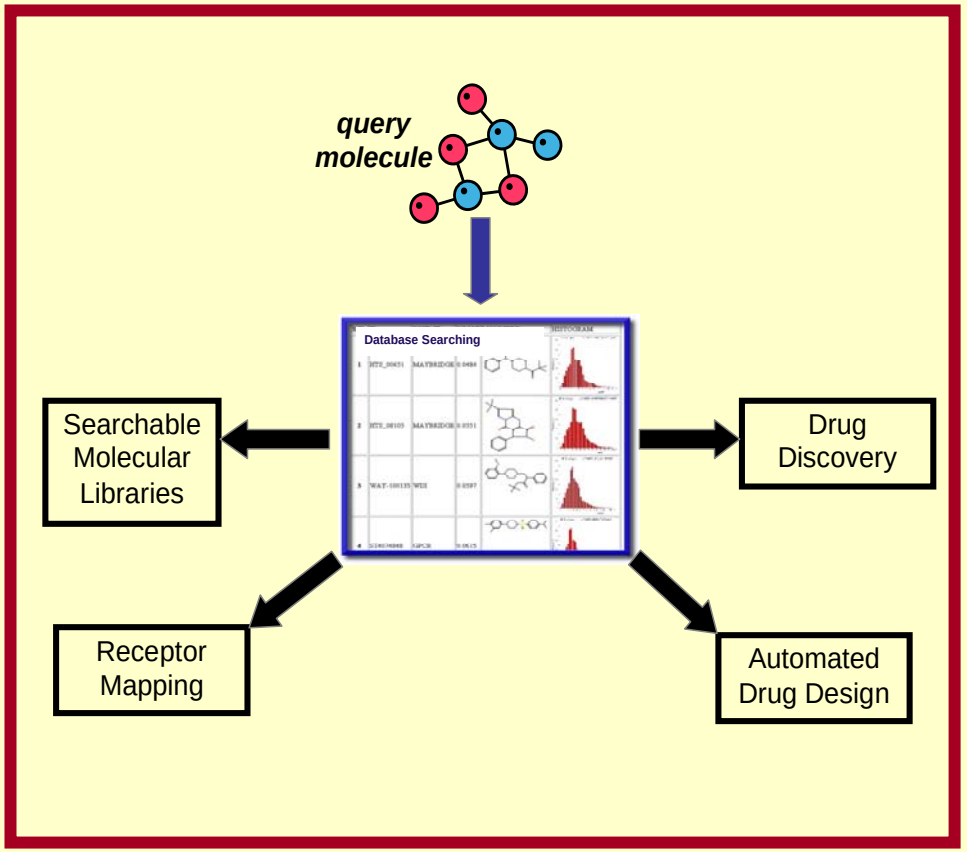


Chemical Structure	CAS Number	1D score	2D score	Toxicity Data (mg/kg-day)	Endpoint(s)

Shape Signatures

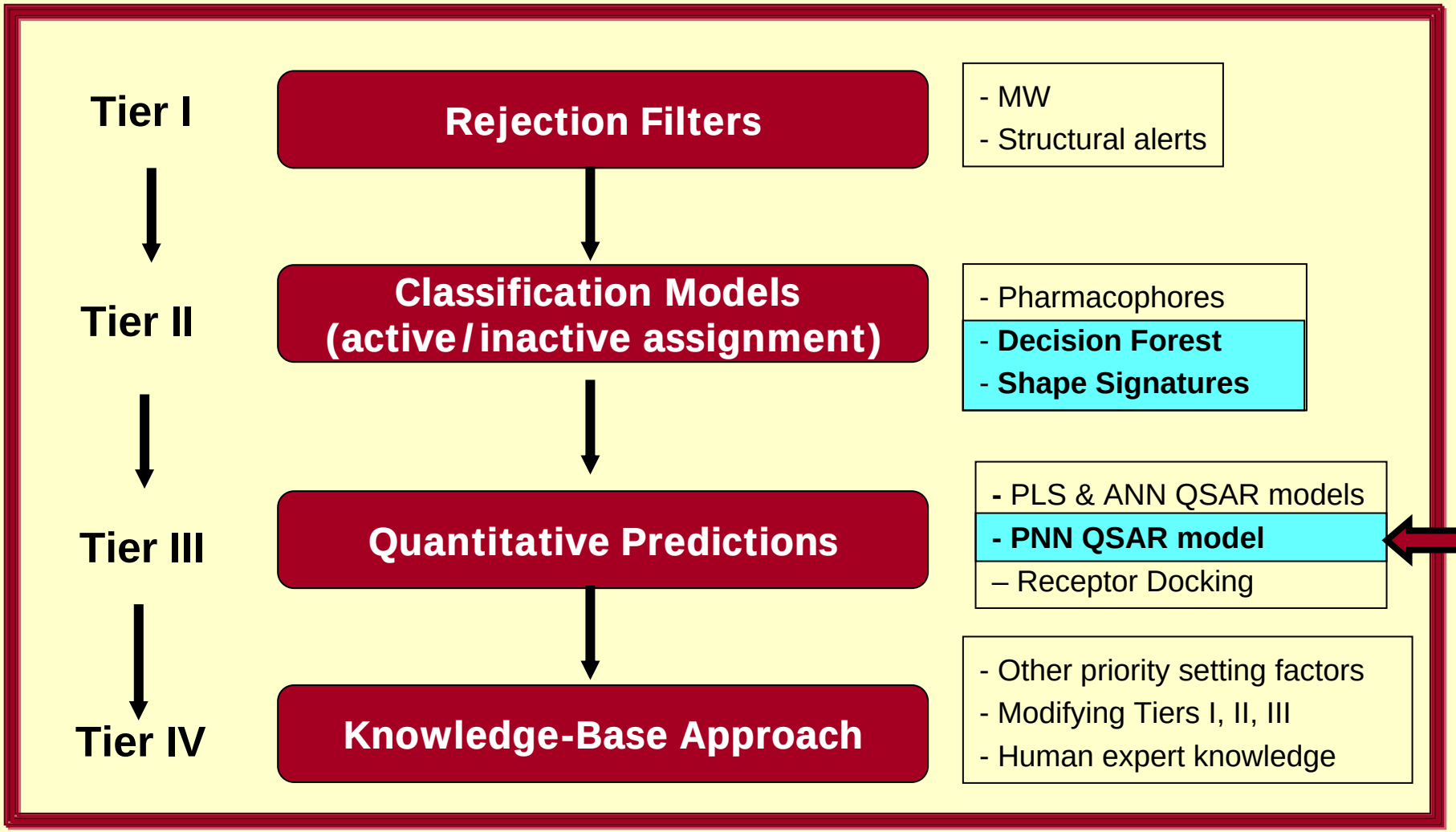
- Key Features -

- Fast
  screens large databases in secs
- Extensible
  works with any kind or number of molecular species
- Portable
  works on any platform
- Versatile
  broad utility, multiple databases



Schematic of Hierarchical Framework

- addresses the need to minimize *false negatives* and *uncertainties* -



Building QSAR Models

target property $\propto$ (molecular descriptors)

$$Y = f(X_i)$$

Types of Molecular Descriptors

Type	Example
Constitutional	Molecular composition (M_w, # of atoms/bonds, # of H-bond donors/acceptors)
Topological	2-D structural formula (Kier-Hall indices, extent of branching)
Geometrical	3-D structure of molecule (molecular volume, solvent accessible surface area, polar and non-polar surface area)
Electrostatic	Charge distribution (atomic partial charges, electronegativities)
Quantum Mechanical	Electronic structure (HOMO-LUMO energies, band gap, dipole moment)

Comparison of Regression Methods

Desirable Features of Methods and Models

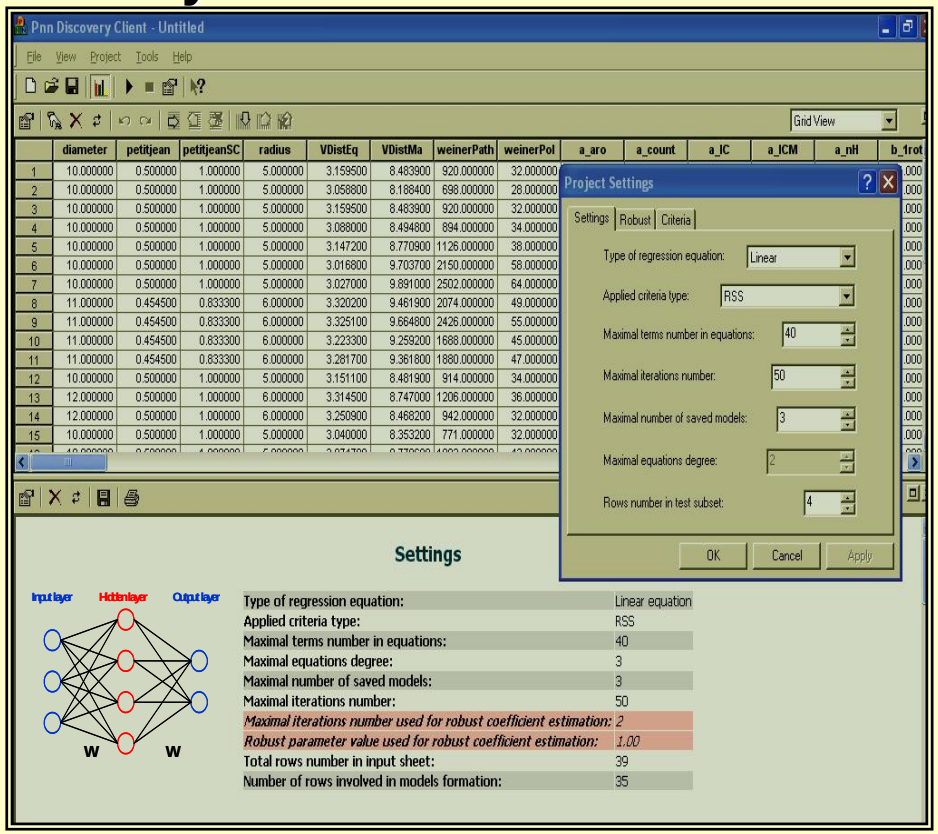
- predictions should be fast
- produces linear or non-linear models (i.e., relationship between obs toxicities and calc'd molecular features may be non-linear)
- models should be physically meaningful, interpretable, and assume parametric form

Method	Speed	Linear Models?	Nonlinear Models?	Regression Equation?	Easy to Interpret?
PLS / MVR	**	Yes	No	Yes	Yes
ANN	*	Yes	Yes	No	Yes
PNN	****	**Yes**	**Yes**	**Yes**	**Yes**

Polynomial Neural Network (PNN)

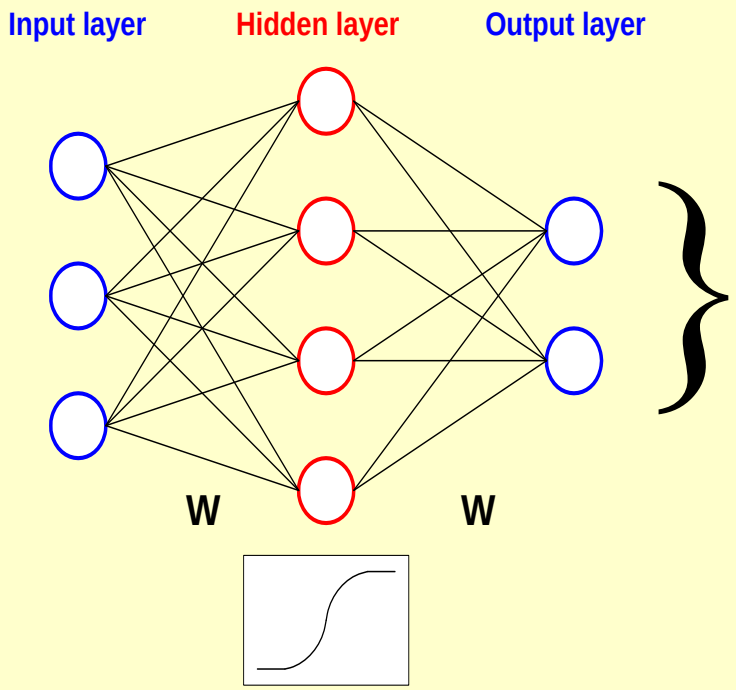
- combines best features of linear multivariate models (parametric form) and ANN models (nonlinearity) -

Polynomial Neural Network



- Produces linear or non-linear QSAR models in parametric form
- User control of model complexity
- Insensitive to irrelevant variables and outliers
- Yields predictive models, even for sparse or noisy data sets
- Trains rapidly, thus amenable to large data sets
- Automatically selects best models
- Customizable to fit user's needs

Polynomial Neural Network (PNN)



1) PNN generates parametric solutions of any desired order ‘n”:

$$\text{Act.} = w_1(SA) + w_2(V) + w_3(\mu) + \ldots$$

$$\text{Act.} = w_1(SA) + w_2(V)^2 + w_3(\mu)^3 + \ldots$$

$$\text{Act.} = w_1(SA)^2 + w_2(V) + w_3(\mu)^2 + \ldots$$

$$\text{Act.} = w_1(SA)^0 + w_2(V) + w_3(\mu)^2 + \ldots$$

$$\text{Act.} = w_1(SA) + w_2(V)^2 + w_3(\mu)^2 + \ldots$$

2) PNN selects best solutions:

$$\text{Act.} = w_1(SA) + w_2(V)^2 + w_3(\mu)^3 + \ldots$$

$$\text{Act.} = w_1(SA) + w_2(V)^2 + w_3(\mu)^2 + \ldots$$

Thank You!

welshwj@umdnj.edu