

Research Theme 1: Information Technologies

Ann Richard

National Center for Computational Toxicology
US Environmental Protection Agency



Agency Problems:

- ✦ Large lists of chemicals to evaluate
- ✦ Many toxicity endpoints to assess
- ✦ Lack of sufficient and relevant data

Need to prioritize and focus limited resources on chemicals and problem areas with potential for greatest health & environmental impact

Inerts

TSCA/PMN

HPV Testing Pgm

Endocrine Disruption
Testing Program

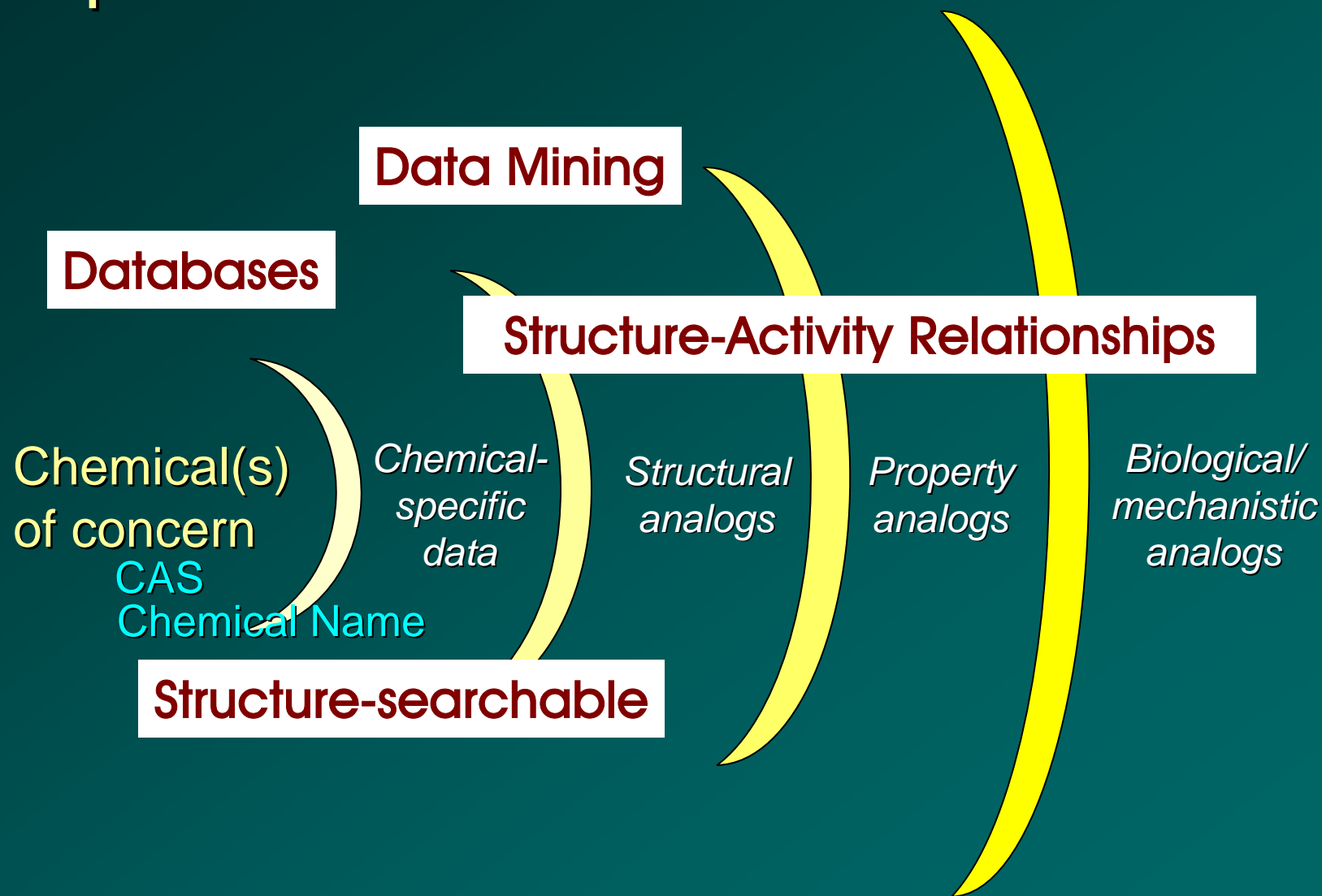
**Water
CCL**

Pesticides

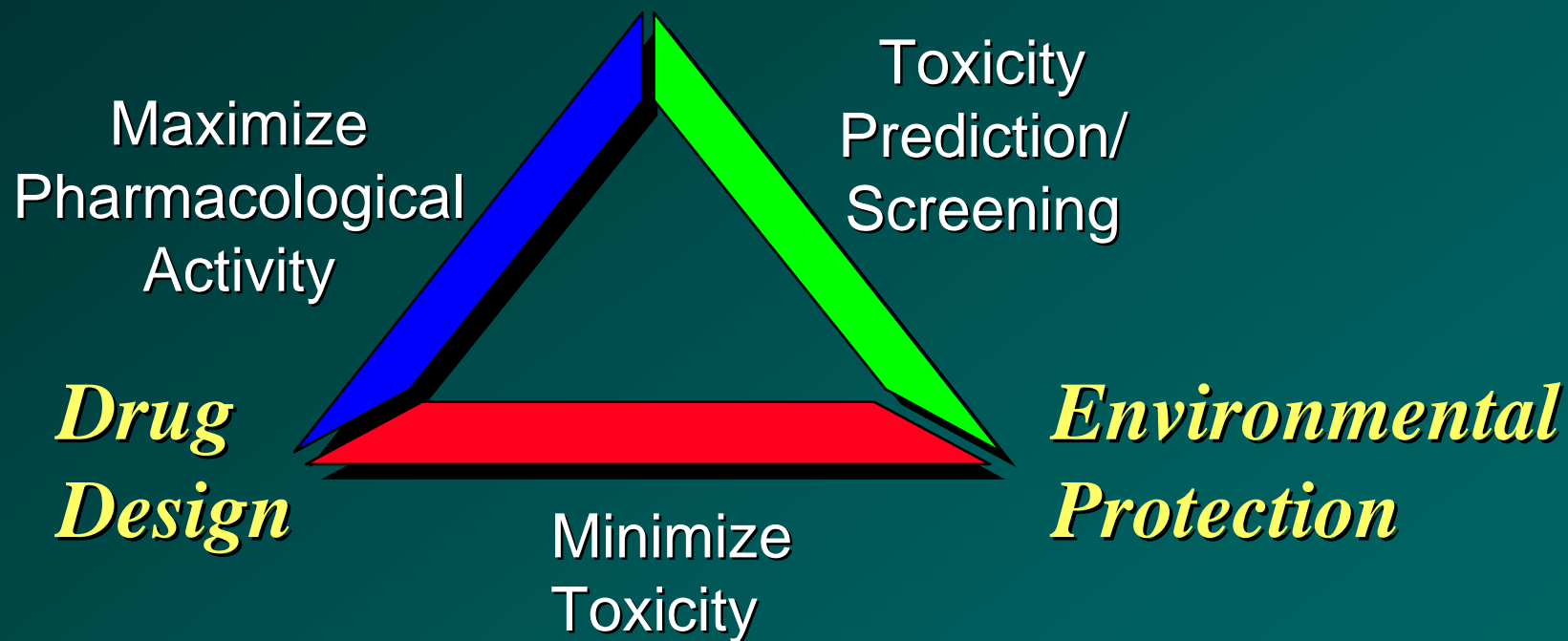
Outline of Presentation:

- ✦ Chemoinformatics & SAR Predictive Challenge
- ✦ DSSTox Project & Data Standards
- ✦ Emerging Chemo-bioinformatic Capabilities

Chemistry-based Data Mining & Exploration:



SAR Application



- Single therapeutic target
- Drug-like chemicals
- Some toxicity anticipated

- Multiple unknown targets
- Diverse structures
- Human and eco endpoints



Chemoinformatics



national
Bioinformatics
institute

What is Chemoinformatics?

Dr. Frank Brown introduced the term “chemoinformatics” in the Annual Reports of Medicinal Chemistry in 1998:

“The use of information technology and management has become a critical part of the ~~drug discovery~~ process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of ~~drug lead identification and organization~~ **toxicity prioritization & screening** **environmental toxicity screening**”

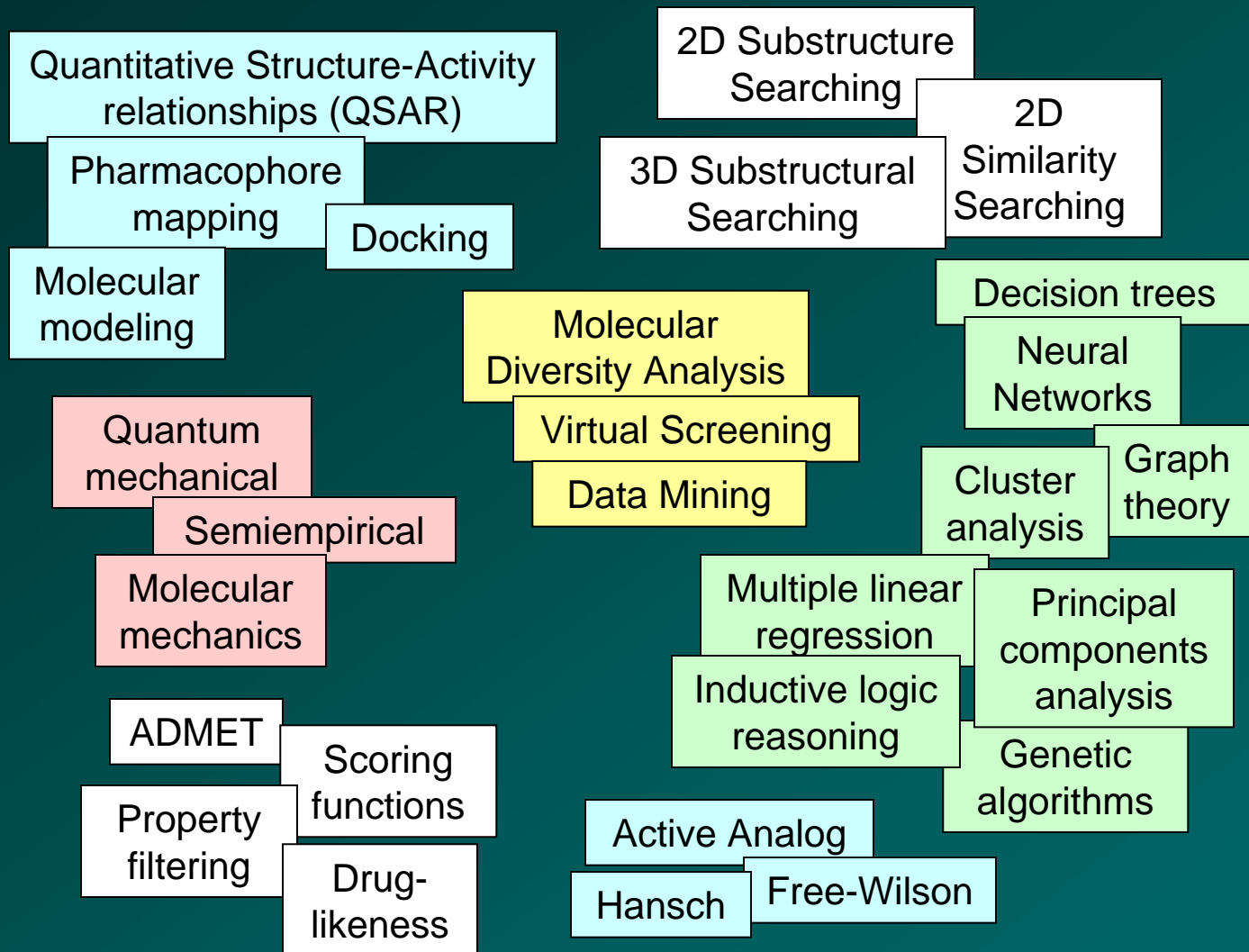
In fact, Chemoinformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information.

<http://www.bioinformatics.com/chemoinfo.htm>

Chemoinformatics

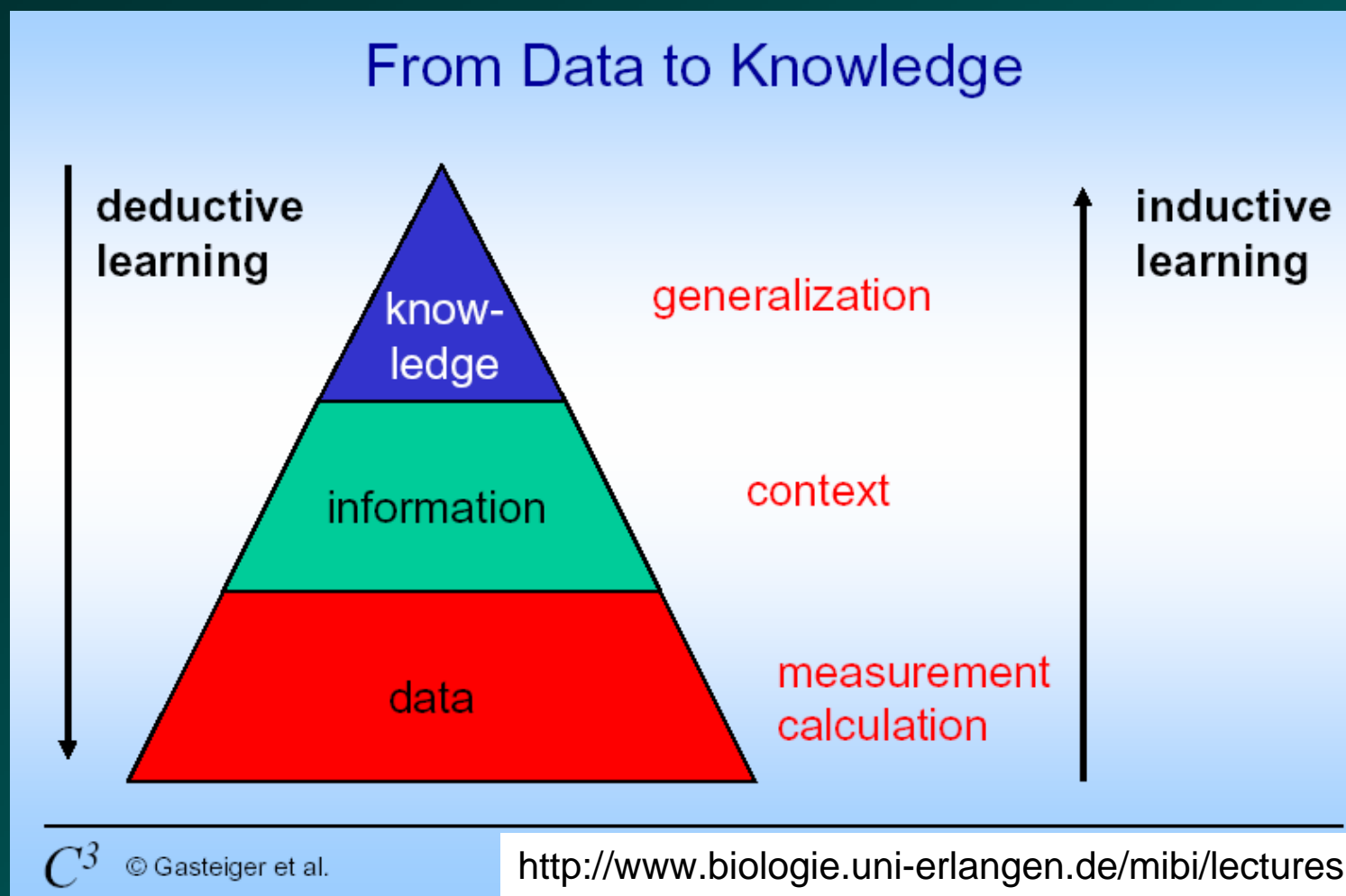


national
BiInformatics
institute



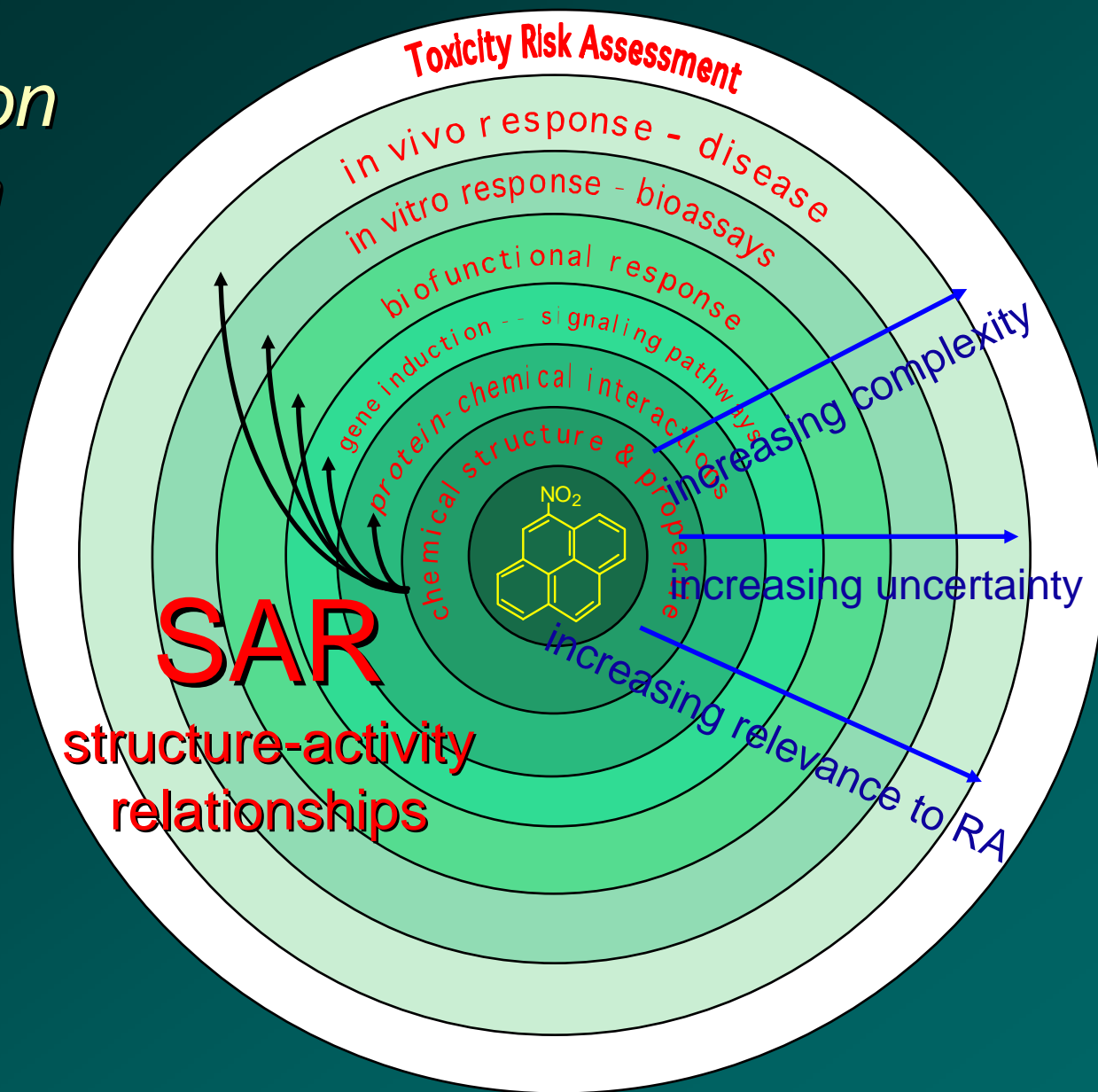
- ✦ Pharmaceutical Sciences
- ✦ Drug Discovery
- ✦ Chemical Design
- ✦ Materials Science
- ✦ Green Chemistry
- ✦ Agricultural Pesticides
- ✦ Food Science
- ✦ Polymers
- ✦ Atmospheric chemistry
- ✦ Environmental Studies
- ✦ Green Chemistry
- ✦ Predictive Toxicology

Chemoinformatics:

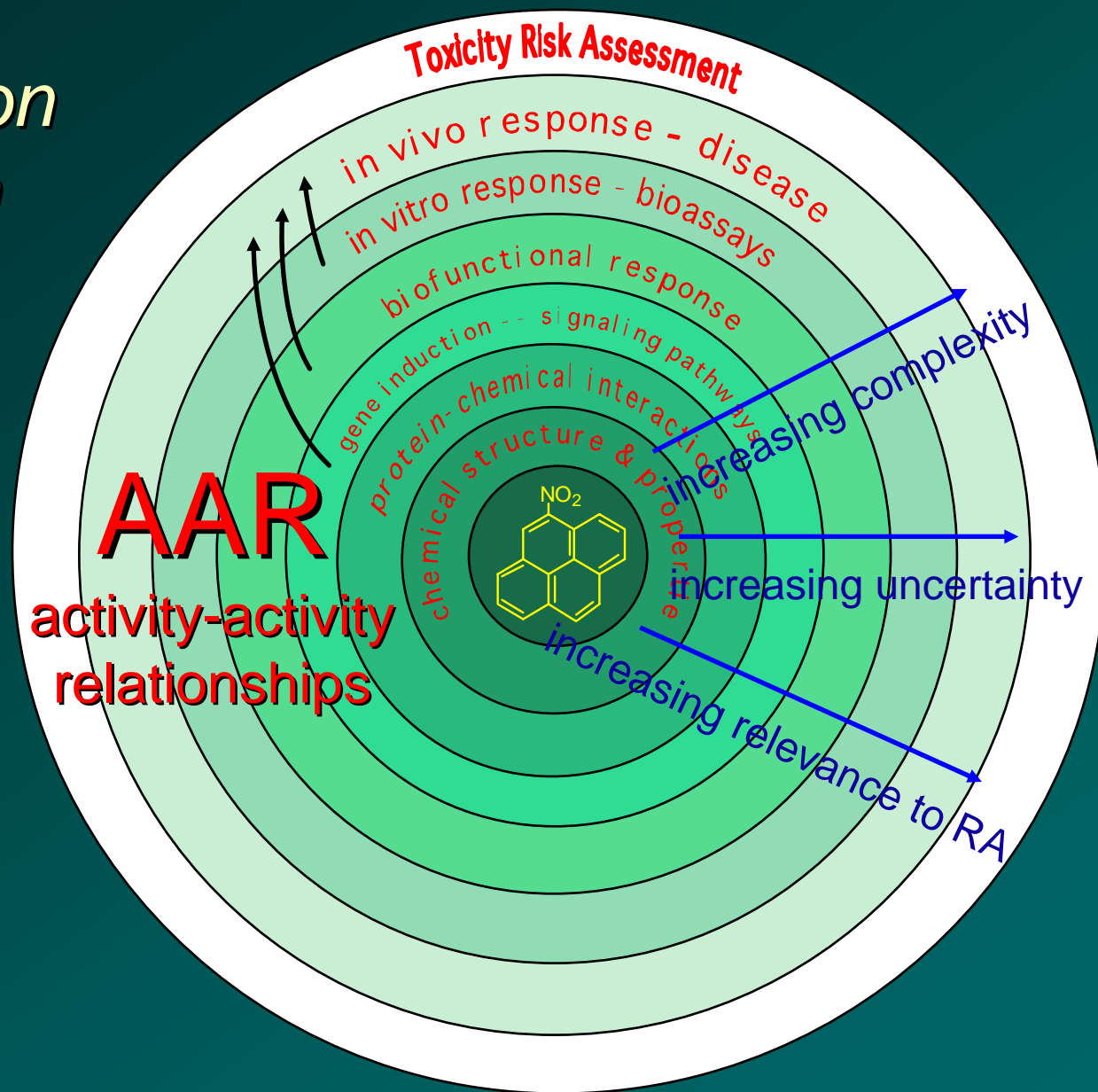


Extracting knowledge from chemical information
– lots of data (structure, activities, genes, etc.)

Toxicity Prediction Problem

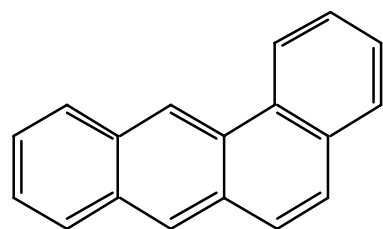


Toxicity Prediction Problem

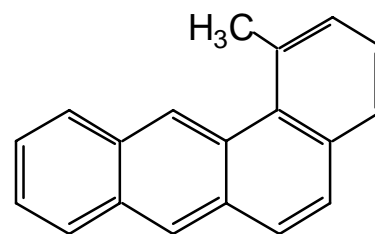


Observation:

Activity = f (Structure)



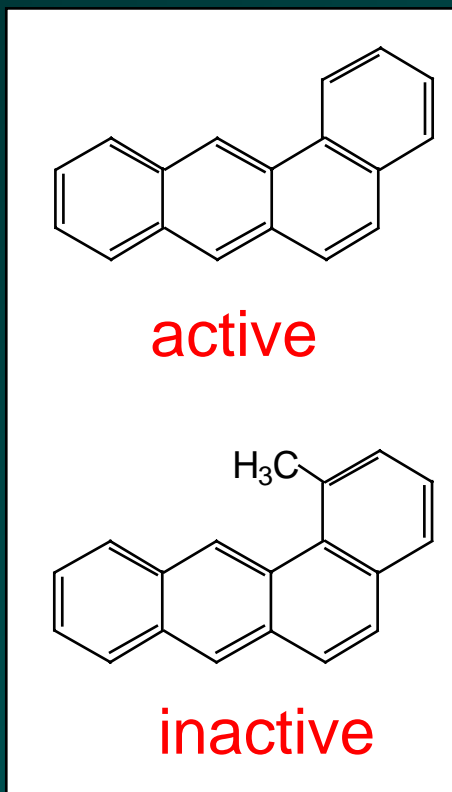
active



inactive

Methyl group leads to loss of carcinogenic activity

SAR: Generalization



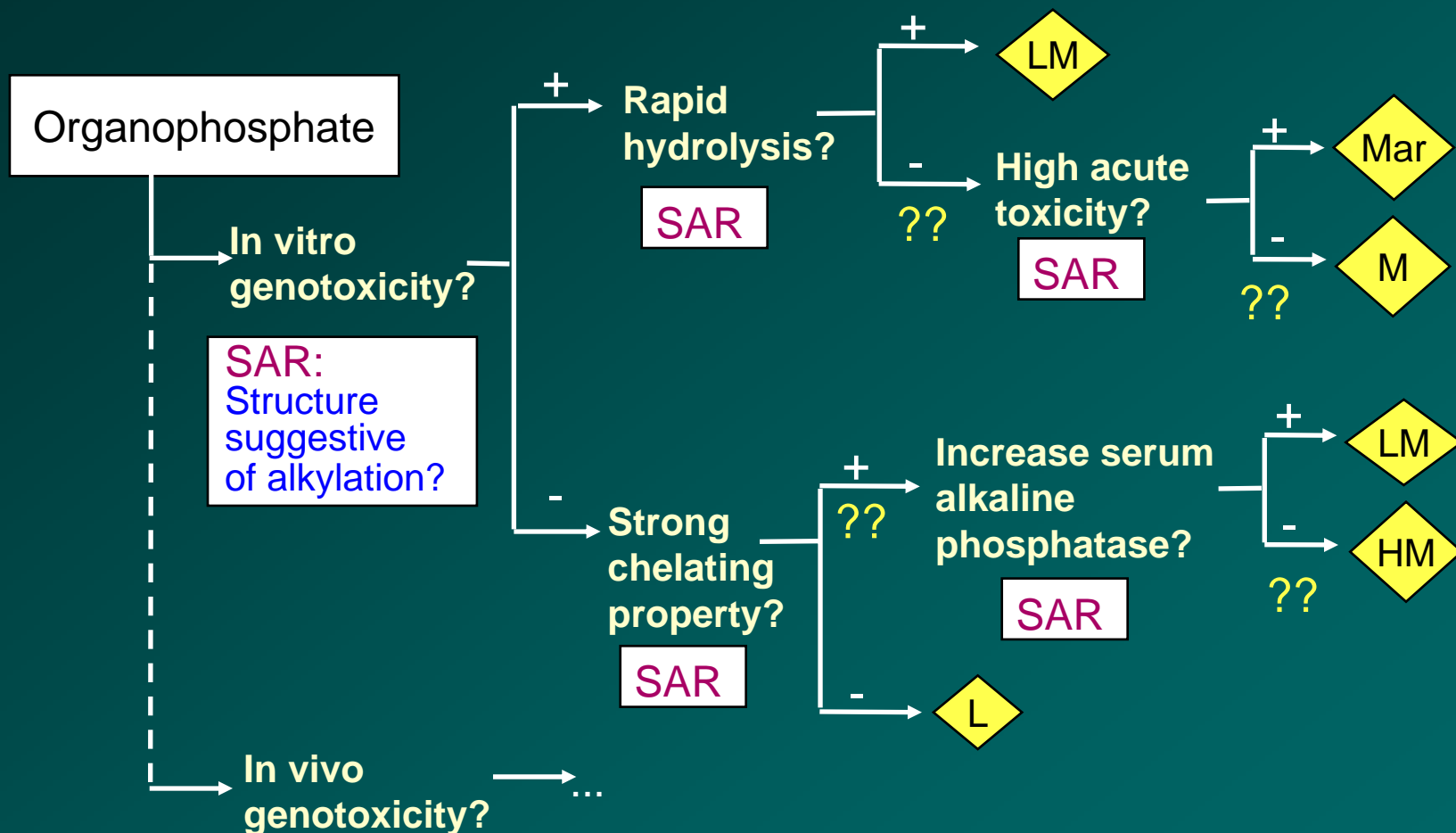
Statistical association
Mechanistic hypothesis



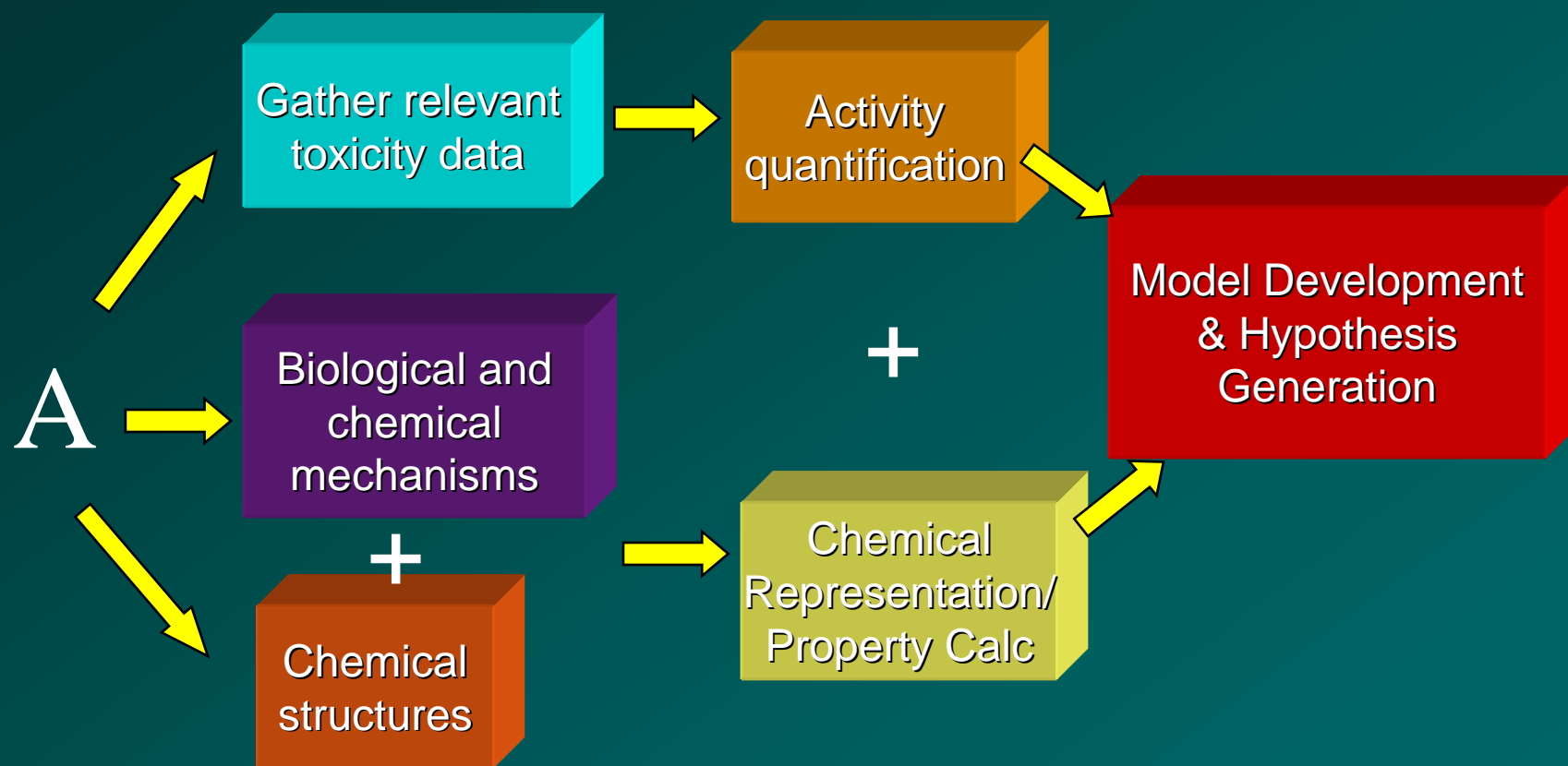
Class → PAHs
activating feature → bay region
modulating feature → steric hindrance

Combining SAR and Biofunctional Information

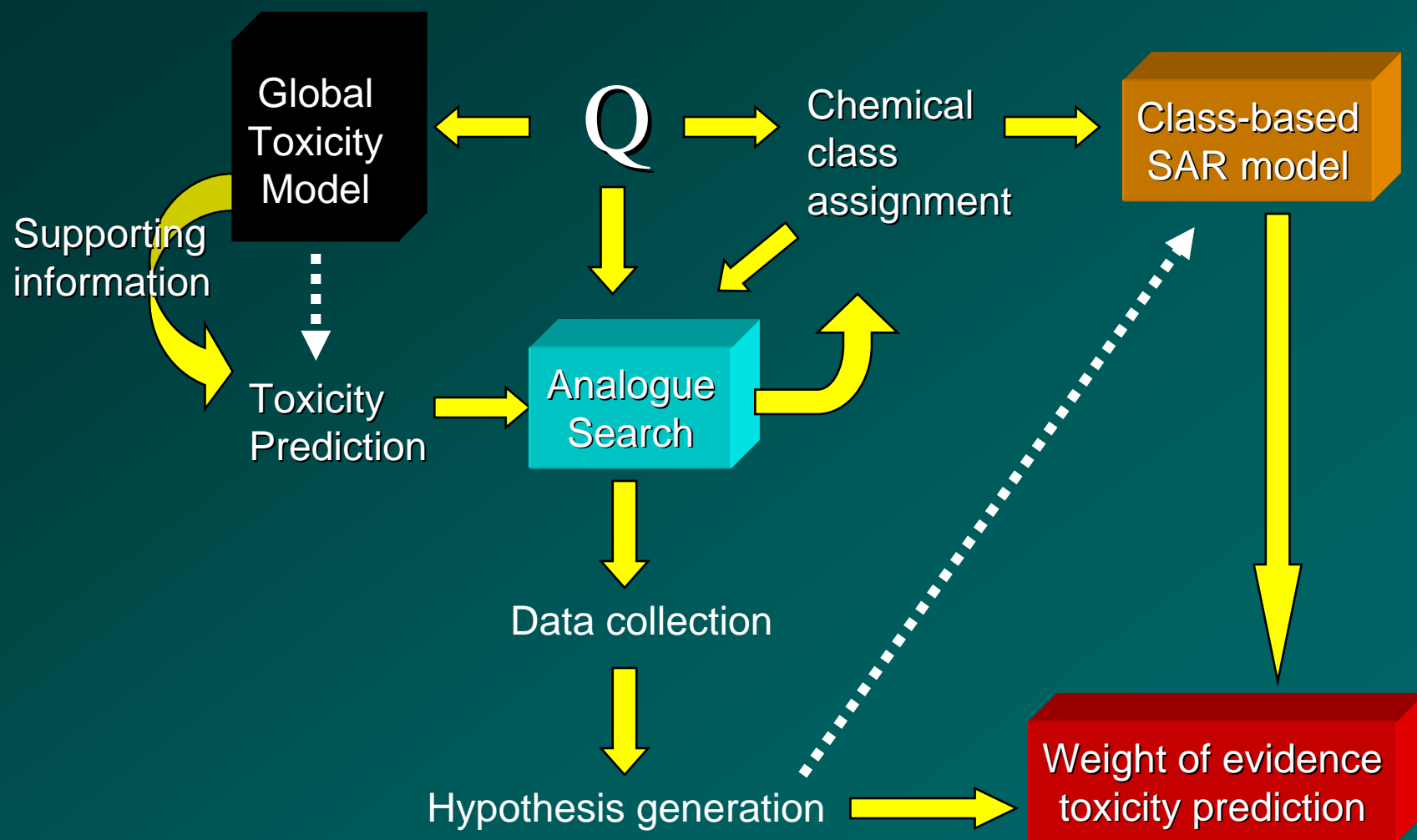
Predicting Carcinogenicity of Organophosphates



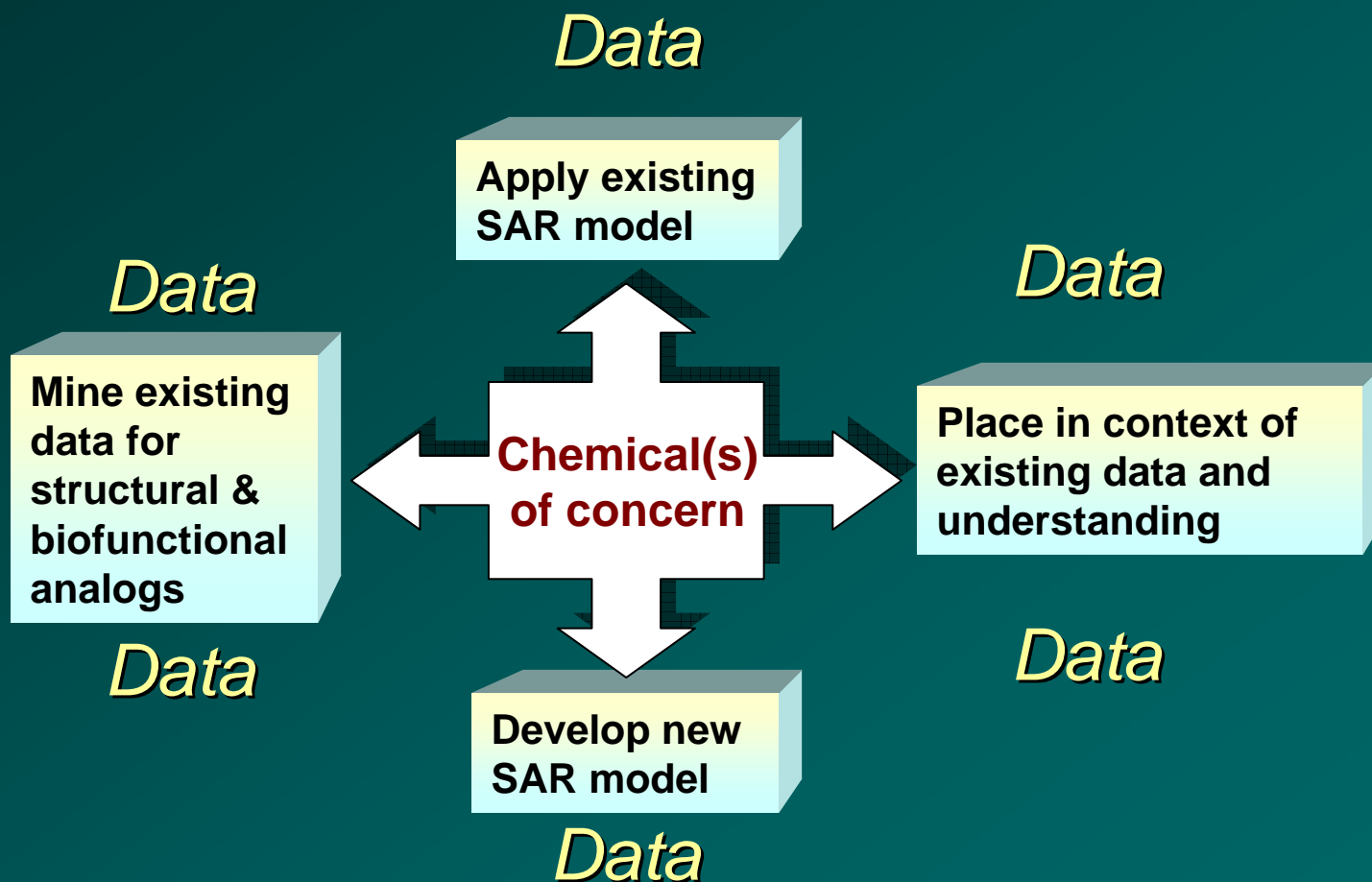
Model Development for Activity A: *Global toxicity prediction*



Toxicity Prediction for Chemical Q



Structure-based Screening & Prioritization:

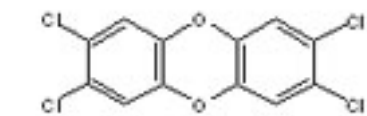


Limitations of Public Toxicity Data for Use in SAR:

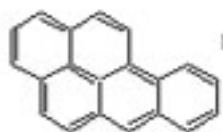
- ◆ Scattered sources
- ◆ Non-standard formats
- ◆ Diverse information content
- ◆ Lack of chemical structure annotation

- ◆ Suitable databases unavailable for many types of tox endpoints

Part II.
DSSTox Project &
Toxicity Database Standards



Chemical Str



Chemical structure-annotation

Distributed
Structure-Searchable
Toxicity
Public
Database
Network

DSSTox
SDF
Files



Import into

Data standards and integration

Prediction
Models



Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network

<http://www.epa.gov/nheerl/dsstox>

[Recent Additions](#) | [Contact Us](#) | [Print Version](#) Search:

[EPA Home](#) > [Research & Development](#) > [Health and Environmental Effects Research](#) > Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network

About DSSTox

Work in Progress

Frequent Questions

Databases

Central Field Definition Table

Apps, Tools & More

DSSTox Community

Site Map

Glossary of Terms

Help

DSSTox

The Distributed Structure-Searchable Toxicity (DSSTox) Database Network is a project of [EPA's Computational Toxicology Program](#), helping to build a public data foundation for improved structure-activity and predictive toxicology capabilities. The DSSTox website provides a public forum for publishing downloadable, standardized toxicity data files that include chemical structures. [More](#)

Recent Additions: 1Mar05

***New Database Additions:

- FDA Maximum (Recommended) Daily Dose Database ([FDAMDD](#)) of 1217 pharmaceuticals - 1Mar05

***Expanded and modified versions:

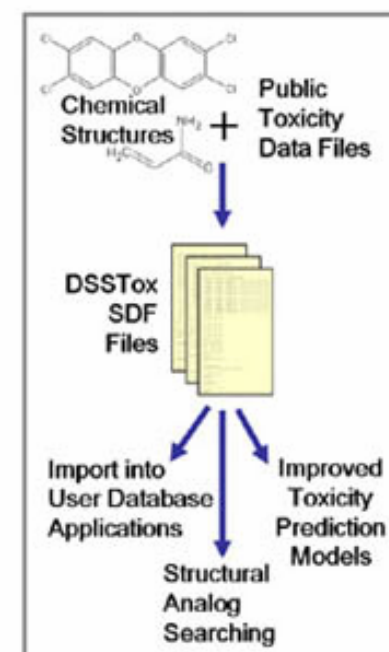
- Consolidated, updated Carcinogenic Potency Database - All Species ([CPDBAS](#)), 1451 compounds: 91 new records added to v2a
- CAS registry numbers added to [EPAFHM](#) and [DBPCAN](#)

***New Standard Fields added to all DSSTox files:

- [INChI](#) (IUPAC/NIST Chemical Identifier) unique structure-text codes
- [IUPAC](#) systematic chemical names (generated by ACD/Name)
- [Standard Toxicity Fields](#): StudyType, Species, Endpoint fields

***New Features of Site:

- [FTP Download Instructions](#) for easy access to archived and new DSSTox data files
- New information pages: [INChI](#), [DSSTox Standard Toxicity Fields](#)
- Links to [External Public Databases](#) adopting DSSTox standards: [ISSCAN](#) *new*



- [DSSTox Graphic Flowchart](#)
- [DSSTox Project Goals](#)
- [DSSTox Publications](#)

DSSTox Databases:

[CPDBAS v2a 1451 1Mar05](#)
[DBPCAN v2a 209 1Mar05](#)
[EPAFHM v2a 617 1Mar05](#)
[FDAMDD v1a 1217 1Mar05**](#)
[NCTRER v2a 232 1Mar05](#)

** new addition



Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network

[Recent Additions](#) | [Contact Us](#) | [Print Version](#) Search: [GO](#)

[EPA Home](#) > [Research & Development](#) > [Health and Environmental Effects Research](#) > [DSSTox](#) > DSSTox Databases

- About DSSTox
- Work in Progress
- Frequent Questions
- Databases
- Central Field Definition Table
- Apps, Tools & More
- DSSTox Community
- Site Map
- Glossary of Terms
- Help

Databases

The following alphabetical list constitutes the most current DSSTox database offerings. Each of the titles below is linked to its main informational/download page.

- See [Recent Additions](#) for a listing of the most recently published DSSTox databases.
- See [Work in Progress](#) for further information on databases currently under development. Upon publication, new DSSTox databases will be added to the listing below.
- See [External Public Databases](#) for links to databases off-site that have adopted DSSTox standards.

- [CPDBAS: Carcinogenic Potency Database Summary Tables - All Species](#)

Tumor target site incidence and TD50 potencies for 1451 chemical substances with data for all species consolidated into single file: rats and mouse, hamsters, dogs, and/or non-human primates; data reviewed and compiled from literature and NTP studies.

(SDF last updated 1Mar05)

- [DBPCAN: Water Disinfection By-Products Database with Carcinogenicity Estimates](#)

Carcinogenicity estimates (high, moderate, low concern) by EPA experts using a mechanism-based analog SAR approach on a set of 209 water disinfection by-products, mostly small halogenated organics.

(SDF last updated 1Mar05)

- [EPAFHM: EPA Fathead Minnow Aquatic Toxicity Database](#)

Acute toxicities of 617 chemicals tested in common assay, with mode-of-action assessments and confirmatory measures.

(SDF last updated 1Mar05)

- [FDAMDD: FDA's Center for Drug Evaluation & Research - Maximum \(Recommended\) Daily Dose Database](#)

Maximum (recommended) daily dose (MRDD) values for 1217 pharmaceuticals in mg/kg-body weight (bw)/day, converted to mmol, and normalized to dataset; MRDD values extracted from public literature sources.

(SDF last updated 1Mar05)

- [NCTRER: FDA's National Center for Toxicological Research - Estrogen Receptor Binding Database](#)

Estrogen receptor relative binding affinities tested in a common in vitro assay for 232 chemicals, listed with chemical class-based structure activity features.

(SDF last updated 1Mar05)

Source SDF Download Page

DSSTox Source SDF Download

NCTRER: National Center for Toxicological Research Estrogen Receptor Binding Database

Description:
 Ligand-based and *in vitro* bioassays are the primary components in assessing the estrogenic activity of chemicals. The National Center for Toxicological Research (NCTR) is a part of the Environmental Protection Agency (EPA). The NCTR is the lead agency for the Environmental Protection Agency and Toxicology Research (EPA/TOX) research on development of predictive toxicology approaches and Tier 1 screening methods. Additionally, the NCTR is the lead agency for the development of predictive toxicology approaches and Tier 1 screening methods. Additionally, the NCTR is the lead agency for the development of predictive toxicology approaches and Tier 1 screening methods.

File List:

File Name	Description	File Size	Download
dsstox.sdf	DSSTox Source SDF	104	[Download]
dsstox.field	DSSTox Field Definition File	204	[Download]
dsstox.log	DSSTox Log File	204	[Download]

DSSTox SDF File

DSSTox SDF File

Chemical structure data in SDF format, showing SMILES strings and associated identifiers.

Structures Data File

Structures Data File

Grid of chemical structures with associated data columns.

No Structures (.xls) Data File

Chemical ID	SMILES	Chemical Name	Chemical Class	Chemical Type	Chemical Status
000001	CC1=CC=CC=C1	Benzene	Aromatic	Chemical	Active
000002	CC1=CC=C(C=C1)O	Phenol	Aromatic	Chemical	Active
000003	CC1=CC=C(C=C1)C	Toluene	Aromatic	Chemical	Active
000004	CC1=CC=C(C=C1)C(=O)O	Benzoic acid	Aromatic	Chemical	Active
000005	CC1=CC=C(C=C1)C(=O)OC	Methyl benzoate	Aromatic	Chemical	Active
000006	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000007	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000008	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000009	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000010	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000011	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000012	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000013	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000014	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000015	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000016	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000017	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000018	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000019	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000020	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active

Field Definition File

DSSTox Field Definition File:
 Carcinogenic Potency Database Summary Tables (CPDBS1, CPDBS2, CPDBS3, CPDBS4)

Description:
 This file defines the fields used in the DSSTox database. It includes information on the source of the data, the type of data, and the units used.

Field Name	Field Type	Field Description	Field Units
Chemical ID	Text	Chemical identifier	
SMILES	Text	Chemical structure	
Chemical Name	Text	Chemical name	
Chemical Class	Text	Chemical class	
Chemical Type	Text	Chemical type	
Chemical Status	Text	Chemical status	

Log File

DSSTox Log File:
 Carcinogenic Potency Database Summary Tables (CPDBS1, CPDBS2, CPDBS3, CPDBS4)

Description:
 This file contains the log of the DSSTox database. It includes information on the source of the data, the type of data, and the units used.

Chemical ID	SMILES	Chemical Name	Chemical Class	Chemical Type	Chemical Status
000001	CC1=CC=CC=C1	Benzene	Aromatic	Chemical	Active
000002	CC1=CC=C(C=C1)O	Phenol	Aromatic	Chemical	Active
000003	CC1=CC=C(C=C1)C	Toluene	Aromatic	Chemical	Active
000004	CC1=CC=C(C=C1)C(=O)O	Benzoic acid	Aromatic	Chemical	Active
000005	CC1=CC=C(C=C1)C(=O)OC	Methyl benzoate	Aromatic	Chemical	Active
000006	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000007	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000008	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000009	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000010	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000011	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000012	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000013	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000014	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000015	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000016	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000017	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000018	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000019	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active
000020	CC1=CC=C(C=C1)C(=O)OC(=O)C	Phthalic anhydride	Aromatic	Chemical	Active

DSSTox SDF files:

csChmFindW05030111462D

```
14 16 0 0 0 0 0 0 0 0999 V2000
0.1283 2.1977 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
0.0000 0.7780 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
1.0347 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
2.3261 0.5213 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
2.4544 1.9411 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
1.4197 2.7191 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
3.6254 0.0000 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
4.5318 1.0347 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
3.8821 2.1977 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
5.9516 1.0347 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
6.7295 2.1977 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
5.9516 3.4891 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
4.5318 3.4891 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
8.0209 2.1977 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
1 2 1 0 0 0 0
1 6 2 0 0 0 0
2 3 2 0 0 0 0
3 4 1 0 0 0 0
4 5 1 0 0 0 0
4 7 2 0 0 0 0
5 6 1 0 0 0 0
5 9 1 0 0 0 0
7 8 1 0 0 0 0
8 9 2 0 0 0 0
8 10 1 0 0 0 0
9 13 1 0 0 0 0
10 11 2 0 0 0 0
11 12 1 0 0 0 0
11 14 1 0 0 0 0
12 13 2 0 0 0 0
M END
> <Last Updated> (1)
5/3/01

> <Source> (1)
http://potency.berkeley.edu/cpdb.html

> <Chemical> (1)
A-alpha-C

> <CAS> (1)
26148-68-5

> <Tested Form> (1)
neutral
```

SAR Model Development “Training Sets”

- ◆ improved predictive tox models
- ◆ more comparable models
- ◆ lowered barriers to use

Chemical Relational Database: *sub-structure, text, property searching*

- ◆ analog searches
- ◆ search across diverse toxicity endpoints
- ◆ search across chemical & toxicity fields

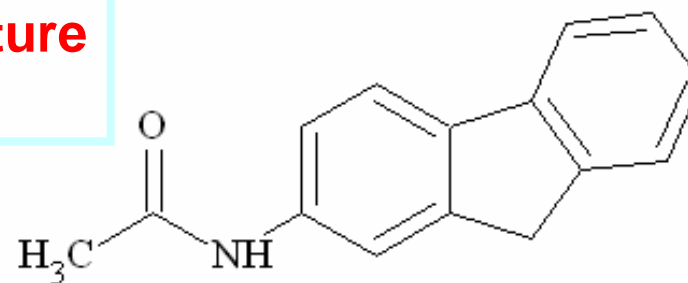
DSSTox Standard Chemical Fields:

ACD
ChemFolder
Application view
after SDF import

ChemName_IUPAC
SMILES_Parent
InChI

StudyType
Species
Endpoint

**Structure
Field**



```
Formula: C15H13NO
FW: 223.2698
DSSTox_ID: 18
DSSTox_FileName: CPDBAS_v2a_1451_4Jan05
MolWeight: 223.2738
StructureShown: tested form
ChemName: 2-Acetylaminofluorene
CAS: 53-96-3
SubstanceType: defined organic
TestedForm: parent
ChemCount: 1
ChemName_IUPAC: N-9H-fluoren-2-ylacetamide
SMILES: C12C3=C(C=CC=C3)CC1=CC(=CC=2)NC(C)=O
SMILES_Parent: C12C3=C(C=CC=C3)CC1=CC(=CC=2)NC(C)=O
INChI: INChI=1.12Beta/C15H13NO/c1-10(17)16-13-6-7-15-12(9-13)8-11-4
StudyType: carcinogenicity
Species: rat, mouse, hamster, rhesus
Endpoint: TD50, Tumor Target Sites
```

**Standard
Chemical
Fields**

**Standard
Toxicity
Fields**

Integrating Diverse Databases from a Chemical Structure Perspective:

CPDB

DBPCAN

EPAFHM

NCTRER

Standard Chemical Fields

Standard Tox Fields: StudyType, Species, Endpoint

SAL CPDB

TD50 Rat

TD50Mouse

Target Sites Rat
Male

Target Sites
Rat Female

Target Sites
Mouse Male

....
Other Species

ChemClass DBP

Concern Level

Rationale

Rational Source

Analog
ChemName

AnalogCAS

AnalogSMILES

ChemClass FHM

MOA

MOACONF

CLOGP

LC50

LC50NOTE

LC50RATIO

MIXMOA

TOXINDEX

FATS

BEHAVIOR

NCTRlogRBA

ER RBA

ChemClass ERB

Activity Group
ERB

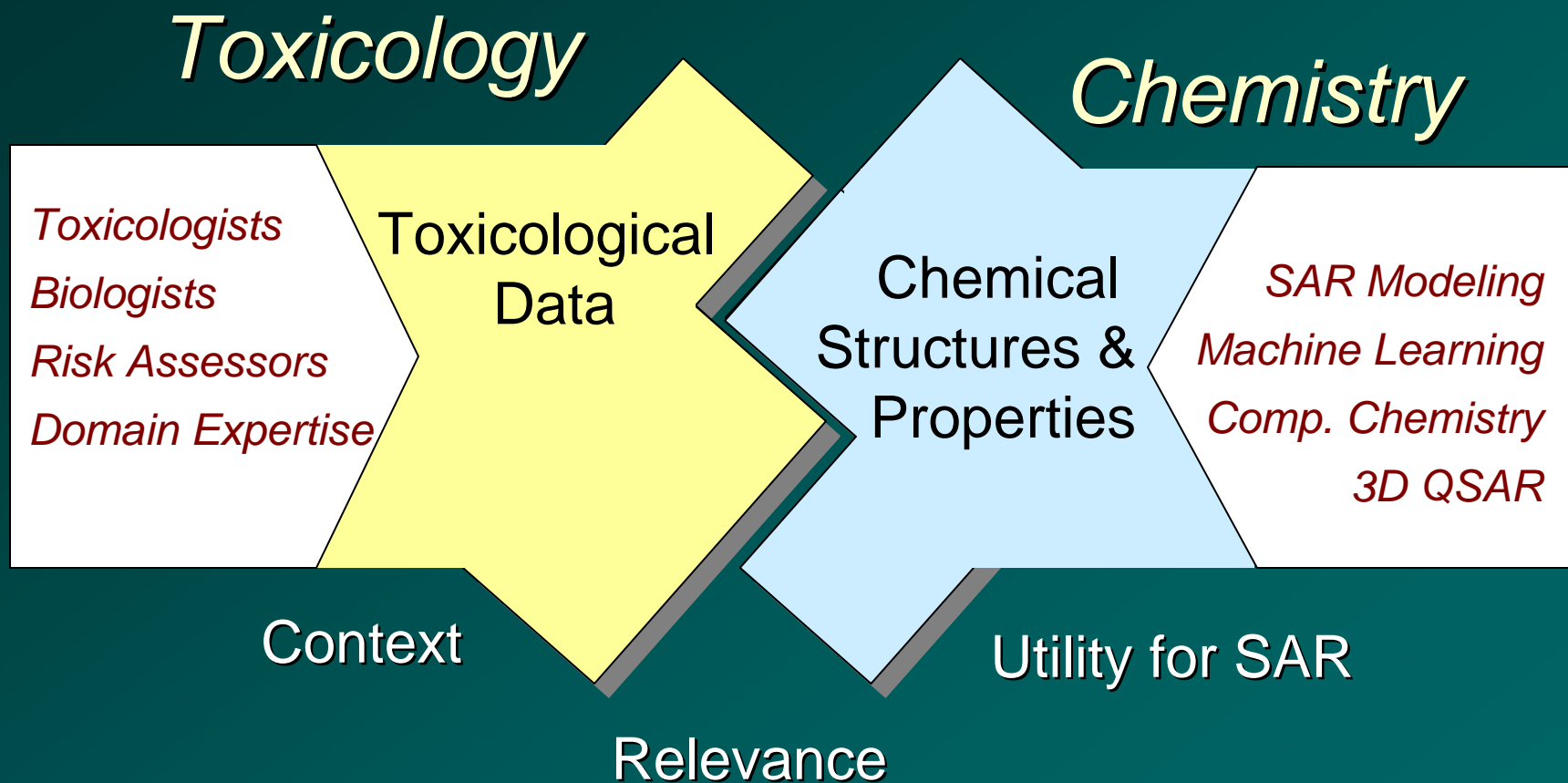
Rationale
ChemClass ERB

MeanChem
Class ERB RBA

LogP

F1, F2, ...F6

DSSTox Database Design:



Where are we heading:

- ✦ Expanded database offerings
- ✦ Automation: DSSTox Master List
- ✦ On-line structure browsing
- ✦ Integration with other public databases
- ✦ Coordination with public data standards
 - *Chemical standards (INChI, etc)*
 - *ToxML*
 - *Miami-Tox*
- ✦ EPA-wide structure browser
- ✦ Linkages to toxicogenomics efforts

Up Next ...

- **NTPIMT**: NIEHS National Toxicology Program Immunotox Database
Based on 1992 publication reporting a battery of immunotox results for 50 chemicals.
- **IRISSI**: EPA's Integrated Risk Information System (IRIS) – Structure Index
Structure index file for EPA IRIS database – initial phase without toxicity data, just structural information.
- **NCTRAR**: FDA's National Center for Toxicological Research - Androgen Receptor Binding Database
Androgen receptor relative binding affinities tested in a common in vitro assay for 202 chemicals, provided with chemical class-based structure activity features.
- **NTPGTZ**: NIEHS National Toxicology Program Gene-Tox Database (E. Zeiger)
Battery of genetic toxicity test results for over 1900 chemicals from historical NTP studies.
- **UNLVSS**: UniLever's Skin Sensitization Database
Skin sensitization results for over 200 chemicals from Unilever studies.

DSSTox Master Chemical Structures File:

1. Consolidate

5. Master Table

6. Create new DSSTox database

CPDBAS →

DBPCAN →

EPAFHM →

FDAMDD →

NCTRER →

IRISSI →

NCTRAR →

UNLVSS →

NTPGTZ →

NTPIMT →

Structure	Formula	Pub	PubID	PubName	PubDate	PubType	PubStatus	PubAccession
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	1	275,239	pubmed	488,72.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	2	242,278	pubmed	527,80.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	3	254,241	pubmed	495,60.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	4	266,269	pubmed	495,72.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	5	288,272	pubmed	1058,28.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	6	319,031	pubmed	3628,02.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	7	380,929	pubmed	117,40.2	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	8	522,424	pubmed	9221,28.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	9	288,264	pubmed	528,40.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	10	224,288	pubmed	487,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	11	302,291	pubmed	528,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	12	302,291	pubmed	528,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	13	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	14	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	15	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	16	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	17	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	18	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	19	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	20	288,264	pubmed	507,20.0	defined organic	pubmed	

QA

2. Identify chemical "replicates"

3. Reconcile replicate information

4. Cull replicate information

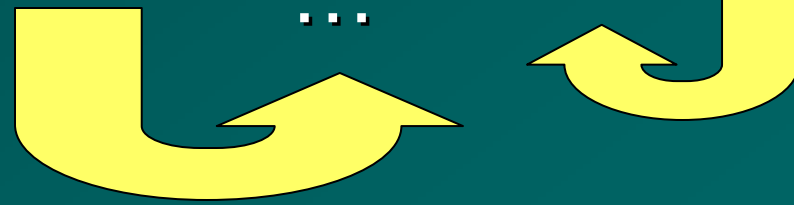
Structure	Formula	Pub	PubID	PubName	PubDate	PubType	PubStatus	PubAccession
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	1	275,239	pubmed	488,72.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	2	242,278	pubmed	527,80.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	3	254,241	pubmed	495,60.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	4	266,269	pubmed	495,72.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	5	288,272	pubmed	1058,28.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	6	319,031	pubmed	3628,02.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	7	380,929	pubmed	117,40.2	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	8	522,424	pubmed	9221,28.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	9	288,264	pubmed	528,40.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	10	224,288	pubmed	487,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	11	302,291	pubmed	528,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	12	302,291	pubmed	528,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	13	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	14	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	15	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	16	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	17	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	18	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	19	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	20	288,264	pubmed	507,20.0	defined organic	pubmed	

CAS cross-referencing

Populate standard chemical fields

Structure	Formula	Pub	PubID	PubName	PubDate	PubType	PubStatus	PubAccession
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	10	224,288	pubmed	487,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	11	302,291	pubmed	528,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	12	302,291	pubmed	528,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	13	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	14	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	15	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	16	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	17	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	18	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	19	288,264	pubmed	507,20.0	defined organic	pubmed	
<chem>C1=CC=C(C=C1)O</chem>	C ₆ H ₆ O	20	288,264	pubmed	507,20.0	defined organic	pubmed	

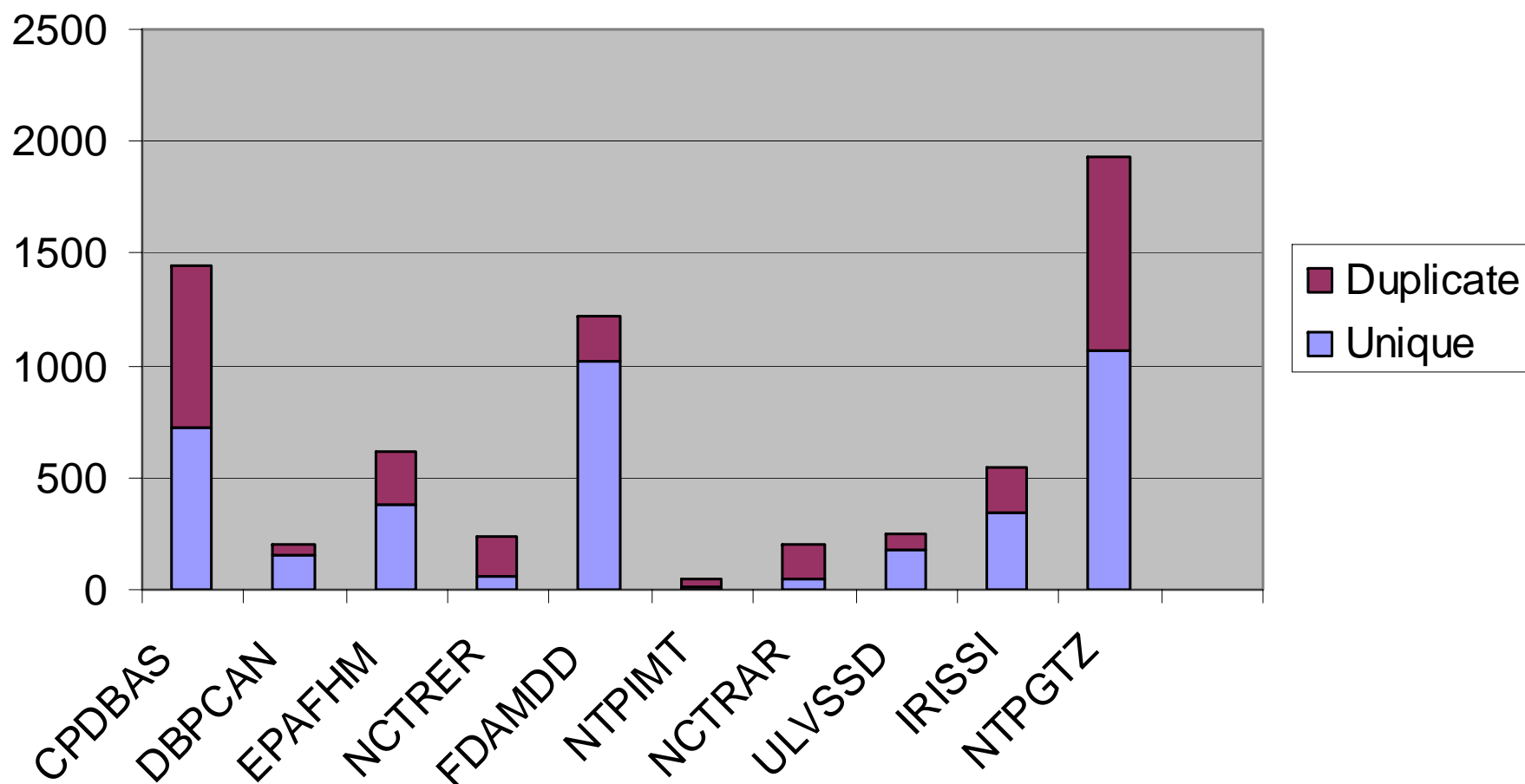
7. Add new structure records to Master



DSSTox Master File

Total Records: 6701

Total Unique Records: 3967 (no replicates)



DSSTox Coordination/Collaborations:

- ✦ Migrate tox data into electronic form
- ✦ Convert raw data to structured data
- ✦ Promote toxicity data standards

IBM's
WebFountain

ILSI Develop. Tox.
Database Initiative

LeadScope's LIST and ToxML
Public Tox Data Initiatives

DSSTox

EPA/SRC PBT Profiler
and Analog Identification
Methodology

EPA – IRIS, OPP,
OPPT, OEI, NHEERL,
Green Chemistry

IUPAC/NIST InChI
Project

Berkeley Carcinogenic
Potency Project

ILS/ICVAM
Databases

FDA CDER/CFSAN
Drug evaluation
reviews/ Food additives

NIH / NLM
TOXNET, PubChem

LHASA VITIC
Toxicity Database

NIH / NCI Structure-
Browser

ACD Labs
WebLibrarian

FDA's NCTR –
Databases, ArrayTrack

NIEHS – NTP,
NCT/CEBS

- ✦ Migrate more tox data to DSSTox format
- ✦ Adopt DSSTox chemical standards
- ✦ Link DSSTox to other public efforts
- ✦ Provide on-line structure-search capability
- ✦ Expand into toxicogenomics

IBM's WebFountain, OmniFind

United States Patent [15] **3,692,776**
Shindo et al. [45] **Sept. 19, 1972**

[54] **PROCESS FOR PREPARING 7-CHLORO-1,3-DIHYDRO-1-METHYL-5-PHENYL-2H-1,4-BENZODIAZEPIN-2-ONE**

[72] Inventors: Minoru Shindo, Tokyo; Kanji Moro, Tokyo; Teizo Shinozaki, Chiba-ken, all of Japan

[73] Assignee: Chugai Seryaku, Kabushiki Kaisha, Tokyo, Japan

[22] Filed: Oct. 27, 1970

[21] Appl. No.: 84,549

Related U.S. Application Data

[62] Division of Ser. No. 841,611, Oct. 27, 1970.

[52] U.S. Cl. 260/239.3 D, 260/562 N

[51] Int. Cl. C07d 53/06

[58] Field of Search 260/239.3 D

[56] **References Cited**

UNITED STATES PATENTS

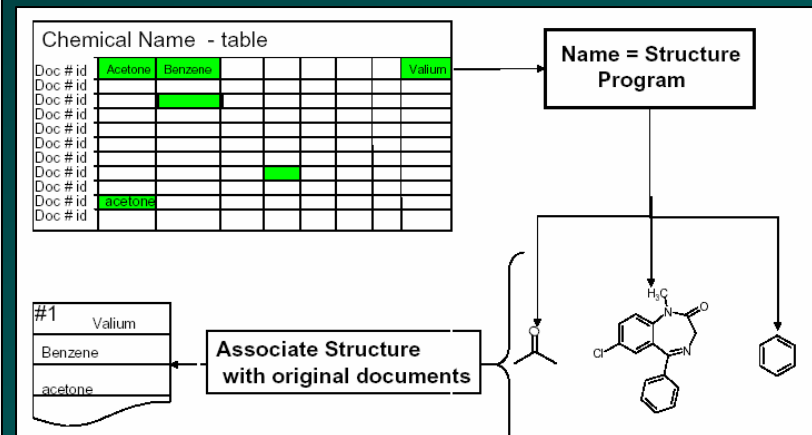
085 2/1968 Reeder et al. 260/239.3 D

Primary Examiner—Henry R. Jiles
Assistant Examiner—Robert T. Bond
Attorney—Otto John Munz

[57] **ABSTRACT**

The known 7-chloro-1,3-dihydro-1-methyl-5-phenyl-2H-1,4-benzodiazepin-2-one which was found to be pharmacologically effective in neuro-psychic disorders is prepared easily and in high yield by an improved process comprising the reaction of a novel intermediate, N-aminoacetyl-5-chloro-N-methylantranilic acid, with phosphorus pentachloride followed by reacting with benzene in the presence of aluminum chloride, the intermediate which is also found to have valuable pharmacological activities being, in turn, prepared by the reaction of 5-chloro-N-methyl-N-phthalimidoacetyl-antranilic acid with hydrazine.

15 Claims, No Drawings



- ◆ US Patent Literature
- ◆ Over 8 million pages processed
- ◆ Searchable by text, content annotators, and chemical structures

- ◆ Recognize chemical names
- ◆ Extract chemical names
- ◆ Convert names to structures
- ◆ Associate chemical structures with original document

- Slide content courtesy of Steve Boyer, IBM

Toxicity data standards:

controlled vocabularies and ontologies

- Leadscope's ToxML public effort
- NIEHS CEBS project
- MIAMI-Tox



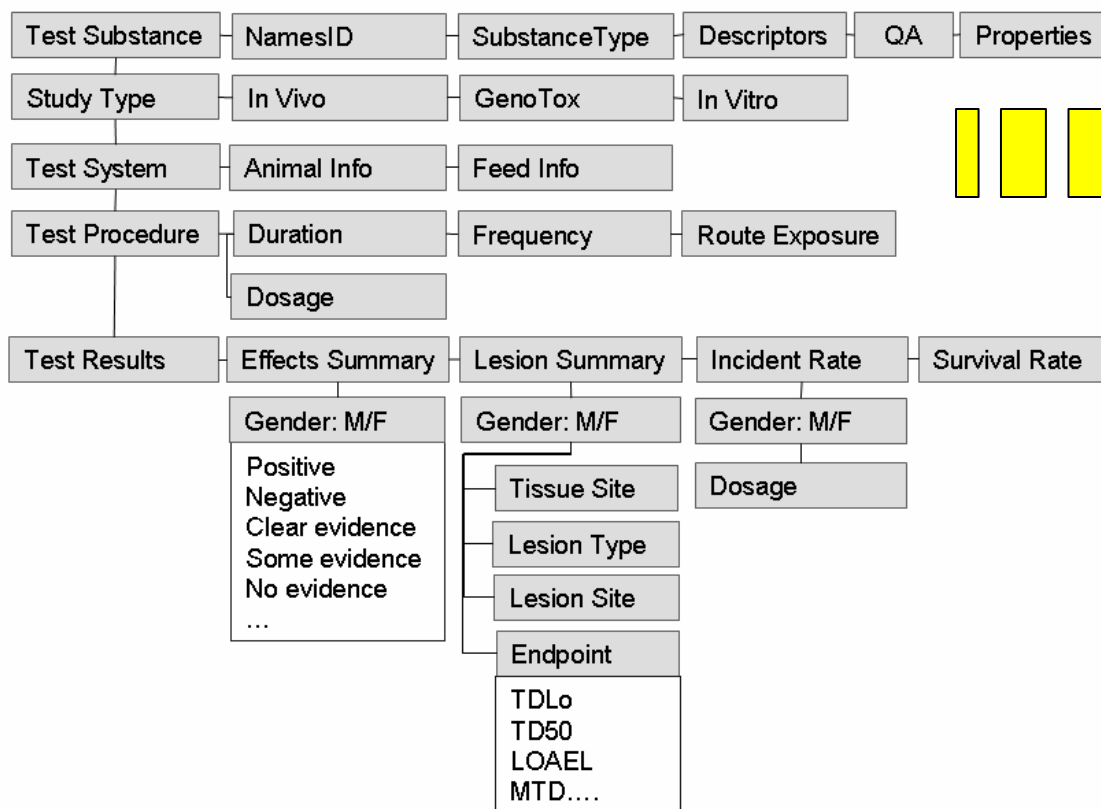
Toxicity databases:

- ILSI Developmental Toxicity DB Project
- Leadscope/FDA – commercial VLC
- LHASA VITIC SAR Toxicity DB Project

Toxicity Experimental Data → Summary Data:

Toxicity Content Model

LIST:ToxML



Intermediate
toxicity
classifications
for SAR



LIST: Leadscope InSilico
Toxicology Consortium

ToxML: Toxicology
controlled vocabulary
for data mining

IUPAC

Trimellitic acid

Current Project

Chemical Nomenclature and Structure Representation Division (VIII)

Number: 2000-025-1-800

Title: IUPAC- International Chemical Identifier

Task Group

Chairman: [A. McNaught](#)

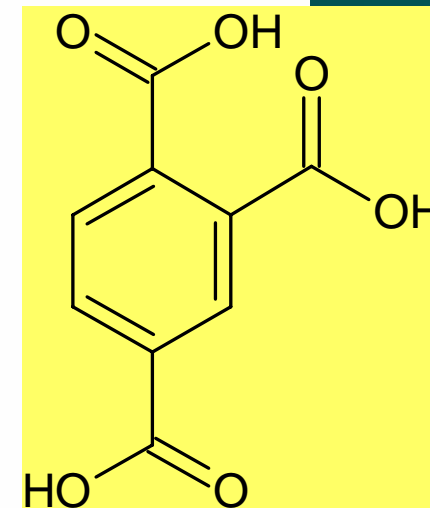
Members: [S. Heller](#) and [S. Stein](#)

Remarks:

- initiated by the [ad hoc Committee on Chemical Identity and Nomenclature Systems](#)
- In July 2004, the Identifier was renamed INChI (formerly IChI) to acknowledge the development work at NIST.
- In November 2004, the Identifier was renamed IUPAC International Chemical Identifier (InChI), to allow trademark, copyright and other issues to be resolved.

Objective:

> 1 July 2002 <
[CAS/IUPAC
Conference on
Chemical Identifiers
and XML for
Chemistry](#)



benzene-1,2,4-
tricarboxylic acid

> INChI=1.12Beta/C9H6O6/c10-7(11)4-1-2-5(8(12)13)6(3-4)9(14)15/h1-3H,(H,10,11)(H,12,13)(H,14,15)



[News &
Notices](#)

[Organizations
& People](#)

[Standing
Committees](#)

[Divisions](#)

[Projects](#)
[..current](#)
[..completed](#)
[..new](#)
[..information](#)

[Reports](#)

[Publications](#)

[Symposia](#)

[AMP](#)

[Links of
Interest](#)

INChI:

- Unique
- Public
- Text XML-based
- Chemically robust
 - ✓ charge state
 - ✓ chiral centers
 - ✓ tautomeric form

```
INChI=1.12Beta/C9H6O6/c10-7(11)4-1-2-5(8(12)13)6(3-4)9(14)15/h1-3H,(H,10,11)(H,12,13)(H,14,15)
```

PubChem (>300K)

ACS Endorsement

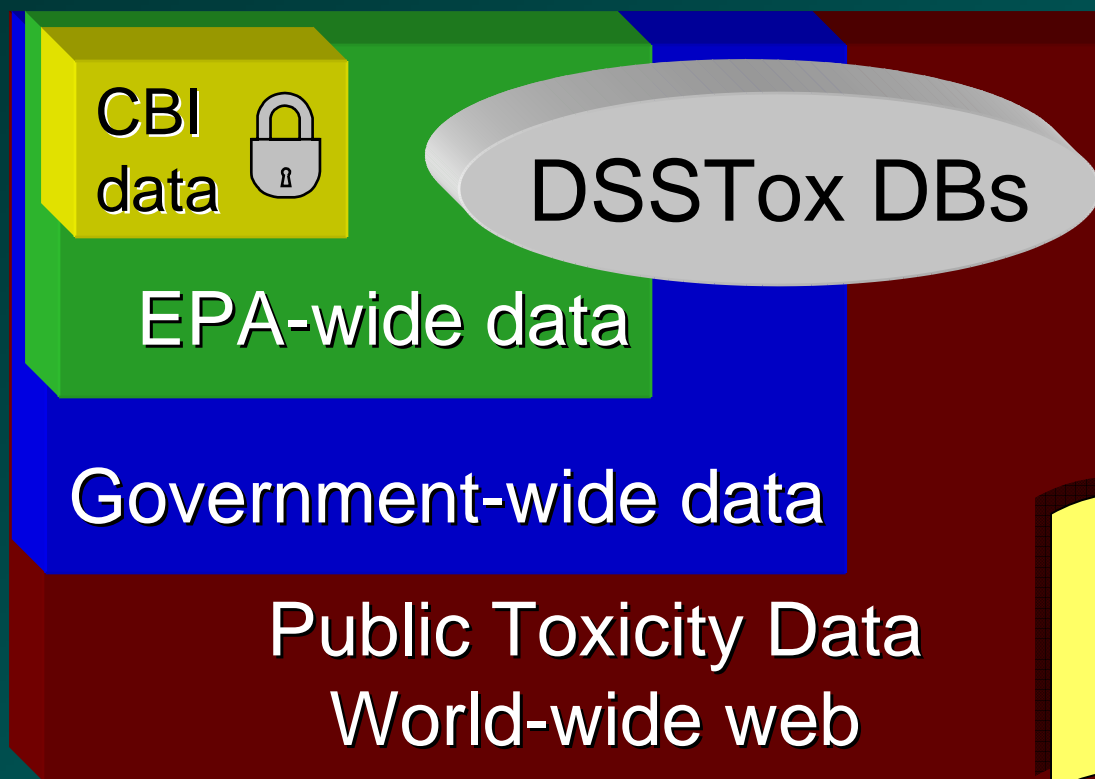
Googling for INChIs; A remarkable method of chemical searching

Peter Murray-Rust^a, Henry S. Rzepa^b and Yong Zhang^a

^aUnilever Centre for Molecular Informatics, University of Cambridge, UK, ^bDepartment of Chemistry, Imperial College London, SW7 2AY,

- ✦ “Electronification” of data
- ✦ Structure-annotation
- ✦ Data standardization
- ✦ Data Integration

EPA Challenges



INChI: Precise
chemical indexing of
information

INChI -> Structures:
Analog searching

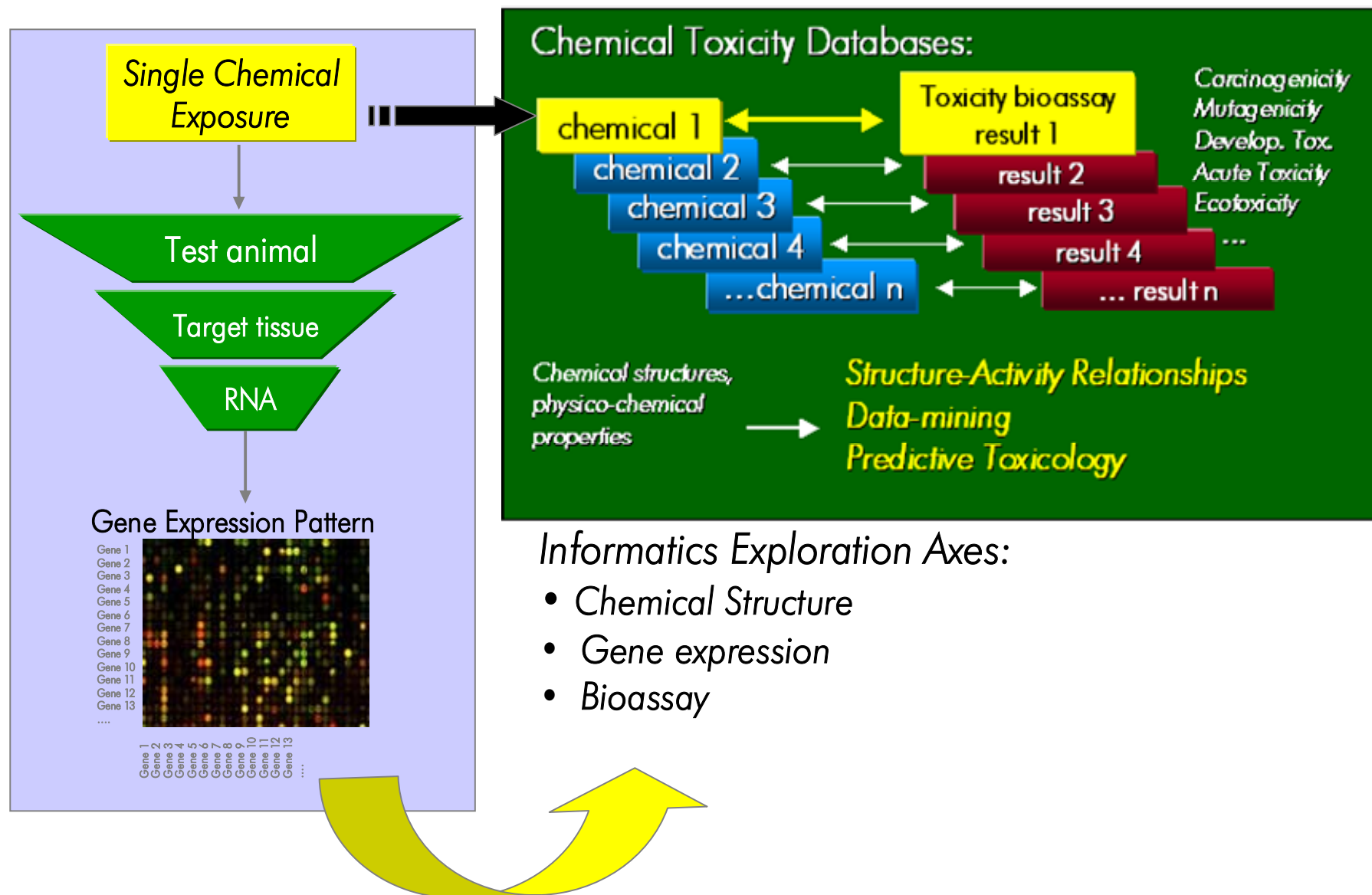
Structure-activity
pattern recognition

**Improved Toxicity
Prediction/Screening
Capabilities**

Part III.

Emerging chemo-bioinformatic
capabilities

Bioinformatics meets Chemoinformatics



NIEHS/National Center for Toxicogenomics: Chemical Effects in Biological System Knowledge- Base *(M. Waters and J. Fostel)*

Gene
expression

Gene
Pathways

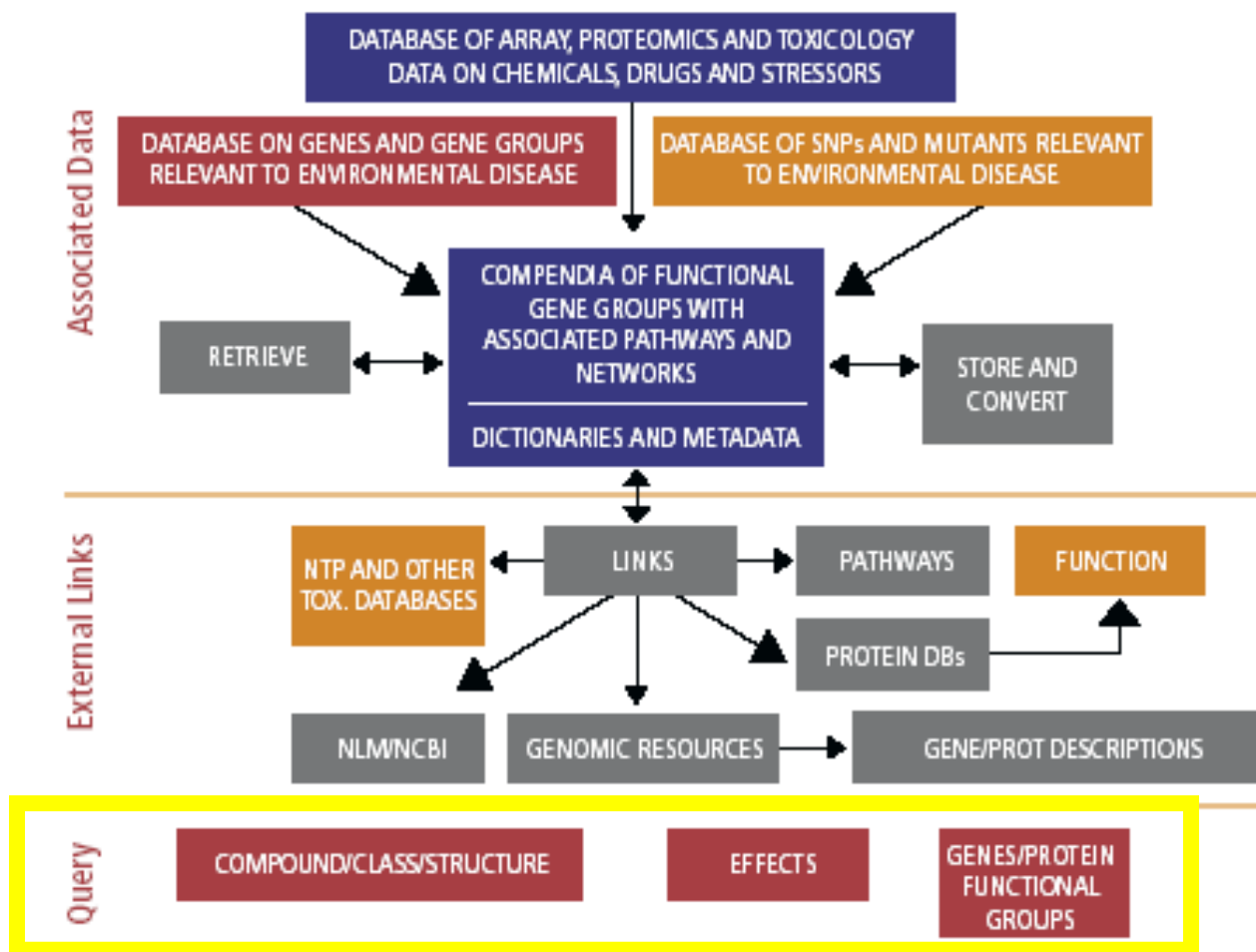
Gene
Function

Proteomics

Metabonomics

Historical NTP
Toxicity data

CEBS Vision - Bioinformatics to Knowledge



DSSTox / CEBS Collaboration:

Part 1 – DSSTox Annotation of CEB

Part 2 – Link CEBS to DSSTox Database Network

DSSTox
Database
Network

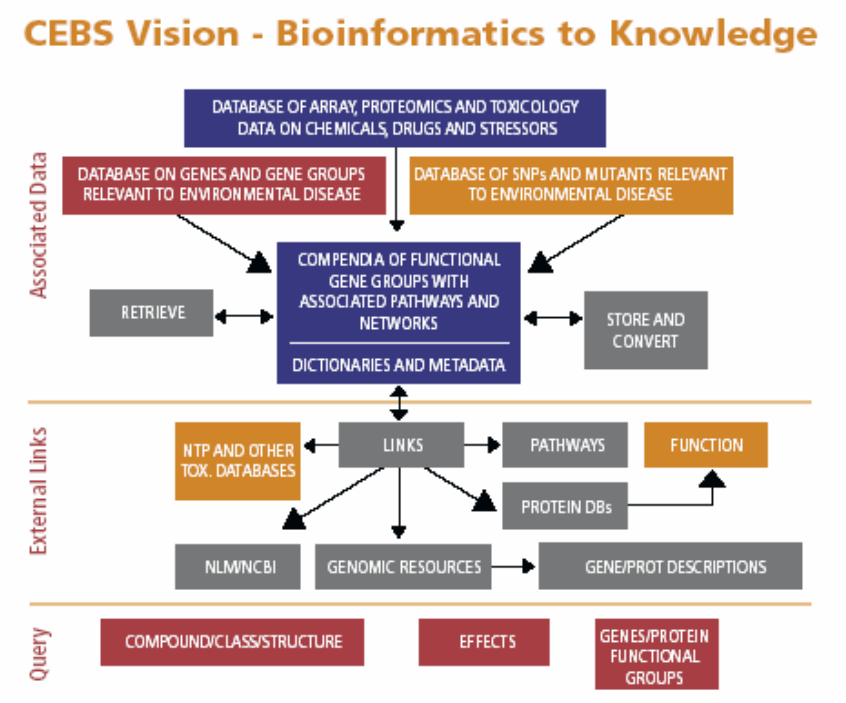
+

- Gene expression
- Gene Pathways
- Gene Function
- Proteomics
- Metabonomics
- Historical NTP Toxicity data

DSSTox Toxicity Data Files



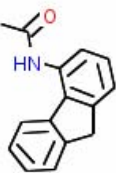
DSSTox Standard Chemical Fields



NCTR: Toxicant Library in ArrayTrack

Linkage of traditional tox data with chemical structure

Compound Structure



Edit Clear

substructure similarity

Structure Search

Compound Fields

Name: contains []

Formula: C [] H [] O []

Mol. ID: CAS_Number []

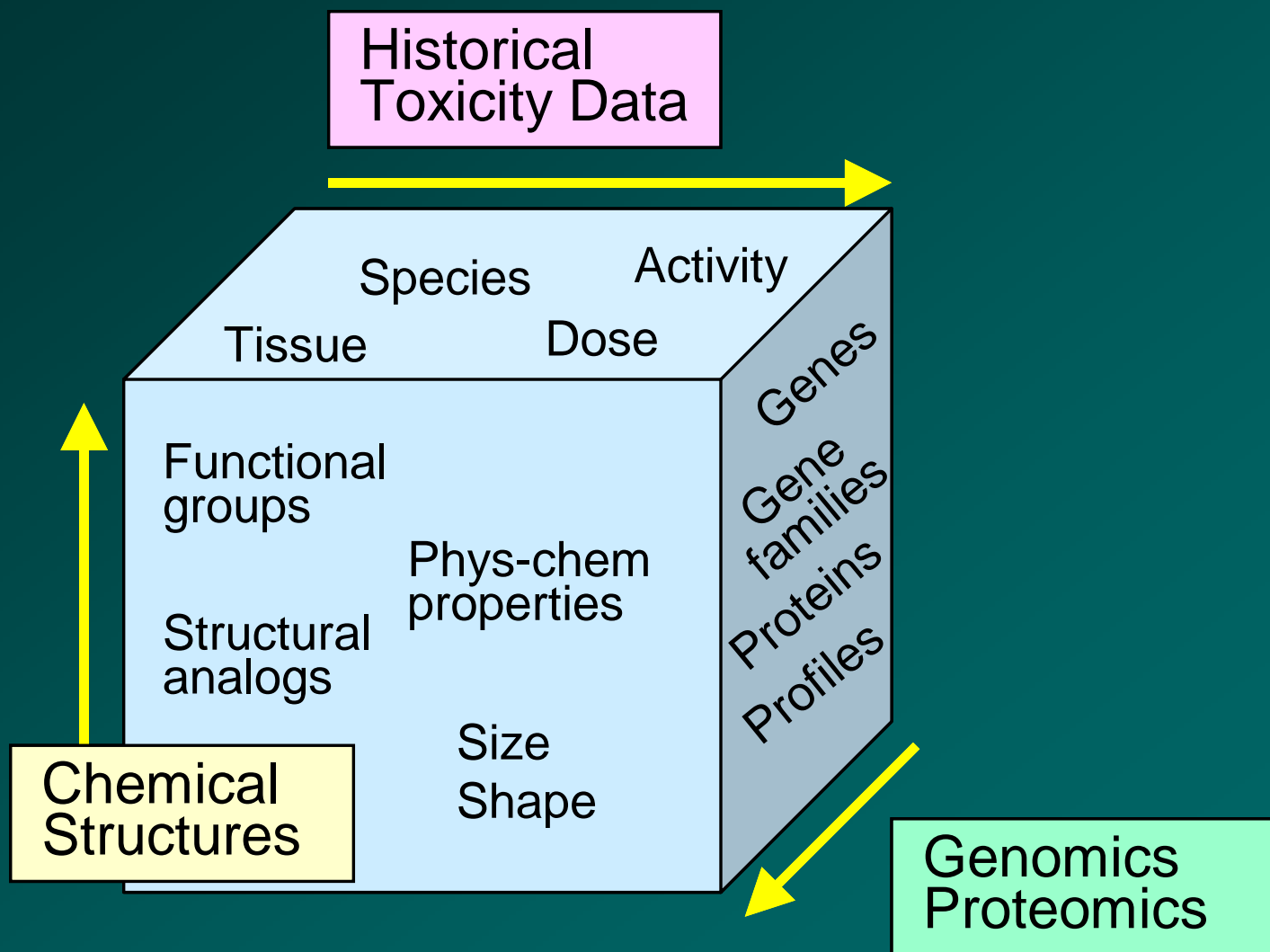
Link To ... Select one -- Go

Pathways ... Customize Table Help

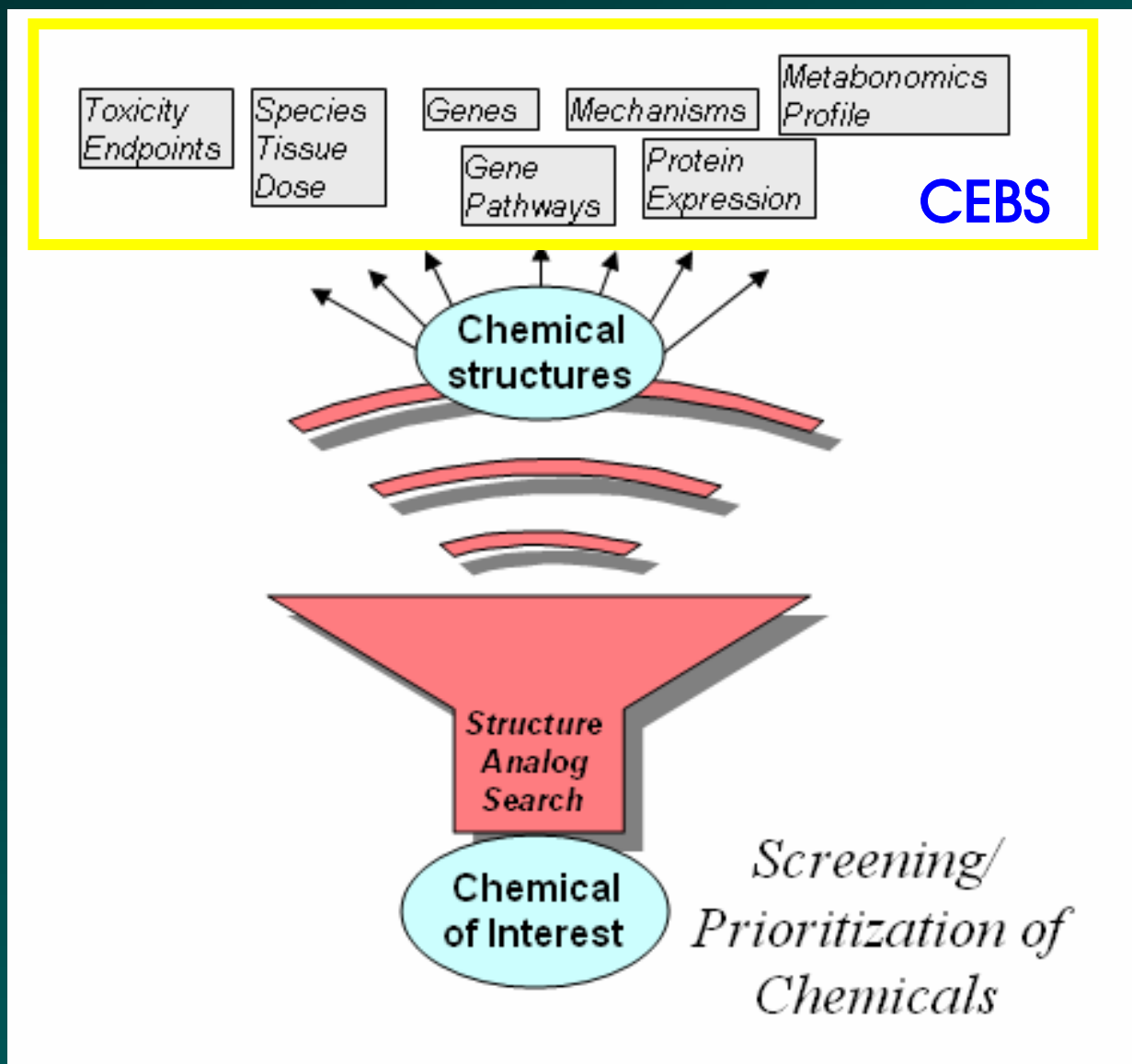
ID	CAS	SIMILA...	NAME	FORMULA	KOWLOGP	ER_CI
1	28322-02-3	1	4-Acetylaminofluorene	C15H13NO	3.16	
2	53-96-3	0.79	2-Acetylaminofluorene	C15H13NO	3.16	
3	591 363-17-7	0.72	"N-(2-Fluorenyl)-2,2,2-trifluoroacetamide"	C15H10F3NO	4.21	
4	17 28314-03-6	0.72	1-Acetylaminofluorene	C15H13NO	3.16	
5	1021 3096-50-2	0.71	N-(9-Oxo-2-fluorenyl)acetamide	C15H11NO2	2.66	
6	361 6301					
7	875 53-9					
8	594 398-					
9	16 4075					
10	1482 243-					
11	874 4463					
12	1647 83-3					
13	1344 128-					
14	1137 2276					
15	178 2475					
16	68 117-					
17	1698 102-					
18	57 82-2					
19	840 90-94-8	0.4	Michler's ketone	C17H20N2O	3.5	
20	816 101-61-1	0.4	"4,4'-Methylenebis(N,N-dimethyl) benzenamine"	C17H22N2	4.37	Misc
21	278 140-49-8	0.4	4'-(Chloroacetyl)-acetanilide	C10H10ClNO2	1.03	
22	111 2465-27-2	0.4	Auramine-O	C17H21N3	2.98	
23	1662 91-88-3	0.39	2-(N-ETHYL-M-TOLUIDINO)ETHANOL	C11H17NO	2.24	
24	516 40762-15-0	0.39	Doxefazepam	C17H15FN2O3	0.58	
25	323 65765-07-3	0.39	Compound 50-892	C19H20N2O2	3.1	
26	1651 87-17-2	0.38	SALICYLANILIDE	C13H11NO2	3.3	
27	1426 15972-60-8	0.38	alachlor	C14H20ClNO2	3.37	Misc
28	1340 2832-40-8	0.38	C.I. disperse yellow 3	C15H15N3O2	3.98	

- Chemical similarity searches
- Linkages to tox information
- Direct links to public databases
- Chemicals mapped to pathways

CEBS Chemo-bioinformatics: Expanded Relational Search Domains



Gathering Data on Chemical Analogues



2005 Catalog
In vitro pharmacology
Early ADME
In vivo models
Standard profiles



- > **Over 530 assays:**
 - > 220 receptors (including 174 GPCRs)
 - > 18 ion channels
 - > 6 transporters
 - > 110 enzymes
 - > 58 specialized cellular assays
 - > 78 functional assays
 - > 59 Early ADME assays
 - > 63 in vivo models
- > **Sub-cellular, cellular and tissue preparations**
- > **60% of human targets**

Home > Catalog online > Selection

Assay by type: enzymes

1 - Family

- Protein-serine/threonine kinases
- Protein-tyrosine kinases
- Phosphatases
- Serine proteases
- Cysteine proteases
- Aspartic proteases
- Metalloproteases
- Phosphodiesterases
- NO synthases

2 - Assays / Catalog reference

<input type="checkbox"/>	AcetylCoA synthetase	757	⌵
<input type="checkbox"/>	Carbonic anhydrase II (h)	762-h	⌵
<input type="checkbox"/>	HMG-CoA reductase	759	⌵
<input type="checkbox"/>	Myeloperoxidase (h)	769	⌵
<input type="checkbox"/>	Xanthine oxidase/ superoxide O2- scavenging	772-1s	⌵

Assay by type

- in vitro pharmacology
 - non-peptide receptors
 - peptide receptors
 - nuclear receptors (steroids)
 - ion channels
 - amine transporters
 - enzymes
 - specialized cellular assays
 - GPCR
- early ADME
- in vivo models

Cerep: In vitro bioactivity profiles

V.N Orechovich Institute
of Biomedical Chemistry

Laboratory
Function

PASS

About PASS

PASS (Prediction of Activity Spectra for Substances)

PASS predicts 900 pharmacological effects, mechanisms of action, mutagenicity, carcinogenicity, teratogenicity, and embryotoxicity. The result of prediction is displayed on your computer automatically via Internet for free.

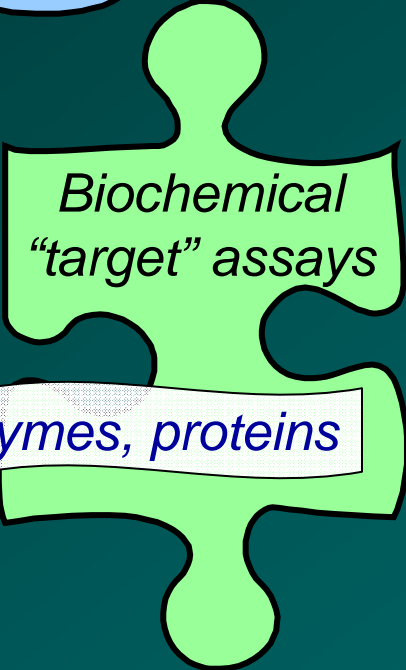
<http://www.ibmh.msk.su/PASS/>

No	Number	MEP, %	Activity Type
5	181	7.494	5 Alpha reductase inhibitor
6	390	7.293	5 Hydroxytryptamine 1 agonist
7	182	8.719	5 Hydroxytryptamine 1 antagonist
76	723	16.416	Alcohol dehydrogenase inhibitor
77	9	9.598	Aldehyde dehydrogenase inhibitor
101	895	22.679	Androgen agonist
102	431	10.859	Androgen antagonist
295	16	26.72	Calcium channel agonist
296	610	15.097	Calcium channel antagonist
300	8	16.719	Cannabinoid receptor agonist
301	4	28.616	Cannabinoid receptor antagonist
308	74	34.815	Cardiotoxic
323	40	17.254	Chloride channel agonist
364	13	16.757	Cytochrome P450 inhibitor
371	125	7.266	DNA intercalator
372	35	6.393	DNA repair enzyme inhibitor
373	288	20.434	DNA synthesis inhibitor
405	656	19.948	Embryotoxic
415	41	1.986	Estradiol 17 beta dehydrogenase stimulant
416	189	5.88	Estrogen agonist
417	144	11.995	Estrogen antagonist
418	326	7.592	Estrogen receptor modulator

Toxico-Chemoinformatics: Data Standardization, Integration, Exploration



Calculated structures, properties



*Biochemical
"target" assays*



Receptors, enzymes, proteins



*In Vitro
assays*




Short-term tests



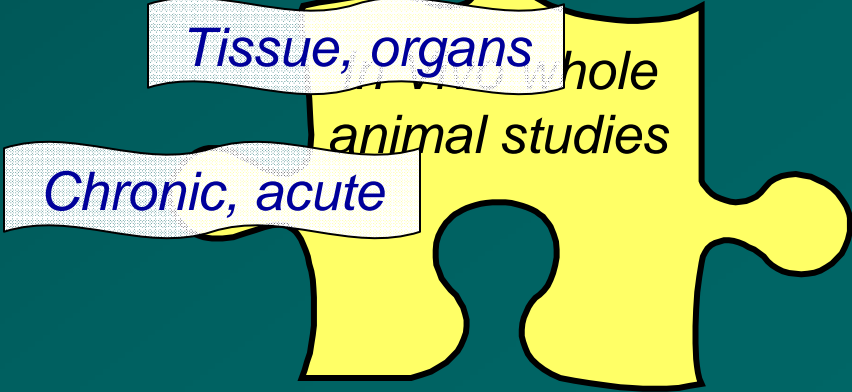
Cell-based assays



Genomics



Tissue, organs



*Whole
animal studies*



Chronic, acute

Toxico-Chemoinformatics: Data Standardization, Integration, Exploration

- ✦ Make better use of ALL available data
- ✦ Overcome data limitations by exploring diverse domains of data from multiple perspectives
- ✦ Develop expanded definitions of “chemical analog”
- ✦ Employ both biological and chemical information to develop predictive toxicity signatures