



NCI  
Center  
for  
Cancer Research

*Reducing the Burden of Cancer Through  
Exploration, Discovery and Translation*



# The IBM Patent Data Donation to NIH, and its Integration in the NCI/CADD Database and Web Services

Marc C. Nicklaus

Computer-Aided Drug Design Group, Chemical Biology Laboratory,  
Frederick National Laboratory for Cancer Research, National Cancer Institute, NIH, DHHS

# CADD Group

Main areas of research:

- *In silico* support of drug development projects at NCI
- Fundamentals of protein-ligand interactions
- Chemoinformatics: methods, tools, and public web services

# CADD Group Web Services and Databases



## CADD Group Chemoinformatics Tools and User Services

---

[Online Services](#) | [Tools & Software](#) | [Downloads](#) | [Links & Data](#) | [Blog](#)

### General Information

---

The NCI/CADD group is a research unit within the [Chemical Biology Laboratory](#), which is part of the [Molecular Discovery Program](#) at the [National Cancer Institute](#). Read [more about](#) the CADD Group's Chemoinformatics Tools and User Services. Or, visit [our blog](#).

### Online Services

---

*To search and display chemical structures here, you will need Java! JavaScript to be enabled on your browser.*

#### Chemical Structure Lookup Service (CSLS) San Diego mirror

Look up whether a structure occurs in many different databases, both public and commercial. Currently loaded pointers to: over 74 million entries from more than 100 databases, representing more than 46 million unique chemical structures.

#### Chemical Identifier Resolver

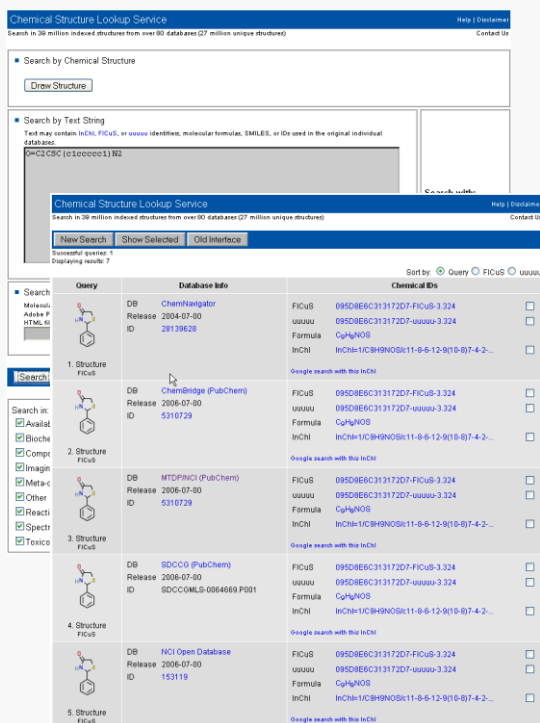
This service works as a resolver for different chemical structure identifiers and allows the conversion of a given structure identifier into another representation or structure identifier. It can be used via a web form or a simple URL API. **Beta version.**

#### Enhanced NCI Database Browser Release 2

A web service to the open NCI database compounds (>250,000 structures) with different kinds of output features and links to other services for continued processing.

Additional information about CADD Group web services and projects:  
<http://cactus.nci.nih.gov/blog>

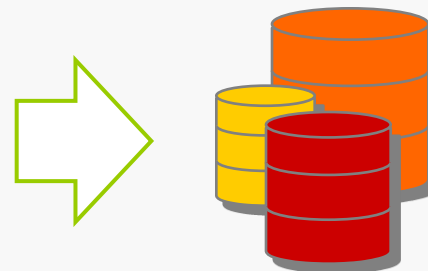
# Chemical Structure Lookup Service (CSLS)



The screenshot displays the CSLS web interface. At the top, it states "Search in 36 million indexed structures from over 80 databases (27 million unique structures)". Below this, there are search options: "Search by Chemical Structure" (with a "Draw Structure" button) and "Search by Text String". A search for "C1=CC=C(C=C1)N" is shown. The results table is as follows:

Query	Database Info	Chemical IDs
1. Structure FICuS	DB ChemNavigator Release 2004-07-00 ID 28739629	FICuS 095DBEC313172D7-FICuS-3.324 uuuuu 095DBEC313172D7-uuuuu-3.324 Formula C <sub>6</sub> H <sub>7</sub> NOS InChI InChI=1C6H7NOSr11-8-6-12-9(10-8)7-4-2...
2. Structure FICuS	DB ChEMBL (PubChem) Release 2006-07-00 ID 5310729	FICuS 095DBEC313172D7-FICuS-3.324 uuuuu 095DBEC313172D7-uuuuu-3.324 Formula C <sub>6</sub> H <sub>7</sub> NOS InChI InChI=1C6H7NOSr11-8-6-12-9(10-8)7-4-2...
3. Structure FICuS	DB MTPD (NCI) (PubChem) Release 2006-07-00 ID 5310729	FICuS 095DBEC313172D7-FICuS-3.324 uuuuu 095DBEC313172D7-uuuuu-3.324 Formula C <sub>6</sub> H <sub>7</sub> NOS InChI InChI=1C6H7NOSr11-8-6-12-9(10-8)7-4-2...
4. Structure FICuS	DB SDCO (PubChem) Release 2006-07-00 ID SDCOMLS-0064669.P001	FICuS 095DBEC313172D7-FICuS-3.324 uuuuu 095DBEC313172D7-uuuuu-3.324 Formula C <sub>6</sub> H <sub>7</sub> NOS InChI InChI=1C6H7NOSr11-8-6-12-9(10-8)7-4-2...
5. Structure FICuS	DB NCI Open Database Release 2006-07-00 ID 153119	FICuS 095DBEC313172D7-FICuS-3.324 uuuuu 095DBEC313172D7-uuuuu-3.324 Formula C <sub>6</sub> H <sub>7</sub> NOS InChI InChI=1C6H7NOSr11-8-6-12-9(10-8)7-4-2...

*Where to find a certain chemical structure?*



link to the original database(s)

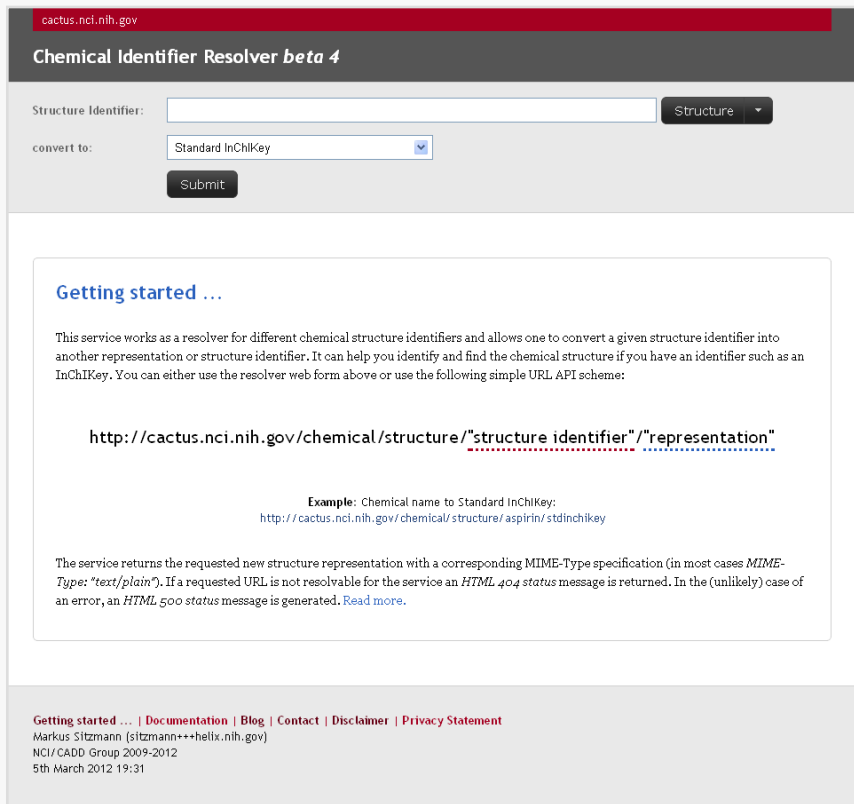
first released August 2006

last updated December 2008

major update to CSLS II imminent

<http://cactus.nci.nih.gov/lookup>

# Chemical Identifier Resolver (CIR)



The screenshot shows the web interface for the Chemical Identifier Resolver (CIR) at cactus.nci.nih.gov. The page title is "Chemical Identifier Resolver beta 4". It features a form with a "Structure Identifier:" input field, a "Structure" dropdown menu, a "convert to:" dropdown menu set to "Standard InChIKey", and a "Submit" button. Below the form is a "Getting started ..." section with the following text:

This service works as a resolver for different chemical structure identifiers and allows one to convert a given structure identifier into another representation or structure identifier. It can help you identify and find the chemical structure if you have an identifier such as an InChIKey. You can either use the resolver web form above or use the following simple URL API scheme:

`http://cactus.nci.nih.gov/chemical/structure/"structure identifier"/"representation"`

**Example:** Chemical name to Standard InChIKey:  
`http://cactus.nci.nih.gov/chemical/structure/aspirin/stdinchkey`

The service returns the requested new structure representation with a corresponding MIME-Type specification (in most cases *MIME-Type: "text/plain"*). If a requested URL is not resolvable for the service an *HTML 404* status message is returned. In the (unlikely) case of an error, an *HTML 500* status message is generated. [Read more.](#)

At the bottom, there are links for "Getting started ...", "Documentation", "Blog", "Contact", "Disclaimer", and "Privacy Statement". The footer includes the name "Markus Sitzmann (sitzmann+++helix.nih.gov)", the text "NCI / CADD Group 2009-2012", and the date "5th March 2012 19:31".

*CIR works as a resolver for different chemical structure identifiers or representations.*

*It allows one to convert a given structure identifier into another representation or structure identifier.*

***(REST-like service)***

<http://cactus.nci.nih.gov/chemical/structure>

# Chemical Structure Representations

SMILES

SYBYL Line Notation

CAS Registry Number

chemical names

GIF image

ChemNavigator SID

SD File

CML

**chemical structure**

FDA UNII

NCI/CADD Identifiers

NSC number

MRV

InChI/InChIKey

PubChem SID/CID

ChemSpider ID

ChEBI ID

Chemical Formula

PDB Ligand ID

# Chemical Identifier Resolver (CIR)

programmatic URL API:

`http://cactus.nci.nih.gov/chemical/structure/"identifier"/"representation"`



# Chemical Identifier Resolver (CIR)

...but also perfectly usable by humans:

<http://cactus.nci.nih.gov/chemical/structure/aspirin/stdinchikey>

returns...

InChIKey=BSYNRYMUTXBXSQ-UHFFFAOYSA-N

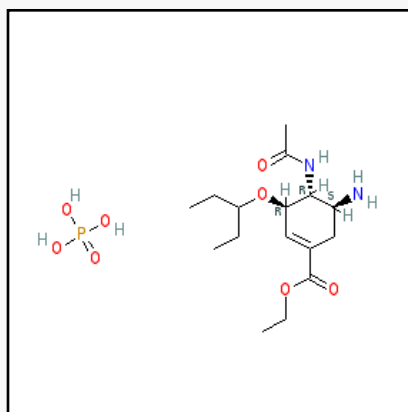
# Chemical Identifier Resolver (CIR)

if a request is not successful: **HTTP404 status message**

otherwise, you get the requested representation, e.g.:

<http://cactus.nci.nih.gov/chemical/structure/PGZUMBJQJWIWGJ-ONAKXNSWSA-N/image>

.....



**MIME type: image/gif**

# Chemical Identifier Resolver (CIR)

chemical names  
IUPAC names (OPSIN)  
CAS Registry Numbers  
SMILES strings  
IUPAC InChI/InChIKeys  
NCI/CADD Identifiers  
CACTVS HASHISY  
NSC number  
PubChem SID  
ZINC Code  
ChemSpider ID  
ChemNavigator SID  
eMolecule VID  
FDA UNII  
chemical formula

“identifier”



<http://cactus.nci.nih.gov/chemical/structure>

/smiles  
/names, /iupac\_name  
/cas  
/inchi, /stdinchi  
/inchikey, /stdinchikey  
/ficts, /ficus, /uuuuu  
/image  
/file, /sdf  
/mw, /monoisotopic\_mass  
/formula  
/twirl  
/urls  
/chemspider\_id  
/pubchem\_sid  
/chemnavigator\_sid

“representation”



# Chemical File Representation in CIR

- available file format representations:

*Aspirin*

<http://cactus.nci.nih.gov/chemical/structure/BSYNRYMUTXBXSQ-UHFFFAOYSA-N/file?format=sdf>

---

**alc** Alchemy format

**cdxml** CambridgeSoft ChemDraw XML format

**cerius** MSI Cerius II format

**charmm** Chemistry at HARvard  
Macromolecular Mechanics file format

**cif** Crystallographic Information File

**cml** Chemical Markup Language

**gjf** Gaussian input data file

**gromacs** GROMACS file format

**hyperchem** HyperChem file format

**jme** Java Molecular Editor format

**maestro** Schroedinger MacroModel  
structure file format

**mol** Accelrys molecule file format

**sybyl2/mol2** Tripos Sybyl MOL2 format

**mrw** ChemAxon MRV format

**pdb** Protein Data Bank

**sdf** Accelrys Structure Data Format

**sdf3000** Accelrys Structure Data Format 3000

**sln** SYBYL Line Notation

**smiles** SMILES

**xyz** xyz file format

# Chemical Properties in CIR

- request molecular weight:

*Aspirin*

<http://cactus.nci.nih.gov/chemical/structure/BSYNRYMUTXBXSQ-UHFFFAOYSA-N/weight>

180.1598

MIME type: text/plain

**/mw** molecular weight

**/formula** formula

**/monoisotopic\_mass** monoisotopic mass

**/h\_bond\_donor\_count** H bond donor count

**/h\_bond\_acceptor\_count** H bond acceptor count

**/h\_bond\_center\_count** H bond center count

**/rotor\_count** number of rotatable bonds

**/effective\_rotor\_count** number of effectively rotatable bonds

**/rule\_of\_5\_violation\_count** number of Rule-of-5 violations

**/xlogp2** octanol–water partition coefficient XLOGP2

**/aromatic** compound is aromatic

**/macrocyclic** compound is macrocyclic

**/heteroatom\_count** heteroatom count

**/hydrogen\_atom\_count** H atom count

**/heavy\_atom\_count** heavy atom count

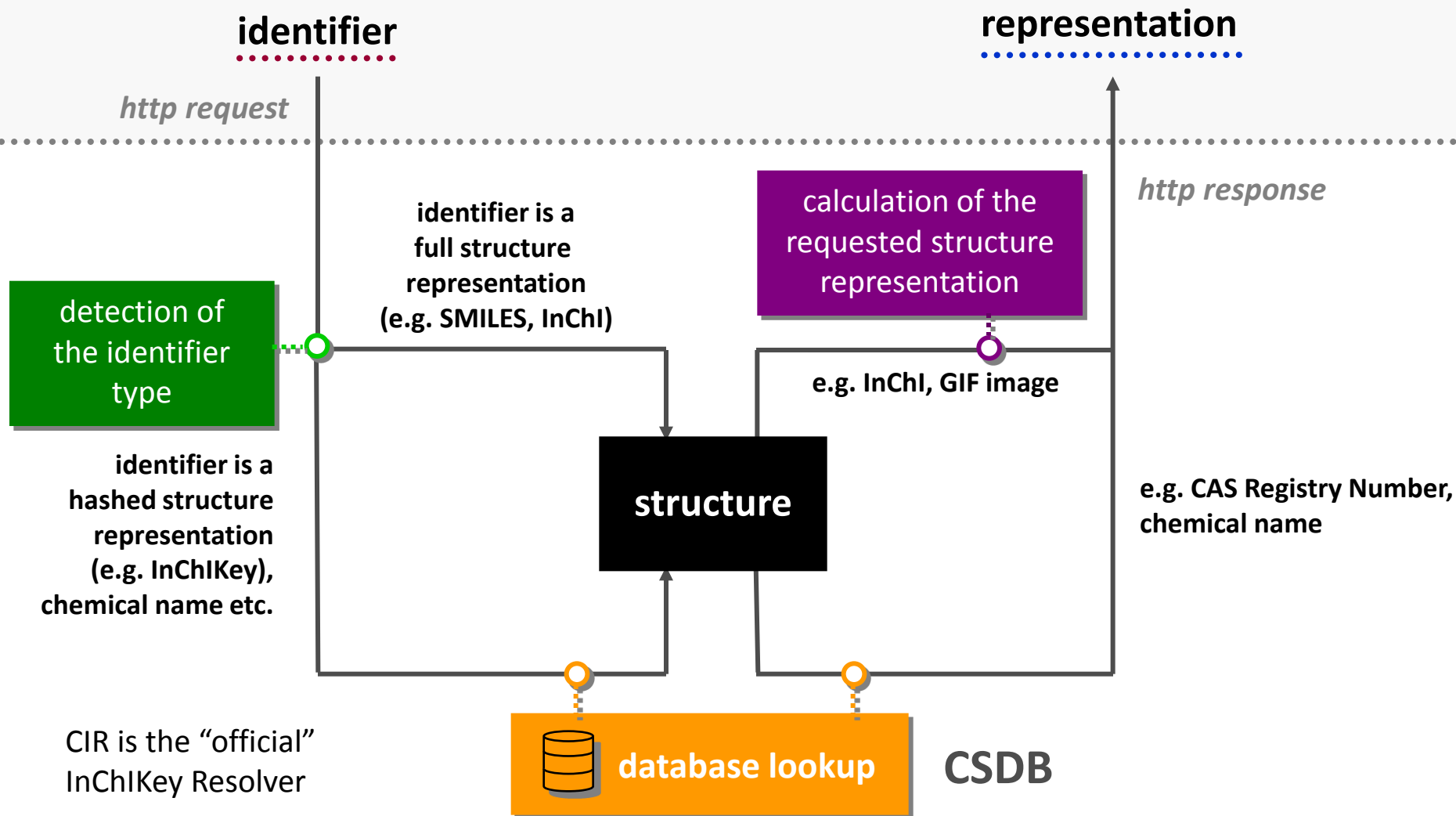
**/deprotonable\_group\_count** number of deprotonable groups

**/protonable\_group\_count** number of protonable groups

**/ring\_count** number of rings

**/ringsys\_count** number of ring systems

# Computation or Database Lookup in CIR



# Small-Molecule Databases

PubChem

DRUGBANK

Open Data Drug & Drug Target Database

eMolecules

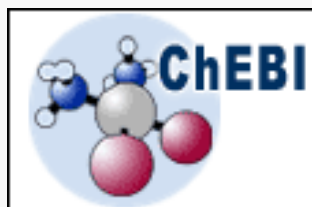


IUPHAR DATABASE  
International Union of Basic and Clinical Pharmacology



ZINC<sup>12</sup>

ChEMBL



ChemSpider  
The free chemical database

*many more ...*

# Chemical Structure Database (CSDB)

## ChemNavigator iResearch Library

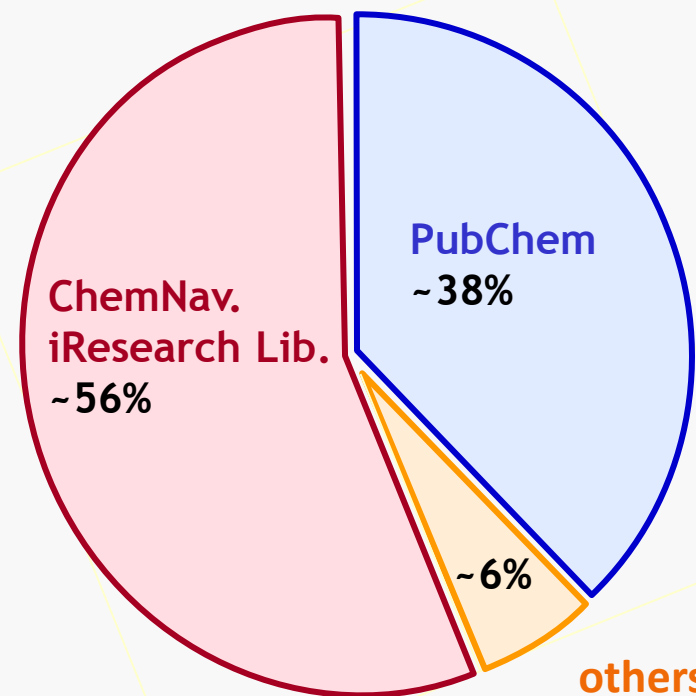
compilation of commercially available screening compounds from ~300 international chemistry suppliers

## PubChem database

including Open NCI database, EPA DSSTox databases, NIAID HIV database, NIST Webbook, NLM ChemIDplus, ChemSpider, ...

## Commercial Sources / others

Asinex, Comgenex, eMolecules, ...



**Currently available in CIR:**  
(CSDB status as of March 2010)

**140** chemical structure databases

**120 million** structure records

**84.6 million** unique structures

**110 million** **Standard InChIKeys** for lookup



# Chemical Structure Database (Update 2012)

Coming soon to CIR:

231 small-molecule database

367 database releases (full, incremental, “historic versions”)

324 million original database records

**Unique structure count:**

**FICTS**

**~118 million**

**FICuS**

**~115 million**

**uuuuu**

**~100 million**

FICTS, FICuS, uuuuu: Identifiers developed by CADD Group, based on CACTVS hash codes.

FICuS: ID for “stuff in the bottle”

Example: C402B3DA608808FC-FICuS-01-61

# IBM Patent Data Donation to NIH

# IBM Patent Data Donation to NIH

What was donated?

Three text files with a total of 43.5 million records, containing data extracted from

- USPTO patents: 23.1 million records
- EPO patents: 14.8 million records
- WIPO patents: 5.5 million records

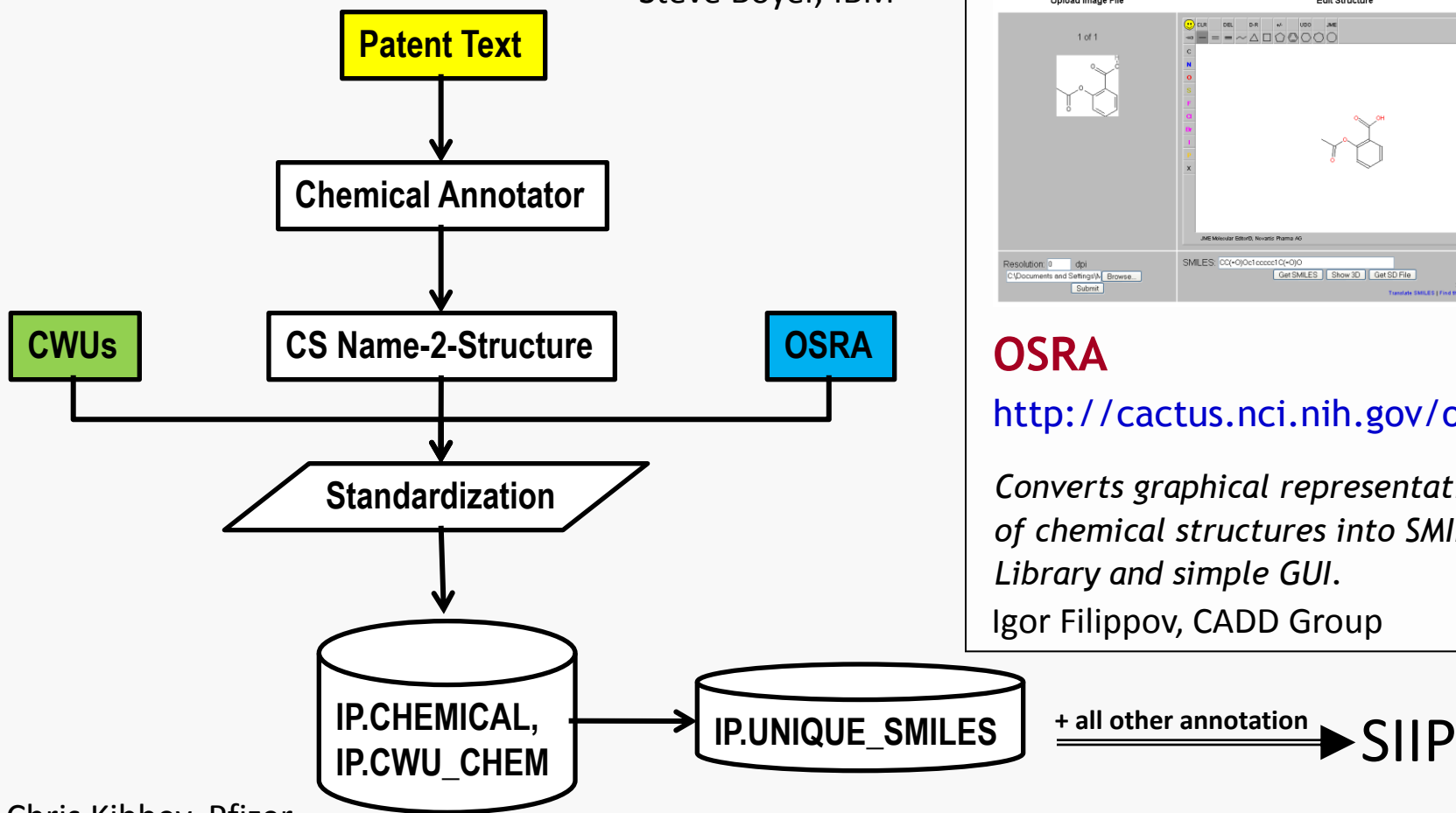
Time range covered:

Patents through 2000

# IBM Patent Data Donation to NIH

## Overall Process by IBM/Consortium

Steve Boyer, IBM



Chris Kibbey, Pfizer

**OSRA: Optical Structure Recognition**

The screenshot shows the OSRA web application interface. It includes an "Upload Image File" section, an "Edit Structure" section with a chemical structure editor, and a "Resolution" section. The SMILES string displayed is CC(=O)Oc1ccccc1C(=O)O. The interface also shows a "Similarity: 1.000" value.

**OSRA**  
<http://cactus.nci.nih.gov/osra/>  
*Converts graphical representations of chemical structures into SMILES. Library and simple GUI.*  
Igor Filippov, CADD Group

# IBM Patent Data Donation to NIH – SIIP

Business consulting > Business Analytics and Optimization services >

## IBM BAO strategic IP insight platform (SIIP)



### Gain Game Changing Insights from Intellectual Property delivered via the Cloud

Insights from the Intellectual Property (IP) in patents and scientific literature can have significant impact on your organization. Examples include:

- Improve R&D productivity
- Increase licensing from IP
- Support open innovation initiatives
- Identify potential collaborators or acquisition targets
- Obtain competitive intelligence
- Increase the win rate for patent and grant applications

The IBM BAO strategic IP insight platform (SIIP) applies deep analytics on patent and scientific literature to provide you with more information and insights than you would be able to derive manually. SIIP has unique capabilities to extract and analyze chemical and biological data to help inform research and development and intellectual property management.

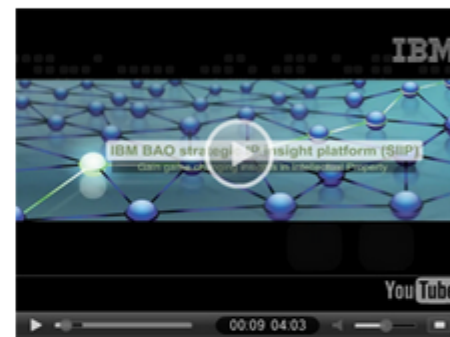
- [Learn more \(766KB\)](#)
- [Contact an expert](#)

### IBM announces Donation of Chemical Information to National Institutes of Health (NIH)

IBM has donated a database containing over 2.5 million unique chemical compounds and their reference in patents and scientific literature. IBM used the strategic IP insight platform solution (SIIP) to develop these insights.

[→ Learn more and go to the IBM NIH donation web site](#)

### Discover game-changing intellectual property insights delivered on the cloud



### Gain insight. Optimize results.



To learn more about IBM's full breadth of Business Analytics and Optimization capabilities, visit [ibm.com/bao](http://ibm.com/bao)

### Learn about IBM Watson and DeepQA

Follow the story of the development of Watson from a modest DeepQA machine to a formidable Jeopardy! contestant. Watch the videos below to see how Watson represents a leap forward in data analytics and how this technology will impact business and industry.



[What is Watson?](#)

# IBM Patent Data Donation to NIH

## Dates and Recipients of the Donation

Donation announced by IBM in late 2011.

Several versions (updates) were prepared by IBM, the first one in late November 2011 (first announcement by PubChem on 8 Dec 2011), the last one on 16 May 2012.

Given to PubChem, CADD Group,....

... ChEMBL, Tropsha group (UNC), Oprea group (UNM), Hopkins group (Dundee, UK)

# IBM Patent Data Donation to NIH

## Format of the Donation Files

"US1255423","CN1CC[C@@]23[C@H]4Oc5c2c(C[C@@H]1[C@@@H]3C=C[C@@H]4O)ccc5O","BQJCRHHNABKAKU-KBQPJGBKSA-N",  
"534DDCA63F163B84DBB31E73BE06632D",0,"http://www.ibm.com/gbs/nih/?sid=534DDCA63F163B84DBB31E73BE06632D&ik=BQJCRHHNABKAKU-KBQPJGBKSA-N","T"

"US1258747","NC#N","XZMCDZFZZKTWFGF-UHFFFAOYSA-N",  
"E03544A04BBBC3FCE272442D86F149CA",0,"http://www.ibm.com/gbs/nih/?sid=E03544A04BBBC3FCE272442D86F149CA&ik=XZMCDZFZZKTWFGF-UHFFFAOYSA-N","T"

"US1265772","OS(=O)(=O)CCN1CCN(CCS(O)(=O)=O)CC1","IHPYMWDTONKSCO-UHFFFAOYSA-N",  
"5DAE9478928DBF41AC48D766F7157BA5",0,"http://www.ibm.com/gbs/nih/?sid=5DAE9478928DBF41AC48D766F7157BA5&ik=IHPYMWDTONKSCO-UHFFFAOYSA-N","T"

"US1266657","CCO","LFQSCWFLJHTTHZ-UHFFFAOYSA-N",  
"C677B6890A9B0247071F07B1A1C3671D",0,"http://www.ibm.com/gbs/nih/?sid=C677B6890A9B0247071F07B1A1C3671D&ik=LFQSCWFLJHTTHZ-UHFFFAOYSA-N","T"

"US1275535","NC#N","XZMCDZFZZKTWFGF-UHFFFAOYSA-N",  
"E03544A04BBBC3FCE272442D86F149CA",0,"http://www.ibm.com/gbs/nih/?sid=E03544A04BBBC3FCE272442D86F149CA&ik=XZMCDZFZZKTWFGF-UHFFFAOYSA-N","T"

"US1281662","COc1ccccc1O","LHGVFZTZFXWLCP-UHFFFAOYSA-N",  
"F6FE90CDB94ABF4B149FC848F5EE37CA",0,"http://www.ibm.com/gbs/nih/?sid=F6FE90CDB94ABF4B149FC848F5EE37CA&ik=LHGVFZTZFXWLCP-UHFFFAOYSA-N","T"

"US1290124","c1ccncc1","JUJWROOIHBZHMG-UHFFFAOYSA-N",  
"06FCB7709728689AD3FDCA3768A17532",0,"http://www.ibm.com/gbs/nih/?sid=06FCB7709728689AD3FDCA3768A17532&ik=JUJWROOIHBZHMG-UHFFFAOYSA-N","T"

"US1350532","c1ccccc1","UHOVQNZJYSORNB-UHFFFAOYSA-N",  
"DCF41741A64B5F9798E62E5599420465",0,"http://www.ibm.com/gbs/nih/?sid=DCF41741A64B5F9798E62E5599420465&ik=UHOVQNZJYSORNB-UHFFFAOYSA-N","A"

# IBM Patent Data Donation to NIH

## Number of Unique Structures

Unique structures by HASHISY (CACTVS): 2.4 million.

Generation of NCI/CADD identifiers (FICTS, FICuS, uuuuu) by database lookup succeeded for 1,921,813 structures.

Unique FICuS: 1,915,806



# IBM Patent Data Donation to NIH

## Loss of Unique Structures (by FICuS Identifier)

Unique structures by “raw” HASHISY hash code: 2.4 million; unique structures by FICuS identifier: 1.9 million => ~500,000 structures lost.

Main reason: unclear situation with hydrogen numbers. E.g.:

CC=[CH]

CCO[C]=O

CCS(=O)(=O)OC(=O)[C@H](N)CC(=O)c1ccc(Cl)c(Cl)c1

[CH2]C(CCO)=C=O

[c]1ccncc1

[c]1cc[nH]c1

...seems to typically indicate some kind of R-group

# IBM Patent Data Donation to NIH

## Overlap with Other Databases

Overlap (by structure) found with 191 databases out of 217 databases in CSDB

% Overlap	Database	
98.7	PubChem Substance	← after IBM Donation
96.3	PubChem Compound	← after IBM Donation
64.6	ChemSpider	← contains SureChem data
17.9	DiscoveryGate	
11.5	iResearch Library	← “commercially available”
9.07	Thomson Pharma	
8.22	ABI Chem	
5.64	ChemFrog	
5.55	eMolecules	
4.91	ChemDB	
4.20	ZINC	
3.55	Ambinter	
3.36	ChEMBL	
3.31	NLM ChemIDplus	
...	...	

# IBM Patent Data Donation to NIH

## Content Analysis – Occurrence Statistics (by SMILES)

Frequency bin	Structure occurrences	Unique structures	Patent count
1-10	5,135,722	2,294,145	365,760
11-20	919,195	64,122	257,360
21-30	524,413	21,078	209,106
31-40	390,170	11,144	184,033
41-50	326,966	7,227	166,131
51-60	287,938	5,203	150,710
61-70	246,024	3,773	140,012
71-80	213,034	2,831	130,033
81-90	200,857	2,354	124,374
91-100	185,067	1,940	117,798
101-1000	5,793,769	19,832	683,289
1001-10000	9,697,745	3,556	927,646
10001-100000	12,681,936	493	1,102,541
100001-1000000	6,912,856	41	1,165,091

# IBM Patent Data Donation to NIH

## Content Analysis – Most Frequent Structures (SMILES)

CADD ID	SMILES	Count
2563229	<chem>CCO</chem>	509594
2563231	<chem>CC(O)=O</chem>	365105
2563201	<chem>OC(O)=O</chem>	333518
2563211	<chem>Cc1ccccc1</chem>	243161
2563219	<chem>CC=C</chem>	242819
2563426	<chem>CC(C)=O</chem>	219641
2563205	<chem>[c]1ccccc1</chem>	208303
2563297	<chem>c1ccccc1</chem>	205048
2563427	<chem>CC(C)O</chem>	182523
2563241	<chem>CICCI</chem>	181793
2563499	<chem>Oc1ccccc1</chem>	178134
2563193	<chem>O1C=C1</chem>	173918
2563274	<chem>OCC(O)CO</chem>	170035
2563209	<chem>C1CCOC1</chem>	167243
2563451	<chem>[CH2]CCC</chem>	162248
2563225	<chem>CCOC(C)=O</chem>	159417

Top 100 represent approx. a quarter of the entire dataset (10 of 43 million patent records) and occur in 1,251,402 of the 1,506,078 patents; top 1000 represent about half of the records.

# IBM Patent Data Donation to NIH

## Content Analysis – Class Codes & Place of Occurrence

### Distribution of IBM class codes

ibm_class_code	count	description
0	37892993	Whole Molecule
1	161953	Markush Scaffold
2	5453586	Fragment
3	3629	Markush + Fragment
4	1960	Polymer
5	1508	Markush + Polymer
6	23	Polymer Fragment
7	30	Markush Polymer Fragment
8	10	Enumerable (Rgmolecule)

### Distribution of occurrences

field	count
A (bstract)	917023
B (ody)	42530466
C (laims)	5107038
T (itle)	159528

# IBM Patent Data Donation to NIH

## Content Analysis – Markush “SMILES” (examples)

```
CC1CCN(CC1)C(=O)OCCC1(C(=O)*CC(F)(F)F)c2ccccc2-c2ccccc12  
CCOC(C)=O.Cc1cn2cc(C)ccc2n1  
CC.CN(C(=O)**c1ccc(cc1)C(=O)\C=C\C=O)c1nc2ccccc2s1  
CCc1ccc(cc1)C1C(C(CN1*))c1ccc2occc2c1)c1nn[nH]n1  
COC(*)=C(C#N)C1CC(C)=NO1  
Cc1c[nH]c(O*)n1  
Cc1ccc(cc1)C1CCC(CCc2ccc(C)cn2)CC1
```

“The ‘\*’ was used for any non-interpretable atom label in the structures, which usually came from CWU data, and typically means some sort of complex label that ChemDraw could not interpret during export to MOL format.”

# IBM Patent Data Donation to NIH

## Content Analysis – Stereochemistry

Database	No. of molecules	Molecules with stereocenter	%	Stereo molecules well defined	%	Average no. of stereocenters/-bonds per stereo molecule
<b>STEREOATOM (chiral center)</b>						
<b>IBM filtered</b>	1,895,955	808,701	<b>42.7</b>	187,476	<b>23.2</b>	2.01
<b>ChEMBL</b>	1,212,539	527,042	<b>43.5</b>	254,186	<b>48.2</b>	3.22
<b>STEREOBOND (E/Z double bond)</b>						
<b>IBM filtered</b>	1,895,955	144,726	<b>7.6</b>	1,064	<b>0.7</b>	1.26
<b>ChEMBL</b>	1,212,539	142,451	<b>11.7</b>	123,291	<b>86.5</b>	1.23
<b>STEREOATOM or STEREOBOND</b>						
<b>IBM filtered</b>	1,895,955	898,205	<b>47.4</b>	171,482	<b>19.1</b>	
<b>ChEMBL</b>	1,212,539	618,206	<b>51.0</b>	328,574	<b>53.1</b>	

# IBM Patent Data Donation to NIH

## Content Analysis – PubChem’s Analysis

Unique chemicals	2,362,985
Rule-of-5 compliant	1,649,589
Biologically tested	94,809
...and declared active	46,790
Found in PDB (as ligands)	3,981
Approved drugs	340
NIH Roadmap chemical “probes”	5

(smaller set is not necessarily subset of previous set)

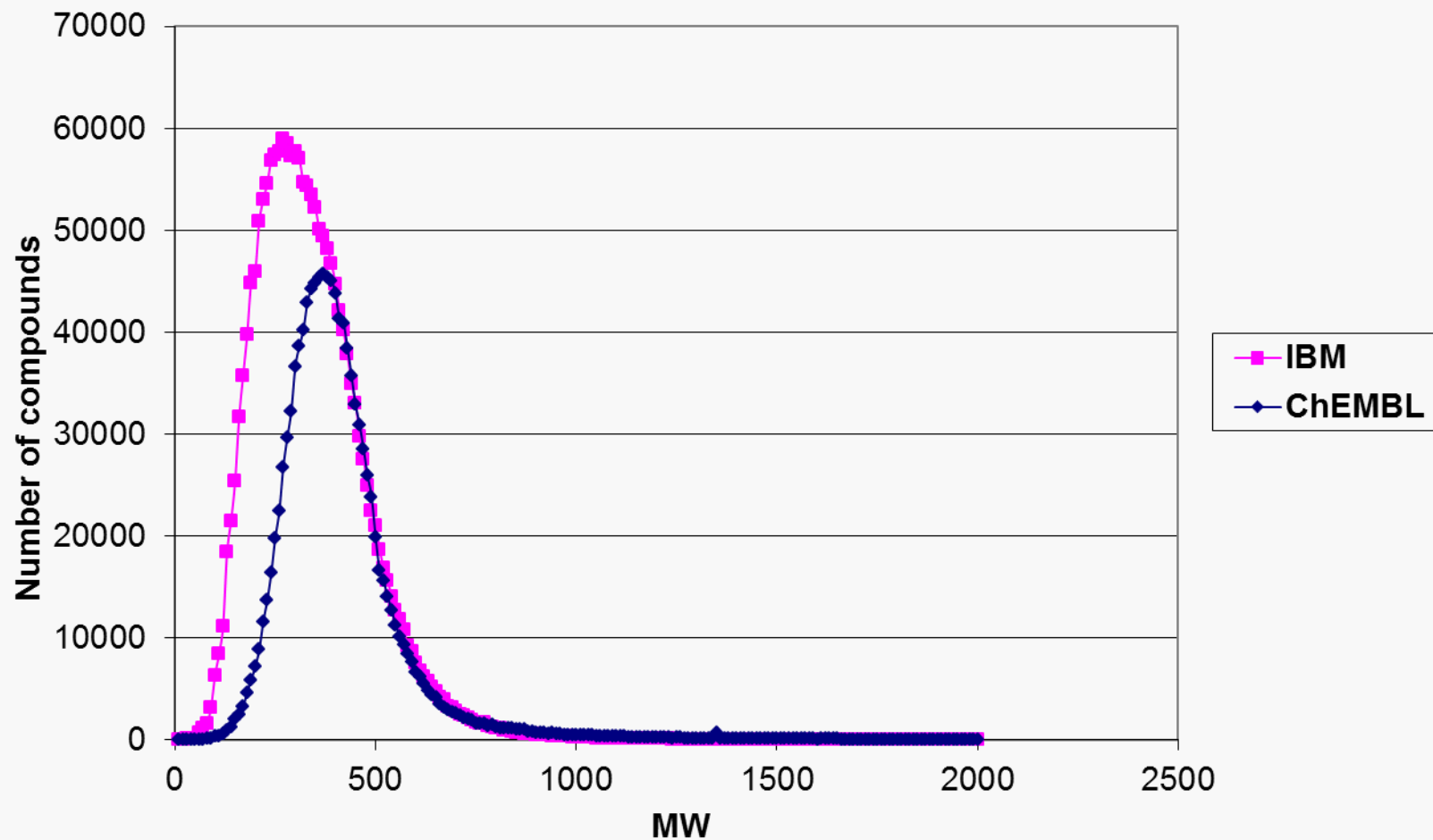
USPTO unique patent IDs: 783,175  
EPO unique patent IDs (minus kind code): 399,276 (535,996 with kind code)  
WIPO unique patent IDs (minus kind code): 186,899 (186,907 with kind code)

Most data on this slide courtesy of Evan Bolton, PubChem, NCBI, NLM, NIH



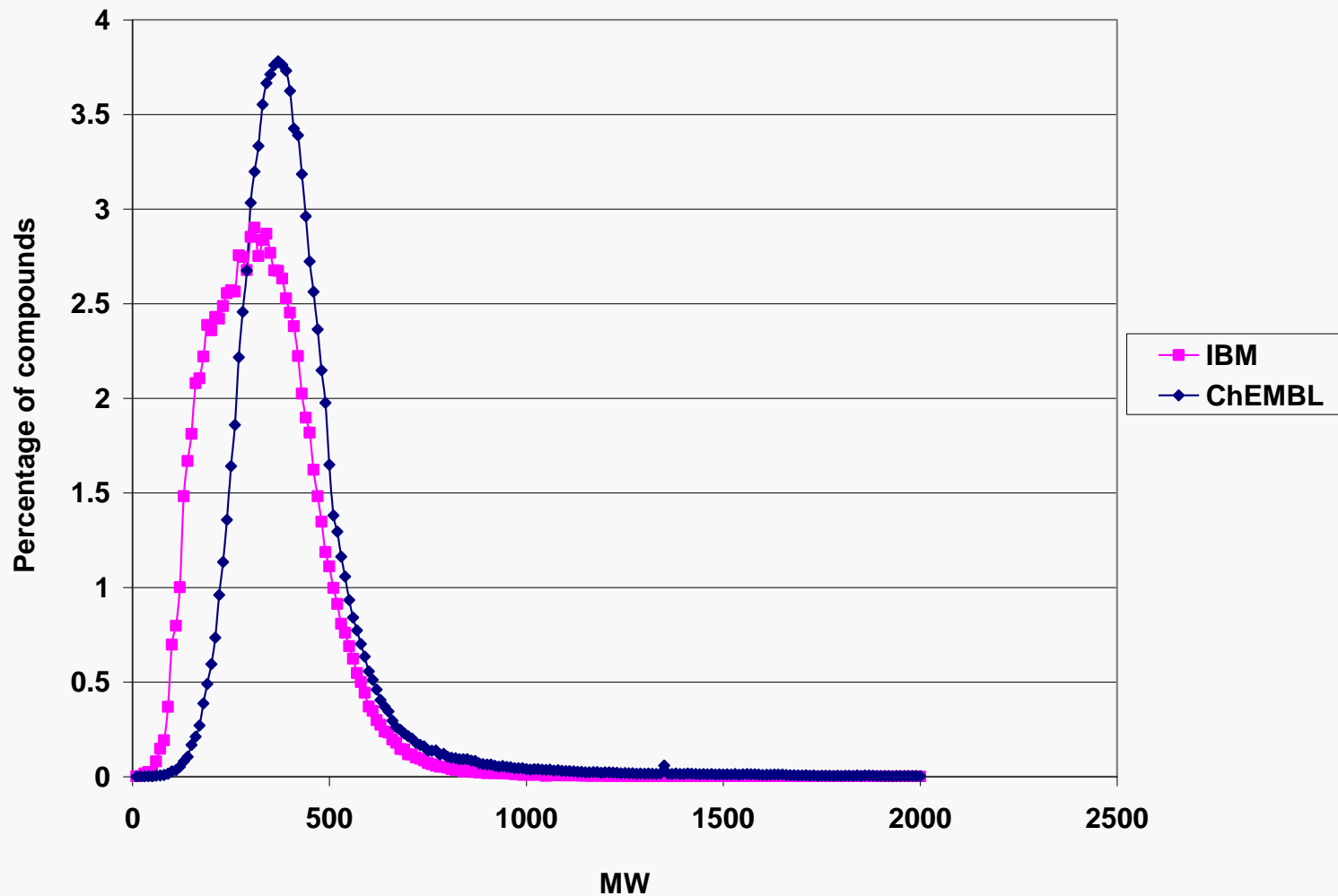
# IBM Patent Data Donation to NIH

## Content Analysis – Size of Molecules



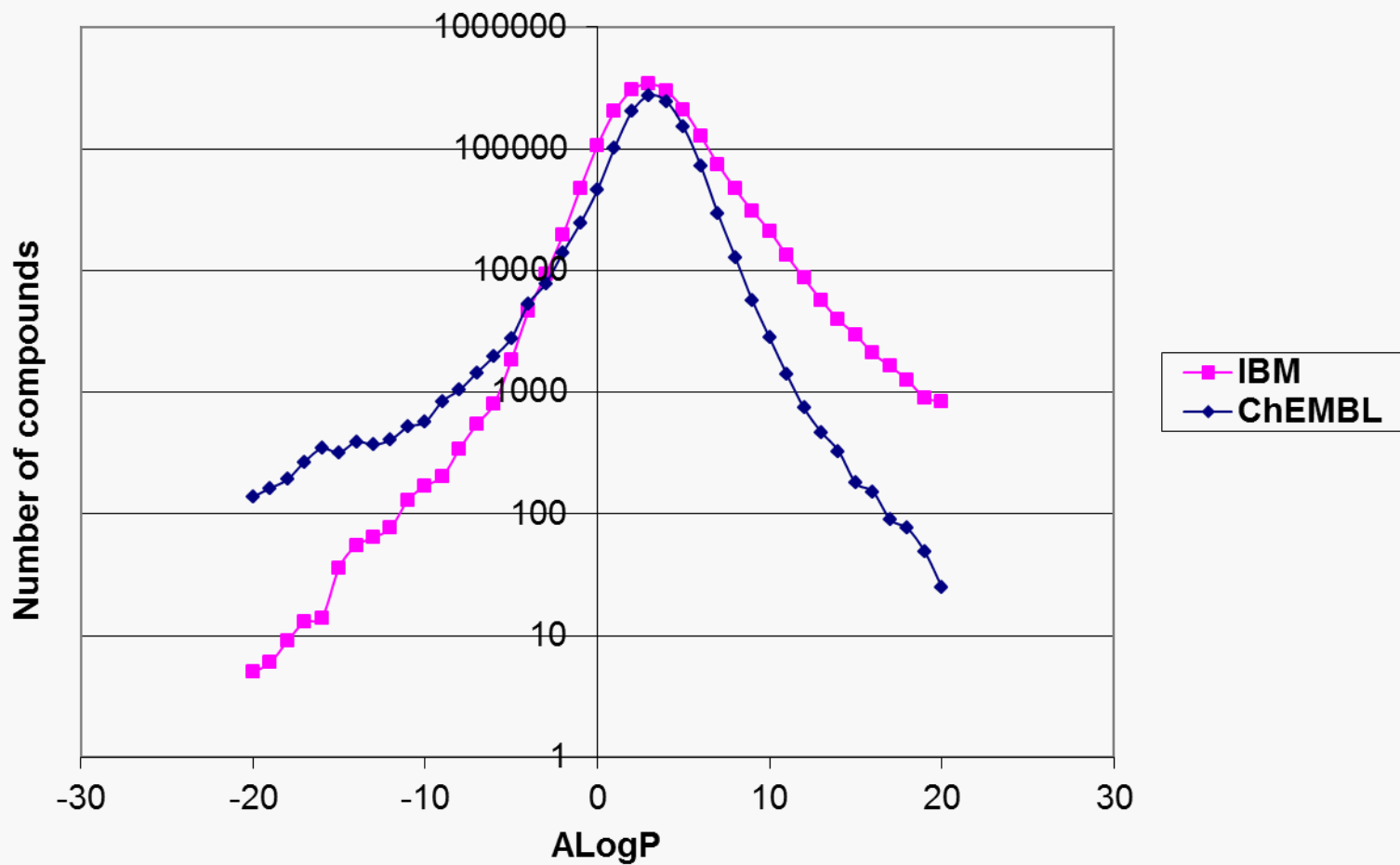
# IBM Patent Data Donation to NIH

## Content Analysis – Size of Molecules



# IBM Patent Data Donation to NIH

## Predictions – Log P



# IBM Patent Data Donation to NIH

## Other Public Chemical Patent Data & Web Sites

### SCRIPDB

JURISICA LAB  
IBM Life Sciences Discovery Center

### Search for a molecule in US Patents (2001-2011)

Search style:  
 Exact match  
 Substructure search  
 Similarity search

Search only syntheses:

From: 2001 To: 2001  
Queries with large ranges may take several minutes to complete.

Molfile:  
 No file chosen

SMARTS:

Display a selection of structures from patents containing the following text:

Chemical structure of acridine is displayed in the central window.

Data from Google's distribution of the USPTO data:  
<http://www.google.com/googlebooks/uspto-patents.html>

### SureChem Open

# IBM Patent Data Donation to NIH

## Status of Integration in NCI/CADD Web Services

- IBM data set coming from (and as processed by) PubChem in CSDB
- Will become available in new releases of our services very soon
- “Raw” IBM data on hold, pending further analysis of issues, future work by IBM/Consortium, and/or possible fix of some problems by us

# IBM Patent Data Donation to NIH

## Conclusions

IBM Donation is useful dataset

...but plenty of room for improvement:

- Stereochemistry, especially E/Z double bonds
- High occurrence rates of small, “meaningless,” molecules (benzene...)
- Distinction between Composition of Matter Patents and Use Patents
- Handling/representation of Markush-type structures
- More recent patents

# Acknowledgments

## **NCI/CADD Team**

Markus Sitzmann

Alexey Zakharov

Laura Guasch

Megan Peach

## **Xemistry GmbH**

Wolf-Dietrich Ihlenfeldt

## **ChemNavigator/Sigma-Aldrich**

Scott Hutton

Tad Hurst

## **IBM**

Stephen Boyer

Thomas Griffin

Alfredo Alba

Ying Chen

## **Pfizer**

Christopher Kibby

## **Dupont**

John Kinney

## **PubChem**

Evan Bolton