# Current Trends in Cheminformatics

Dr. Wendy A. Warr

http://www.warr.com

# Overview

- Literature analysis
- Hardware and infrastructure
- Web 2.0
- The Semantic Web
- Text mining
- Combichem
- Evaluation of scientific methods
- Conclusions

# Bibliometric Analysis

Willett, P. A bibliometric analysis of the literature of chemoinformatics. *Aslib Proc.* **2008**, *60*(1), 4-17.

# Citation Analysis - *Caveat*

- “Must cites”
- Reviews and perspectives
- Bias toward older papers
- Etc.

# Trend Toward Applications

Willett, P. A bibliometric analysis of the *Journal of Molecular Graphics and Modelling. J. Mol. Graphics Modell.* **2007**, *26*(3), 602-606.

# Highly Cited

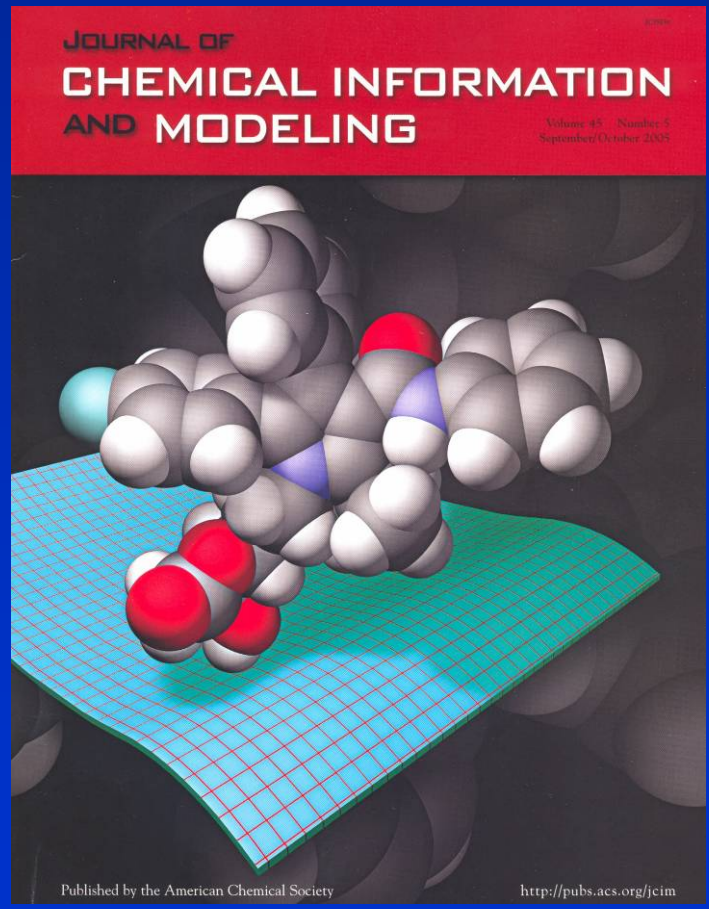
1. Koradi *et al*. MOLMOL: a program for display and analysis of of macromolecular structures. *J. Mol. Graphics Modell.* **1996**, *14*, ***51-55***. (3298 citations)

10. Dunker *et al*. Intrinsically disordered protein. *J. Mol. Graphics Modell.* **2001**, *19*, ***26-59***. (242)

11. Golbraikh, A.; Tropsha, A. Beware of $q^2$! *J. Mol. Graphics Modell.* **2002**, *20*, ***269-276***. (172)

# *JCIM*

*J. Chem. Inf. Comput. Sci.* (*JCICS*)
is now
*J. Chem. Inf. Model.* (*JCIM*)



# Most Cited JCICS-JCIM Paper

Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The development of versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187-204.

# Most Cited *JCICS-JCIM* since 1996

1. Fletcher, D. A.; McMeeking, R. F.; Parkin, D. The United Kingdom chemical database service. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 746-749. **HIGHER**
2. Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983-996. **HIGHER**
3. Brown, R. D.; Martin, Y. C. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572-584.

**(compare 2005 rank)**

# Most Cited *JCICS-JCIM* May 2008 *versus* July 2005

4. Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1308-1315. **HIGHER**
5. Hann, M. M.; Leach, A. R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41,* 856-864. **HIGHER**
6. Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726-735. **HIGHER**

# Most Cited *JCICS-JCIM* May 2008 *versus* July 2005

7. Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1-9. **HIGHER**
8. Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. Estimation of Molecular Linear Free Energy Relation Descriptors Using a Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1999**; *39*, 835-845. **HIGHER**
9. Irwin, J. J.; Shoichet, B. K. ZINC - A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177-182. **NEW**

# Most Cited *JCICS-JCIM* May 2008 *versus* July 2005

10. Pearlman, R. S.; Smith, K. M. Metric Validation and the receptor-relevant subspace concept**.** *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28-35. **HIGHER**
11. Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868-873. **NEW**
12. Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118-127. **NEW**
13. Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1-12. **NEW**

# Most Cited *JCICS-JCIM* May 2008 *versus* July 2005

14. Randic, M.; Vracko, M.; Nandy, A.; Basak, S. C. On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235-1244. **NEW**
15. Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally diverse quantitative structure-property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1-18. **NEW**
16. Wang, R.; Fu, Y.; Lai, L. A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comput. Sci.* **1997**, 37, 615-621. **NEW**

# Most Cited *JCICS-JCIM* May 2008 *versus* July 2005

17. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP-retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511-522. **NEW**
18. Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017-1026. **NEW**

# Most Cited *JCICS-JCIM* May 2008 *versus* July 2005

19. Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J.; Goddette, D. W.; Heritage, T. W.; Sheehan, D. M. Evaluation of quantitative structure-activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 669-677. **NEW**
20. Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 773-777. **NEW**

# Fearless Predictions…

"Where a calculator on the Eniac is equipped with 18,000 vacuum tubes and weighs 30 tons, computers in the future may have only 1,000 vacuum tubes and perhaps weigh 1.5 tons."

Popular Mechanics, March 1949

# 1984 VAX 11/750

- Clock speed 6 MHz
- 2 Mb memory
- 134 Mb fixed disk
- Two 67 Mb exchangeable disk drives
- Shared peripherals
- £100,000 (1984 price)

# 1992 PC

- IBM-compatible, MS-DOS 2.1 or higher machine
- 640K memory
- a floppy drive (3.5-in. 1.44MB or 5.25 in. 1.2MB)
- hard disk with at least 6MB available
- EGA or Hercules display

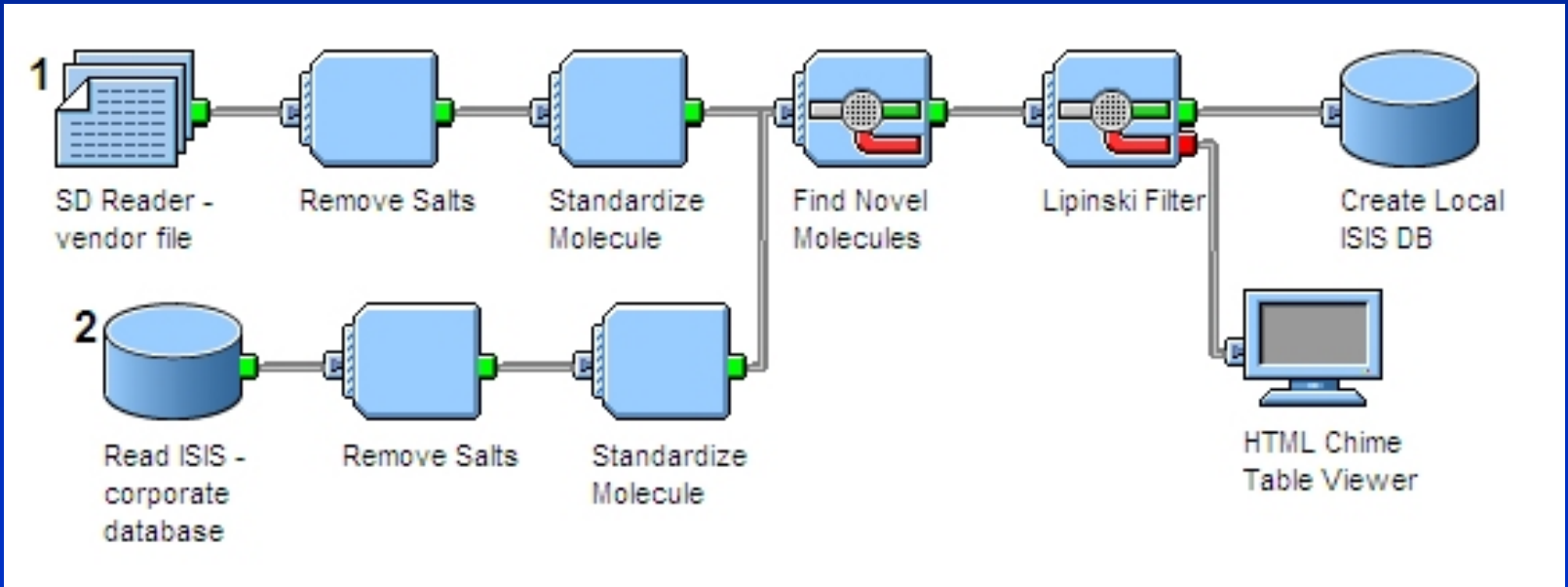
http://www.warr.com/15years.html
Bretherick's Reactive Chemical Hazards Database Version 1.00

# Infrastructure 2008

- Cyberinsfrastructure/e-science
- Grid computing
- Web 2.0
- The Semantic Web

# A Mobile CAS Service

# SciTegic Pipeline Pilot



# Workflow and Pipelining

- Kepler
- Taverna
- SOMA
- Pipeline Pilot
- InforSense
- KNIME

http://www.qsarworld.com/qsar-workflow1.php

# Data Cartridges

- Symyx Direct
- Accelrys Accord Chemistry Cartridge
- Daylight DayCart
- ChemAxon JChem Cartridge
- CambridgeSoft Oracle Cartridge
- InfoChem IC Cartridge
- Tripos AUSPYX
- IDBS ChemXtra
- Dotmatics Pinpoint
- Digital Chemistry Torus

Source: DeltaSoft Inc.

# Social Software

- Instant messaging
  - MSN Messenger, Yahoo Messenger, AOL Instant Messenger, Skype
- Text chat
  - Internet Relay Chat
- Internet forums
- Blogs
- Wikis
  - Wikipedia

# Social Software

- Blog search engines
  - Google blogsearch, Technorati
- Web feeds
  - RSS
    - Rich Site Summary/Really Simple Syndication
  - Atom
- Podcasts
  - Can be syndicated, subscribed to etc.

# Social Software

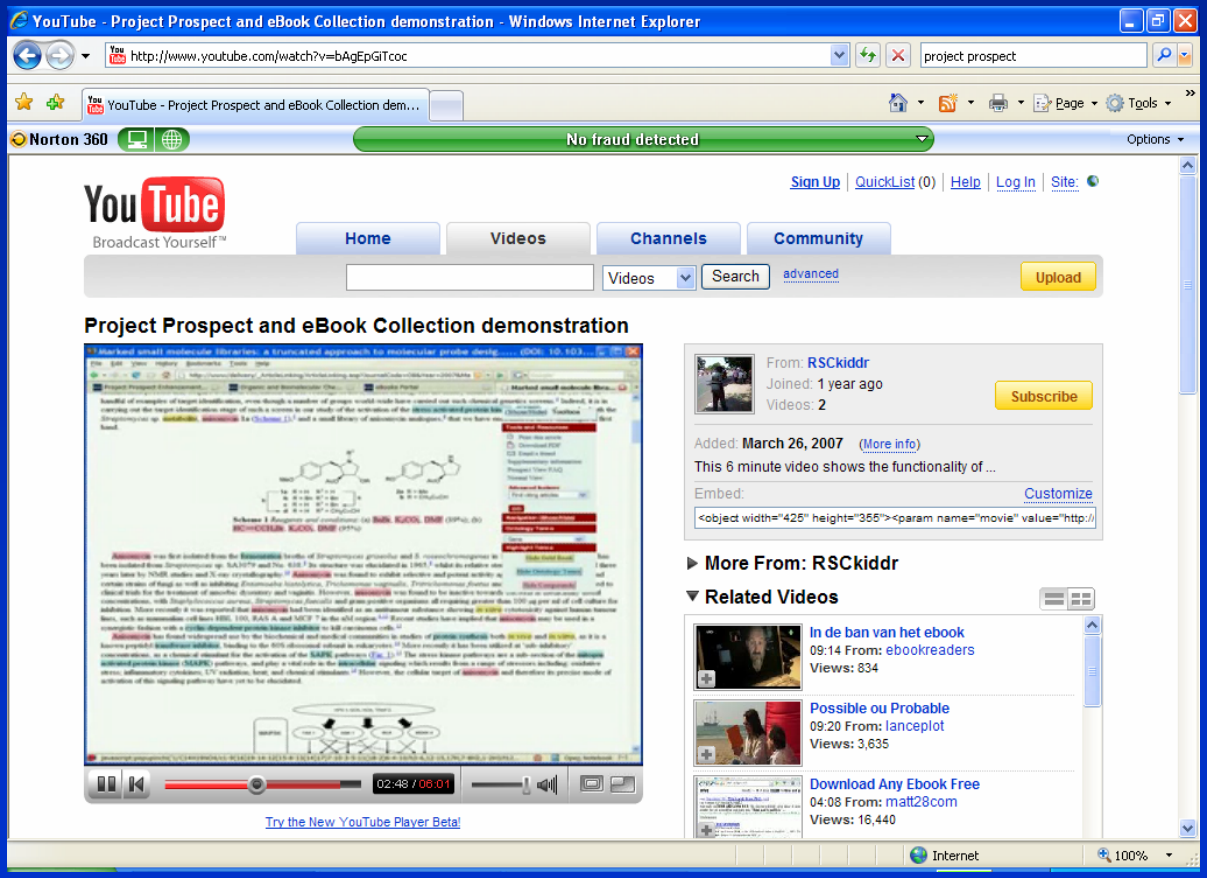
- Social network services
  - Facebook, MySpace, Flickr, YouTube
  - LinkedIn, Ryze, XING
- Social guides
  - WikiTravel, TripAdvisor
- Social bookmarking
  - Tagging, tag clouds, folksonomy
- Social citations
  - CiteULike, Connotea, BibSonomy
- Virtual worlds

# Most Popular Web Sites

1. Yahoo
2. Google
3. MSN
4. YouTube
5. Live
6. MySpace
7. Orkut
8. Facebook
9. Wikipedia
10. Hi5

English language sites, Alexa, October 7, 2007

# RSC Project Prospect



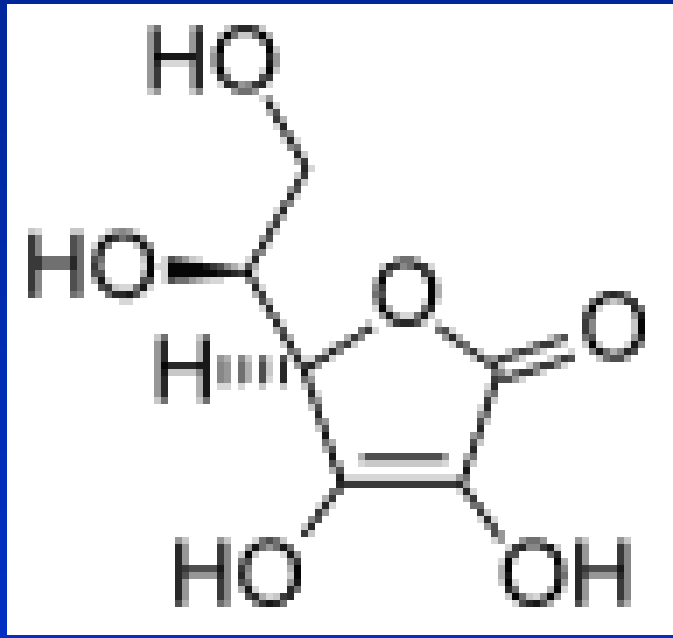

Wendy Warr & Associates

# RSC Project Prospect

# InChI Goal

- Provide a unique string representing a chemical substance of known structure
  - independent of specific depiction
  - derived from conventional connection table
  - freely available
  - extensible

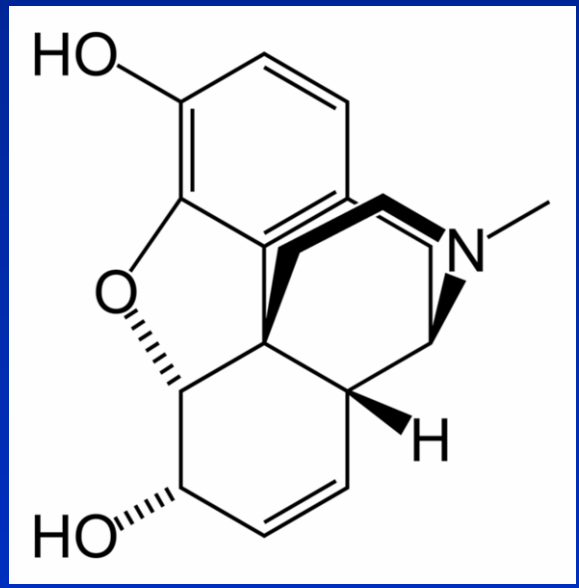
# InChI Example




L-ascorbic acid

InChI=1/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-10H,1H2/t2-,5+/m0/s1

# InChIKey




InChI=1/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11-,13-,16-,17-/m0/s1

InChIKey = BQJCRHHNABKAKU-XKUOQXLYBY

Wendy Warr & Associates

# Words, words, words

- The Lord's Prayer uses 56 words
- The Ten Commandments, 297 words
- The American Declaration of Independence,  300 words
- The EEC Directive on the import of caramel and caramel products uses 26,911 words

# Information Extraction Technologies in Chemistry

http://www.infonortics.com/chemical/ch07/slides/hofmann.pdf

Martin Hofmann-Apitius

# Chemical Named Entity Recognition (NER)

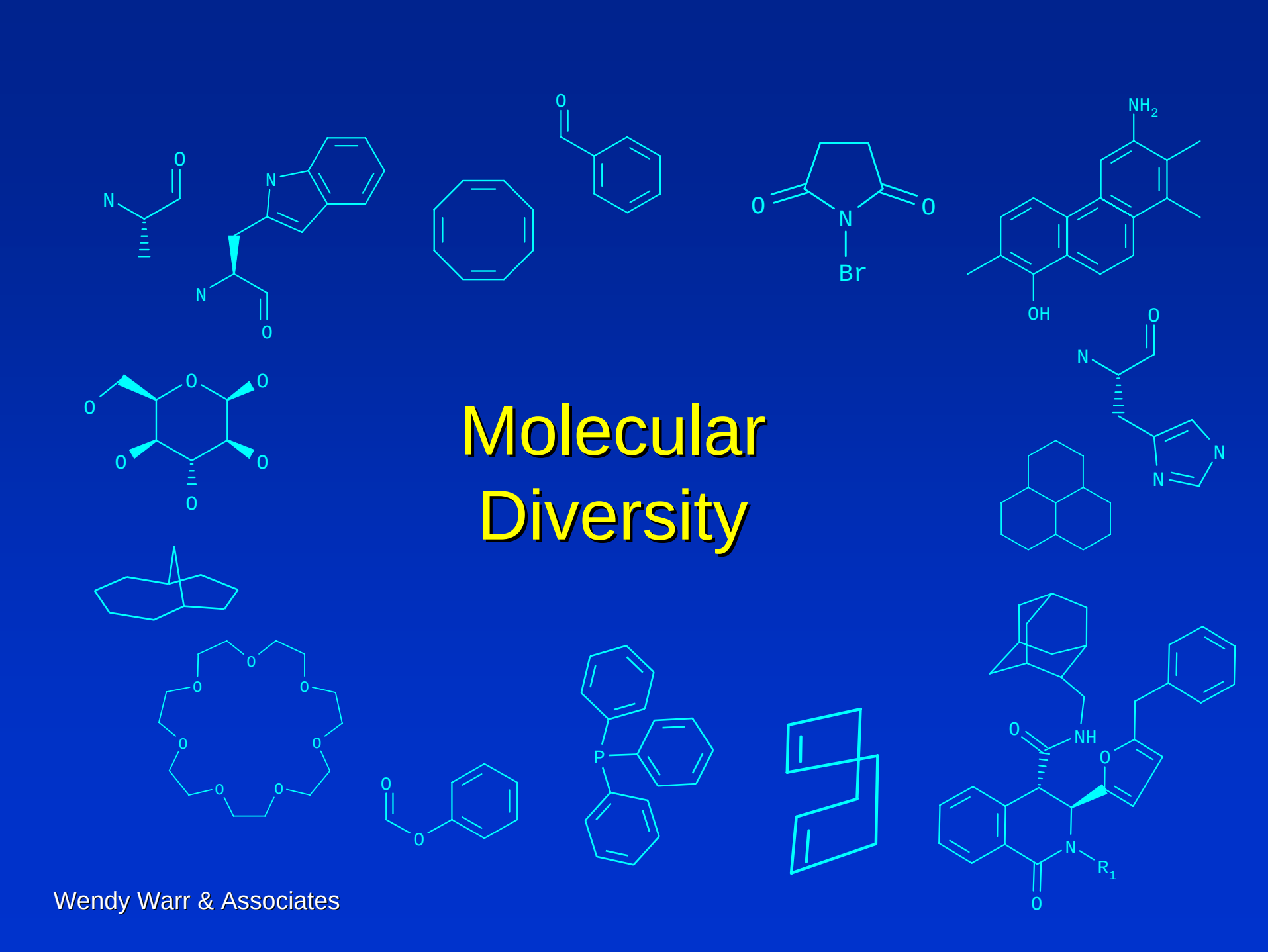
- TEMIS
- InforSense
- SciTegic
- Open Source Chemical Analysis Routines (OSCAR)
- IBM Chemical Annotator

# Name to Structure

- CambridgeSoft Name=Struct
- ACD/Labs Name to Structure
- OpenEye Lexichem
- InfoChem ICN2S

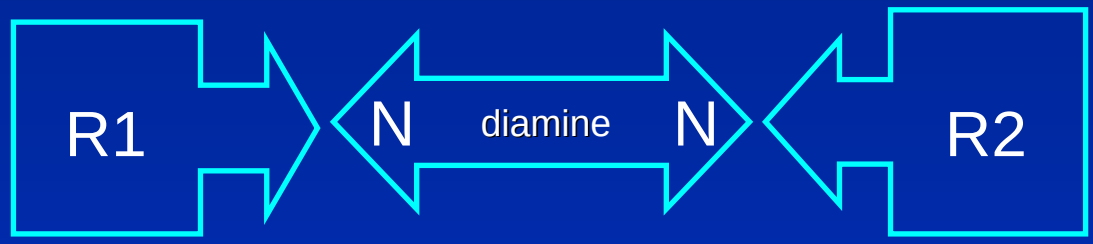
# Structures from Images

- Kekulé (*JCICS* 1992)
- CLiDE (Peter Johnson Key Module)
- OSRA (NIH Cactus)
- chemoCR (Fraunhofer/InfoChem)




Wendy Warr & Associates

# How Big is "Readily Accessible" Chemistry Space?




68934 *x* 4145 *x* 68934 = $2.0 \times 10^{13}$

R1 = R2 = acylating reagents, cleanly displaceable halides, amines (to ureas via ClC(=O)Cl), reductive amination substrates

***Counts** refer to Available Chemicals Directory (ACD)*

Source: Tripos (as of 1995?)

$2.0 \times 10^{13}$

= 40 million pharma "decks"

= 1,000,000 x compounds ever reported (Chemical Abstracts)

= 60,000 years of screening @ rate of one million per day

# Library Design

- Random screening is too expensive
- Its hit rate is low
- False positives may be a problem
- HTS consumes expensive compounds
- There are too many possible compounds
- How to choose those most likely to be hits?

# Selection

$10^{180}$ possible drugs...$10^{18}$ likely drugs...

$10^{7}$ known compounds...$10^{6}$ commercially available...$10^{6}$ in corporate databases...

$10^{4}$ in drug databases...$10^{3}$ commercial drugs...

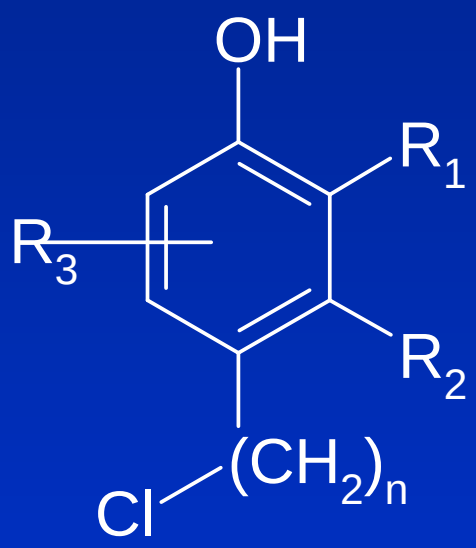
$10^{2}$ profitable drugs

Source: David Weininger

# Instead of blockbusters and “me-toos”

- Translational research
- Pharmacogenomics
- Biomarkers
- Personalized medicine

# Generic Structure



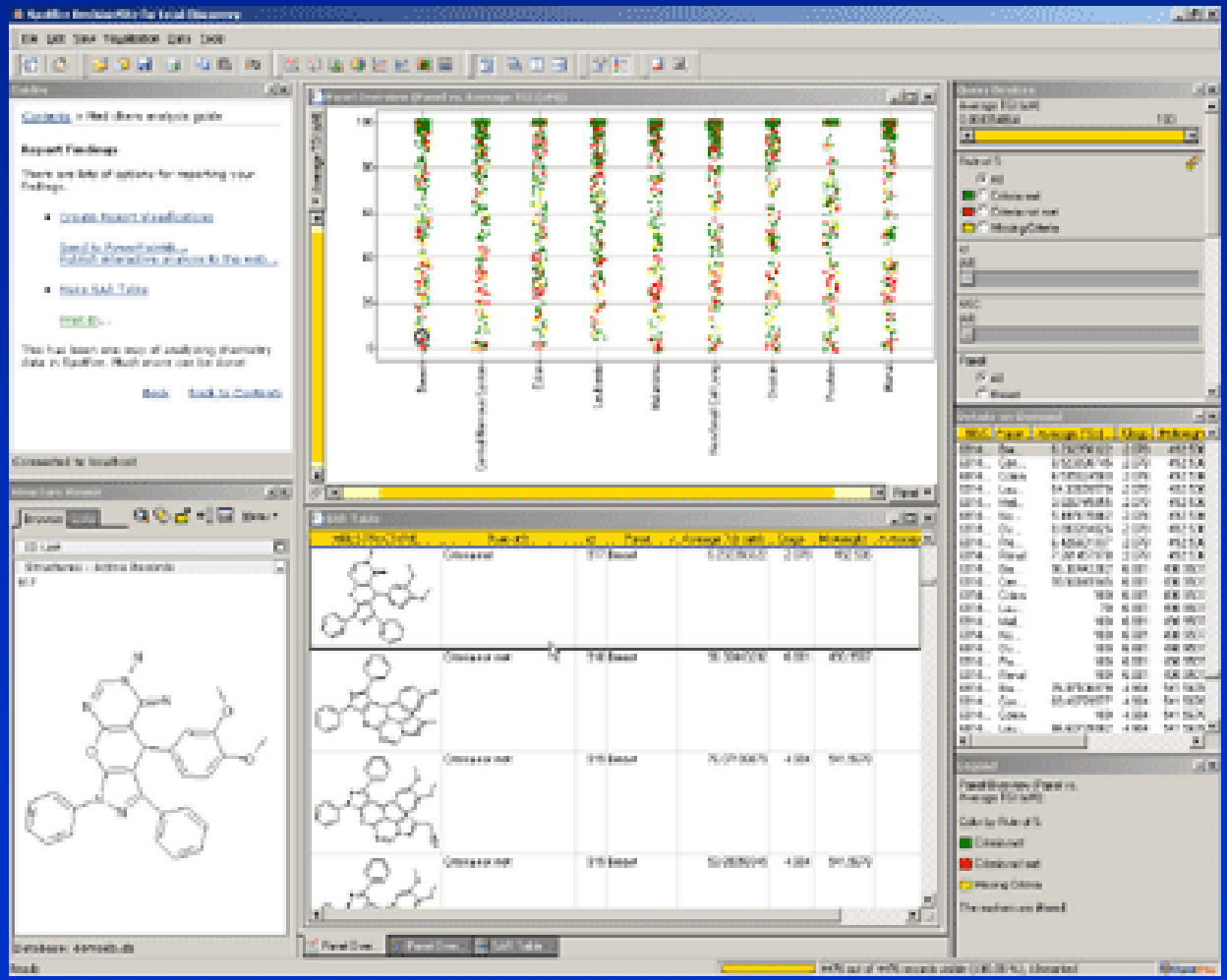

- substituent variation
  $R_1$ is methyl or ethyl
- homology variation
  $R_2$ is alkyl
- position variation
  $R_3$ is amino
- frequency variation
  n is 1-3

Source: John Barnard

# Compound characterisation


- Barreiro, G.; Guimaraes, C. R. W.; Tubert-Brohman, I.; Lyons, T. M.; Tirado-Rives, J.; Jorgensen, W. L. Search for non-nucleoside inhibitors of HIV-1 reverse transcriptase using chemical similarity, molecular docking, and MM-GB/SA scoring. *J. Chem. Inf. Model.* **2007**, *47*, 2416-2428.
- Barreiro, G.; Kim, J. T.; Guimaraes, C. R. W.; Bailey, C. M.; Domaoal, R. A.; Wang, L.; Anderson, K. S.; Jorgensen, W. L. From docking false-positive to active anti-HIV agent. *J. Med. Chem.* **2007**, *50*, 5324-5329.

# Data Mining and Visualization



# Spiral View for SAR Interpretation

.

Smellie, A. General purpose interactive physico-chemical property exploration. *J. Chem. Inf. Model.* **2007**, *47*, 1182-1187.

# Evaluation of Computational Methods

*J. Comput.-Aided Mol. Des.* **2008**, March/April special issue

# Open Data. Reusable Chemistry

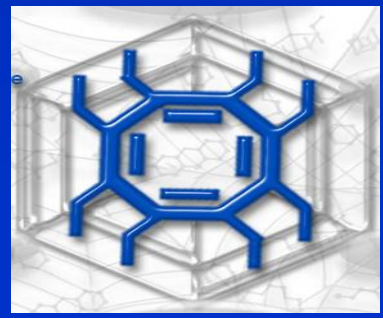

PubChem

DUD

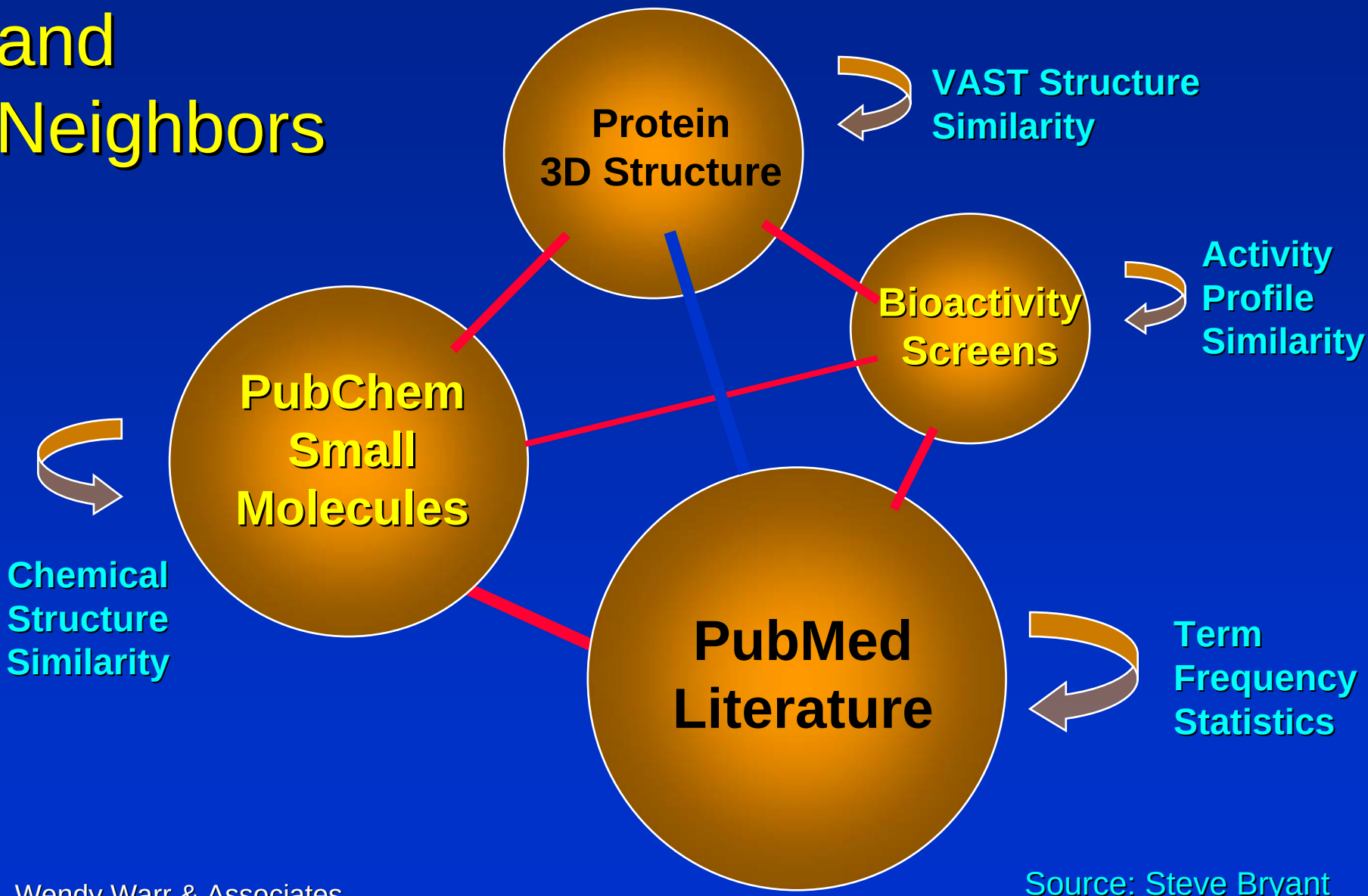
A Directory of Useful Decoys

CDD COLLABORATIVE DRUG DISCOVERY

eMolecules



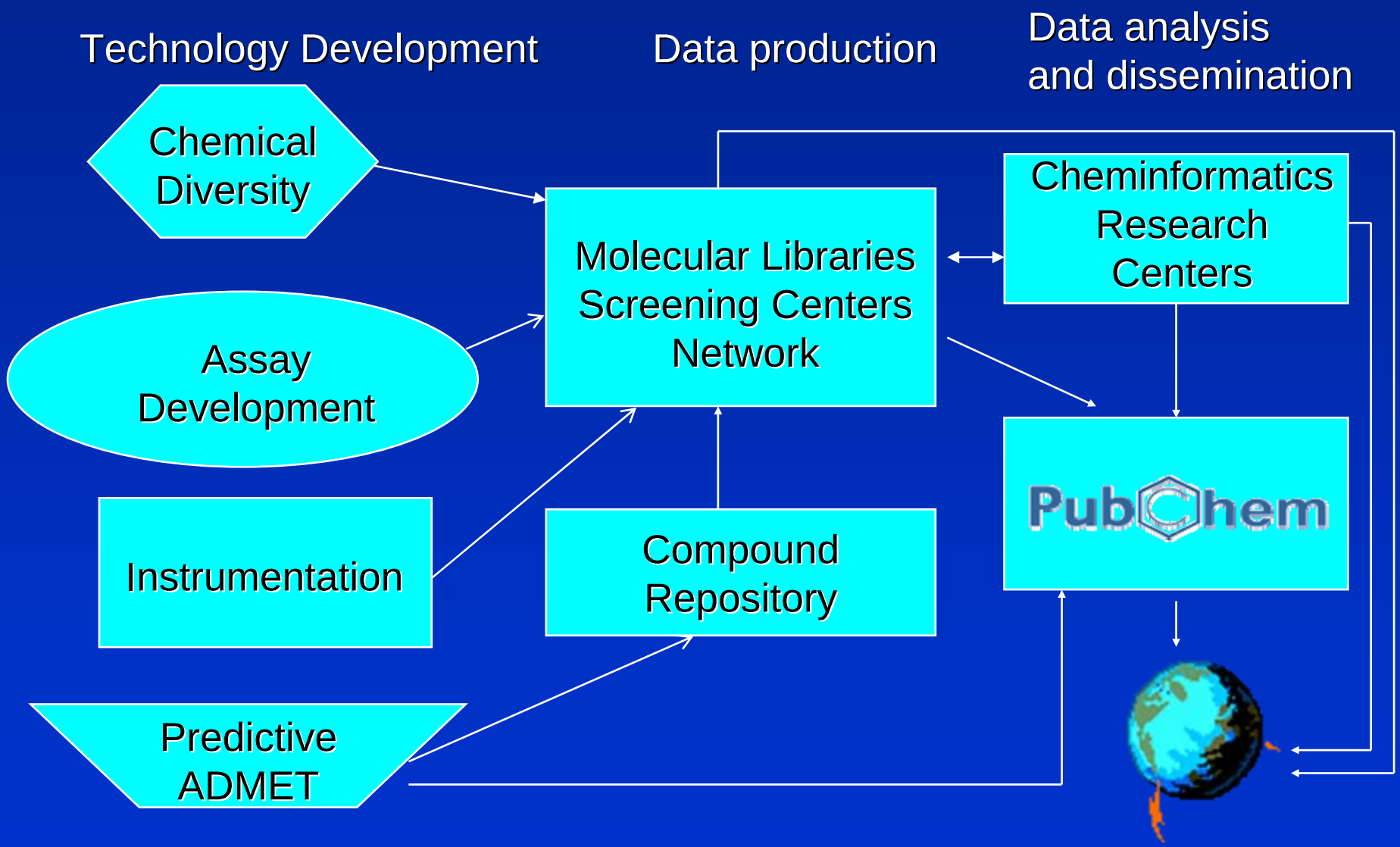
# Entrez Links and Neighbors




Wendy Warr & Associates

Source: Steve Bryant

# Molecular Libraries Roadmap




Wendy Warr & Associates

# Conclusion

The good news and the bad news…