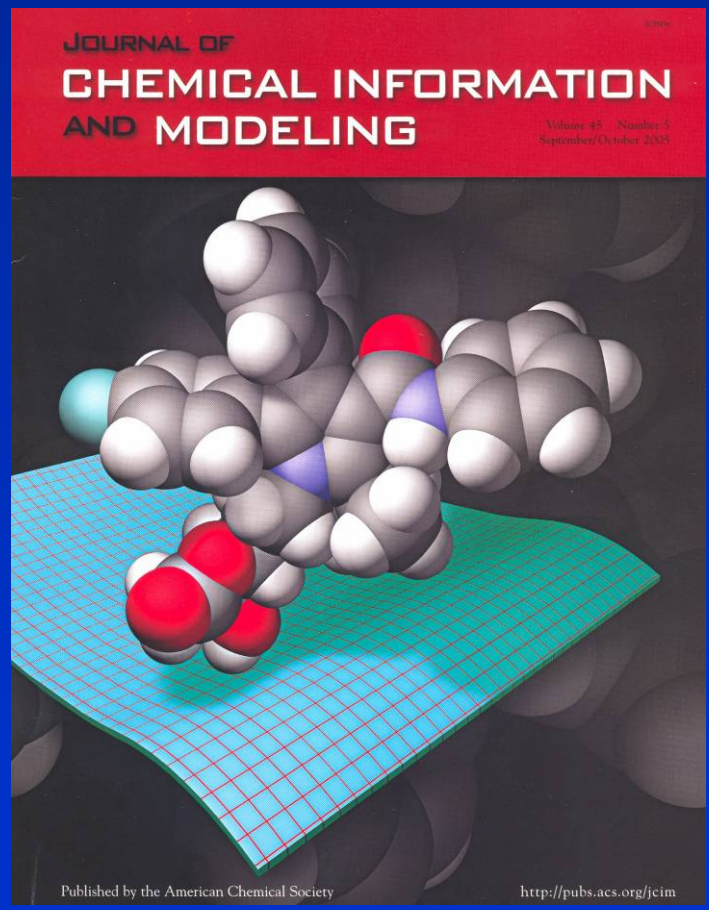
Twenty Five Years of Progress in Cheminformatics

Dr. Wendy A. Warr

http://www.warr.com

JCIM

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



The Chemist

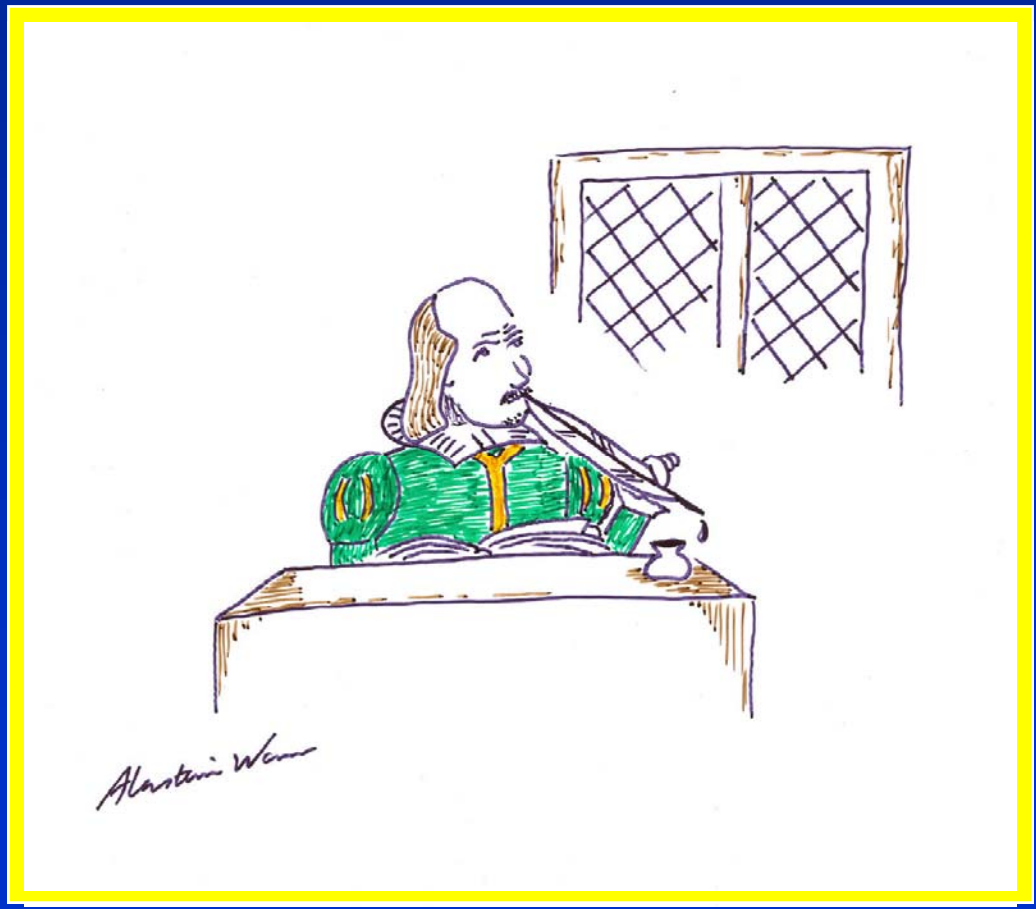
Author

Reader

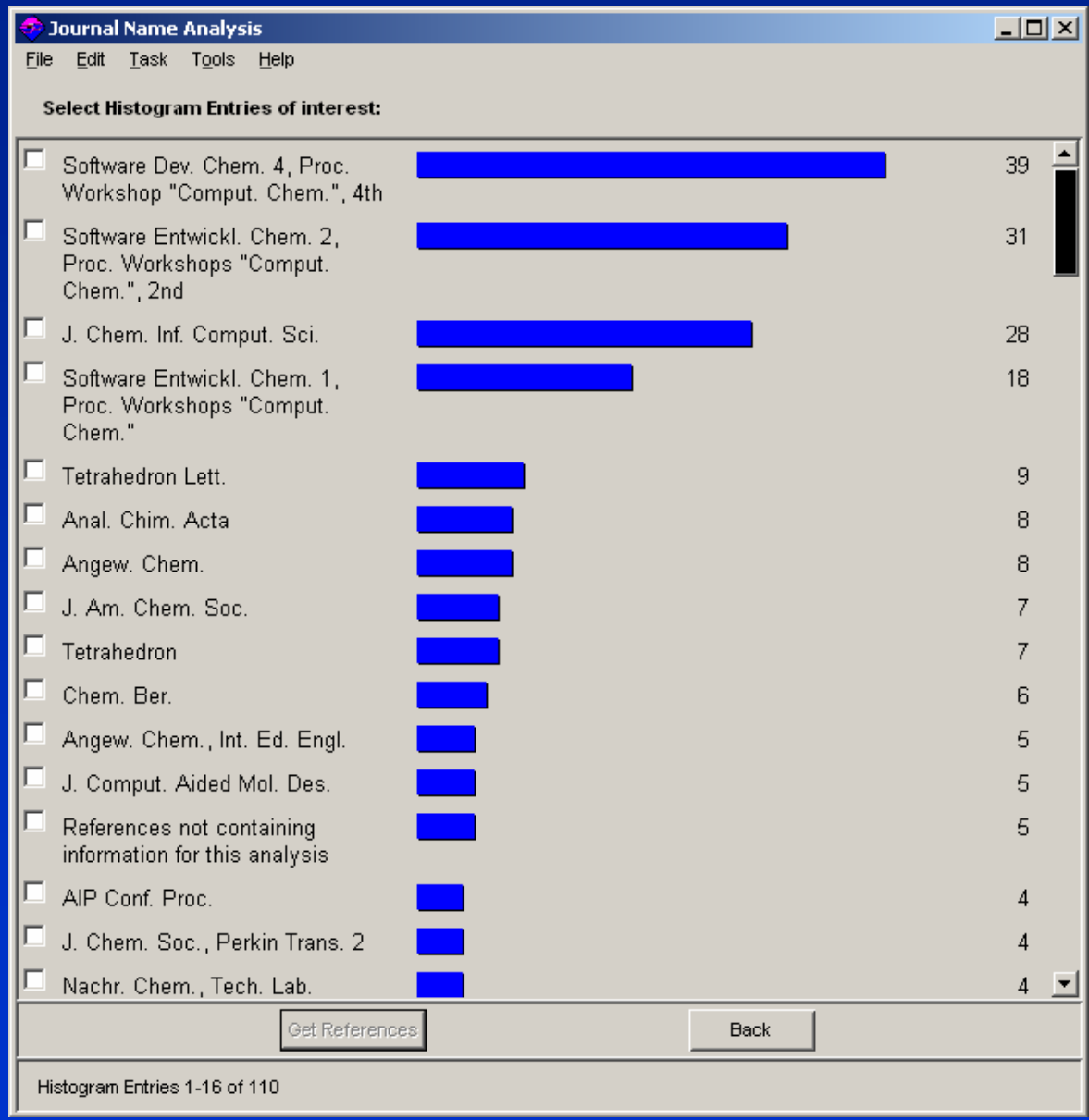
Reviewer

Editor

Our Authors



SciFinder Gasteiger



Wendy Warr & Associates

ACS Award for Computers in Chemical and Pharmaceutical Research

1986 Raymond E. Dessy
1987 W. Todd Wipke
1988 W.A. Goddard III
1989 Christie G. Enke
1990 Peter C. Jurs
1991 John A. Pople
1992 Ernest R. Davidson
1993 W. Clark Still
1994 Michael J.S. Dewar
1995 Peter A. Kollman
1996 Norman L. Allinger
1997 Harold A. Scheraga
1998 William L. Jorgensen
1999 Corwin H. Hansch
2000 Donald G. Truhlar
2001 Martin Karplus
2002 Irwin D. Kuntz
2003 Kendall N. Houk
2004 W. Graham Richards
2005 Peter Willett
2006 Johann Gasteiger

ACS Division of Chemical Information Herman Skolnik Award

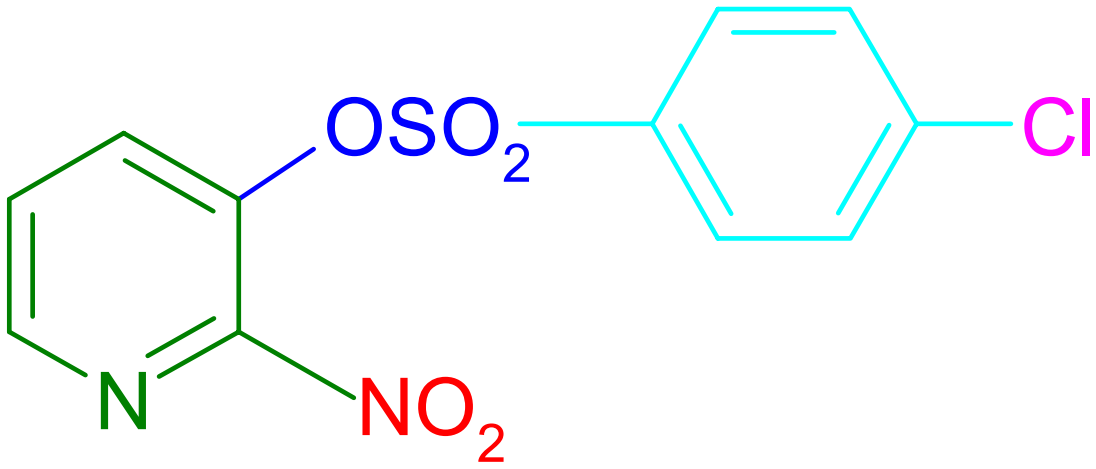
Herman Skolnik (1976)
Eugene Garfield (1977)
Fred A. Tate (1978)
William J. Wiswesser (1979)
Ben H. Weil (1981)
*Robert Fugmann (1982)
Russell J. Rowlett, Jr. (1983)
Montagu Hyams (1984)
Dale B. Baker (1986)
William Theilheimer (1987)
David R. Lide, Jr. (1988)
Michael F. Lynch (1989)
Stuart A. Marson (1989)
*Ernst Meyer (1990)
W. Todd Wipke (1991)
Jaques-Emile Dubois (1992)
Peter Willett (1993)
Alexandru T. Balaban (1994)
*Reiner Luckenbach (1995)
*Clemens Jochum (1995)
Milan Randic (1996)
*Johann Gasteiger (1997)
Gary Wiggins (1998)
Stuart Kaback (1999)
Stephen R. Heller (2000)
G.W.A. (Bill) Milne (2000)
Guenter Grethe (2001)
Peter Norton (2002)
Frank Allen (2003)
Peter Johnson (2004)
Lorrin Garson (2005)
Hugo Kubinyi (2006)

ACS Division of Chemical Information Herman Skolnik Award

- Robert Fugmann (1982)
- Ernst Meyer (1990)
- Reiner Luckenbach (1995)
- Clemens Jochum (1995)
- Johann Gasteiger (1997)

- Guenter Grethe (2001)
- Hugo Kubinyi (2006)

Wiswesser Line Notation



T6NJ B NW C OSW R D G

CROSSBOW

Computerized
Retrieval of
StructureS
Based
On
Wiswesser

Structure Representation

- Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36.
- Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97-101.

Structure Representation

- Morgan, H. L. The generation of a unique machine description for chemical structures - a technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107-113.
- Wipke, W. T.; Dyott, T. M. Stereochemically unique naming algorithm. *J. Am. Chem. Soc.* **1974**, *96*(15), 4834-4842.

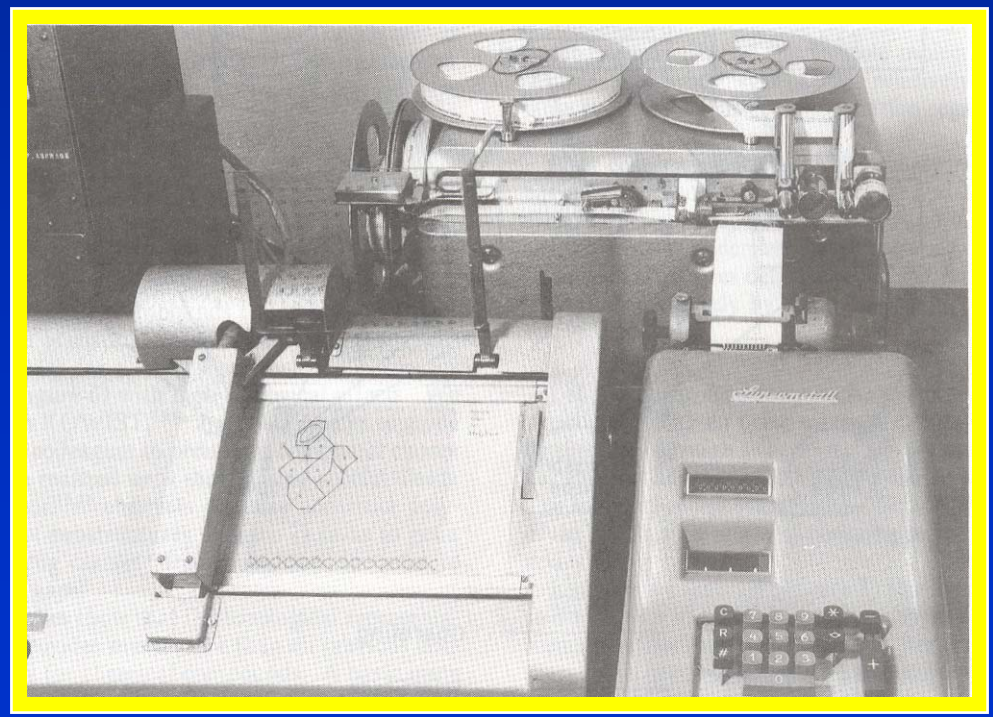
Structure Representation

- Canonical numbering for InChI: modified from McKay, B. D. Practical graph isomorphism. *Congressus Numerantium* **1981**, *30*, 45–87.

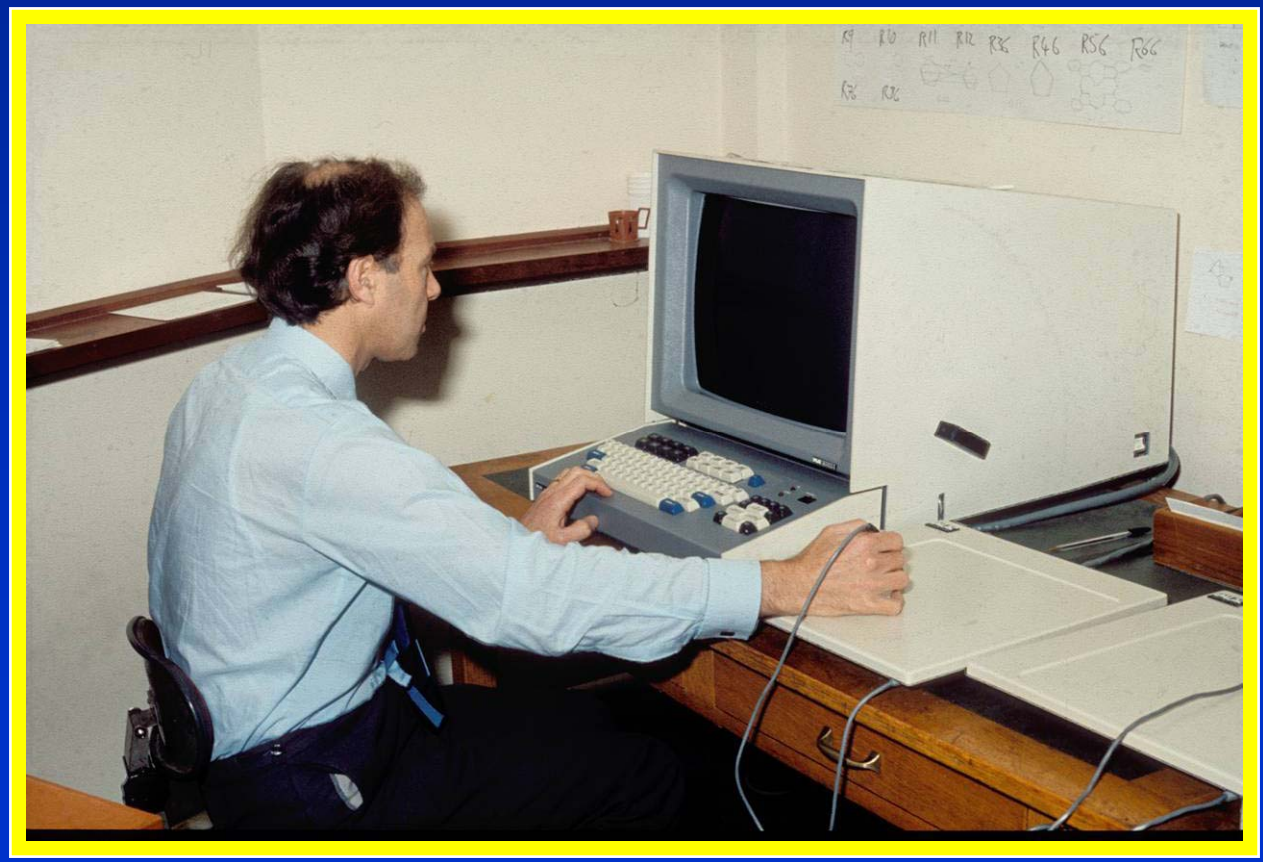
Substructure Searching

- NIH/EPA Chemical Information System
- DARC
- CAS ONLINE
- MACCS/ISIS

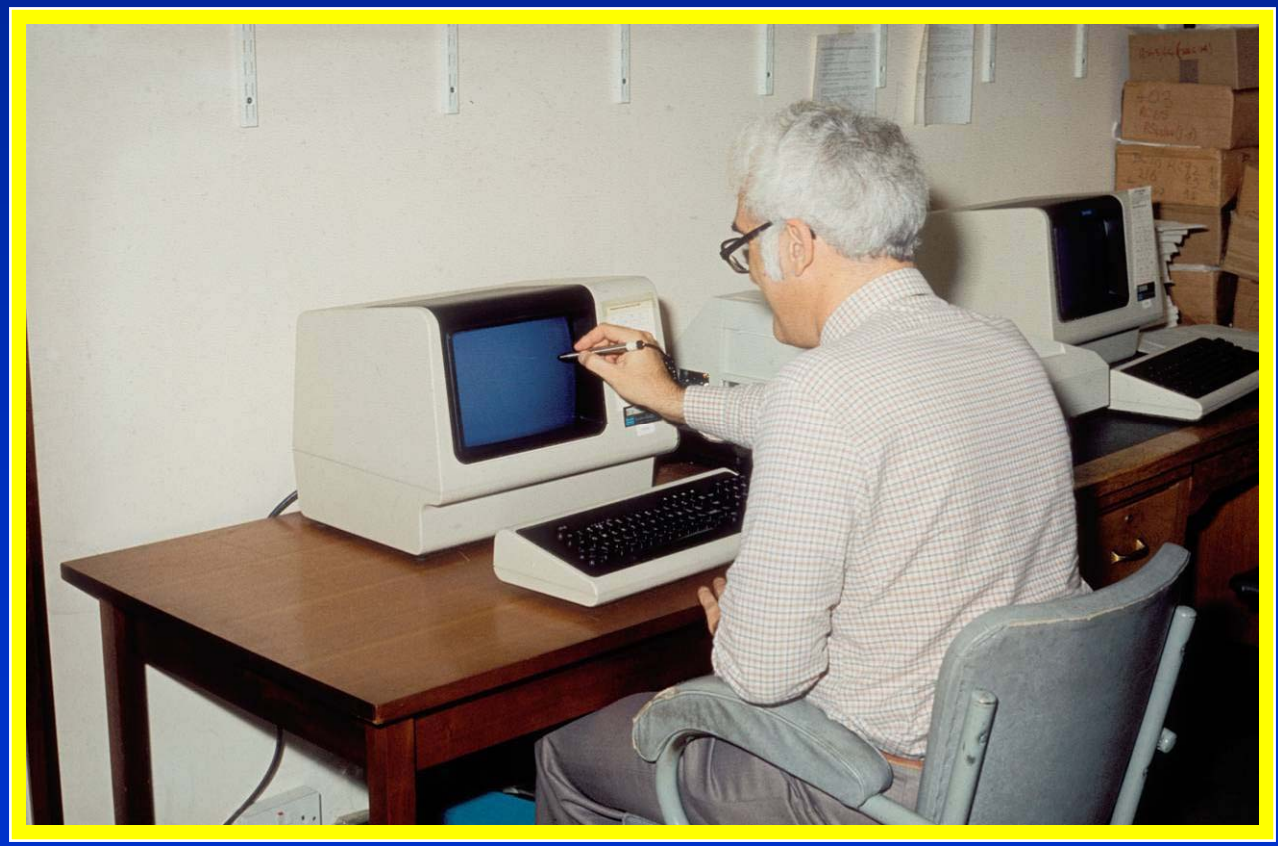
Meyer's Formula Reading Machine



Imlac Terminal



VT640 Terminal



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)

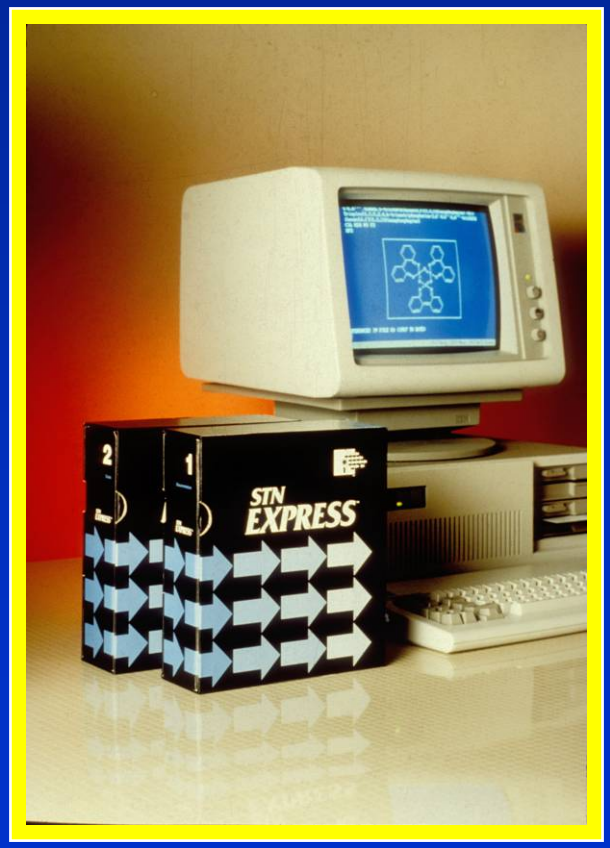
Windows

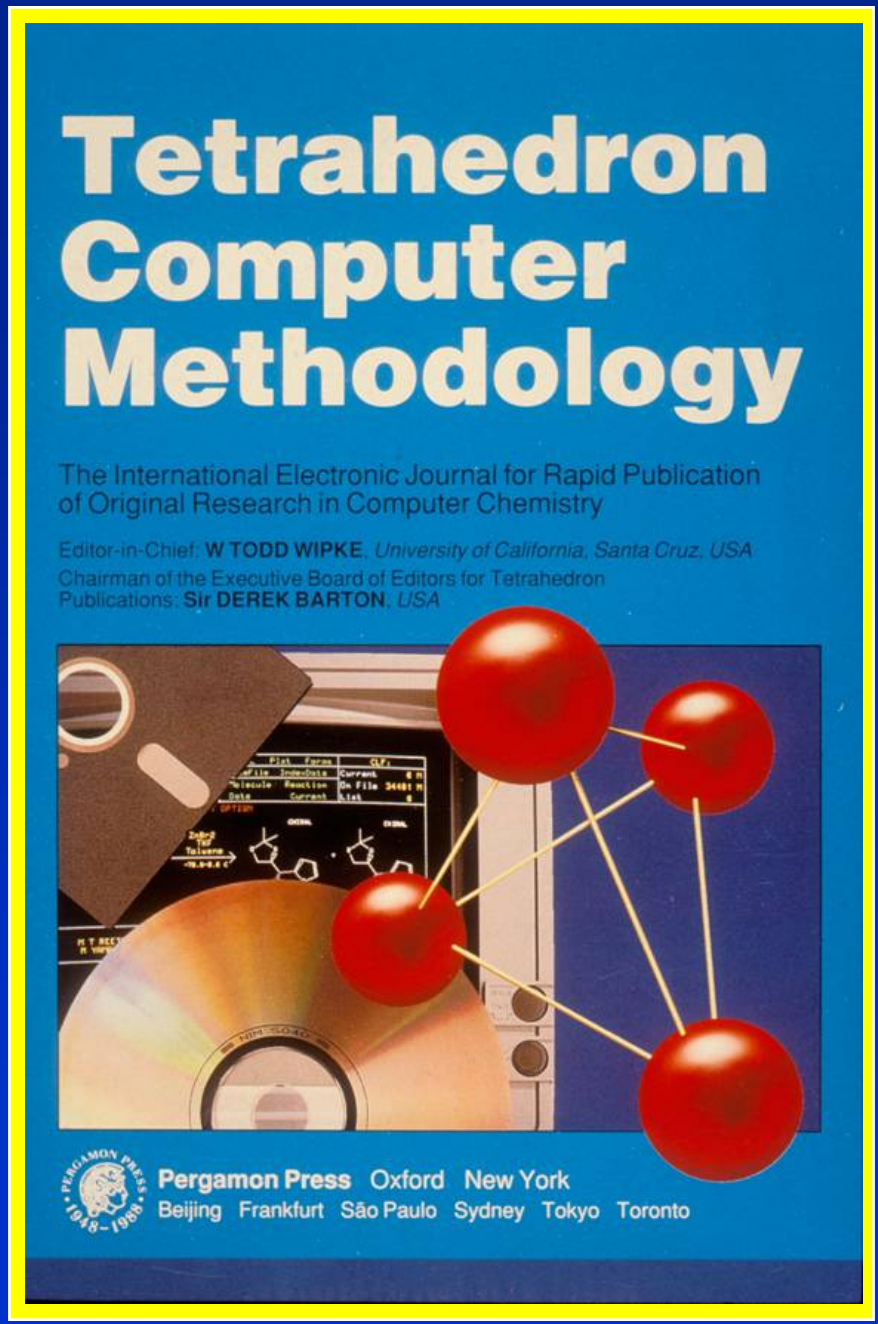
Icons

Mice

Pointers

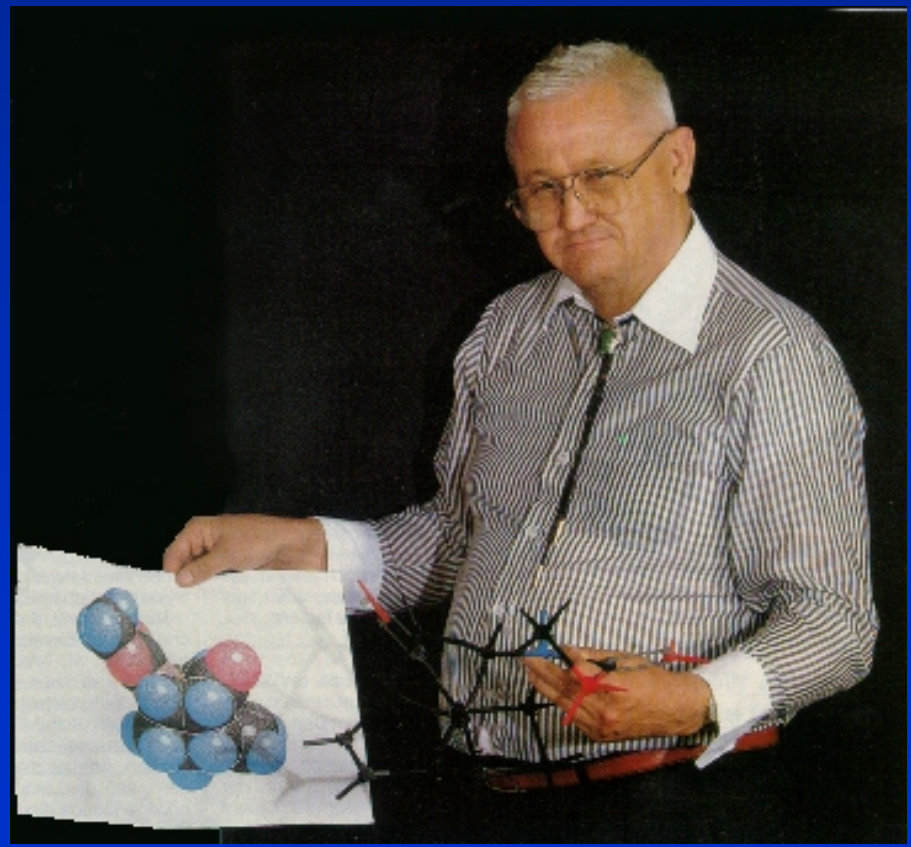
Microcomputers





Wendy Warr & Associates

Ivar Ugi
1930 - 2005



CAOS

- Corey, E.J.; Wipke, W.T. Computer-assisted Design of Complex Organic Syntheses. *Science*, **1969**, *166* (*3902*), 178-192.
- Dugundji, J.; Ugi, I. Algebraic Model of Constitutional Chemistry as a basis for Chemical Computer Programs. *Top. Curr. Chem.* **1973**, *39*, 19-64.
- Bauer, J.; Herges, R.; Fontain, E.; Ugi, I. IGOR and Computer Assisted Innovation in Chemistry. Chimia, 1985, 39, 45-53.

Reaction Programs

- Reaction retrieval
- Synthetic analysis
  - Synthesis design
  - Reaction prediction
  - Mechanism elucidation

Reaction Retrieval

- REACCS/ISIS
- ORAC
- CASREACT

Synthetic Analysis

- Logic-oriented
  - IGOR
  - EROS
  - SYNGEN
  - CAMEO
  - WODCA
- Information-oriented
  - LHASA
  - SECS/CASP

Reaction Classification

InfoChem Classification Algorithm

CLASSIFY

Cambridge Structural Database

- Allen, F. H. The Cambridge Structural Database: a quarter of a million structures and rising. *Acta Cryst. Section B* **2002**, *58*, 380-388.
- 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.

3D Structures

- Pearlman, R. S. Rapid generation of high quality approximate 3D molecular structures. *Chemical Design Automation News*, **1987**, *2,* 1, 5-7.
- Hiller, C.; Gasteiger, J. Ein automatisierter Molekülbaukasten. In *Software-Entwicklung in der Chemie*, Gasteiger, J., Ed.; Springer: Berlin, 1987; Vol. 1; pp. 53-66.
- Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Computer Methodology* **1990**, *3*, 537-547.

Gasteiger Most Cited

Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000-1008.

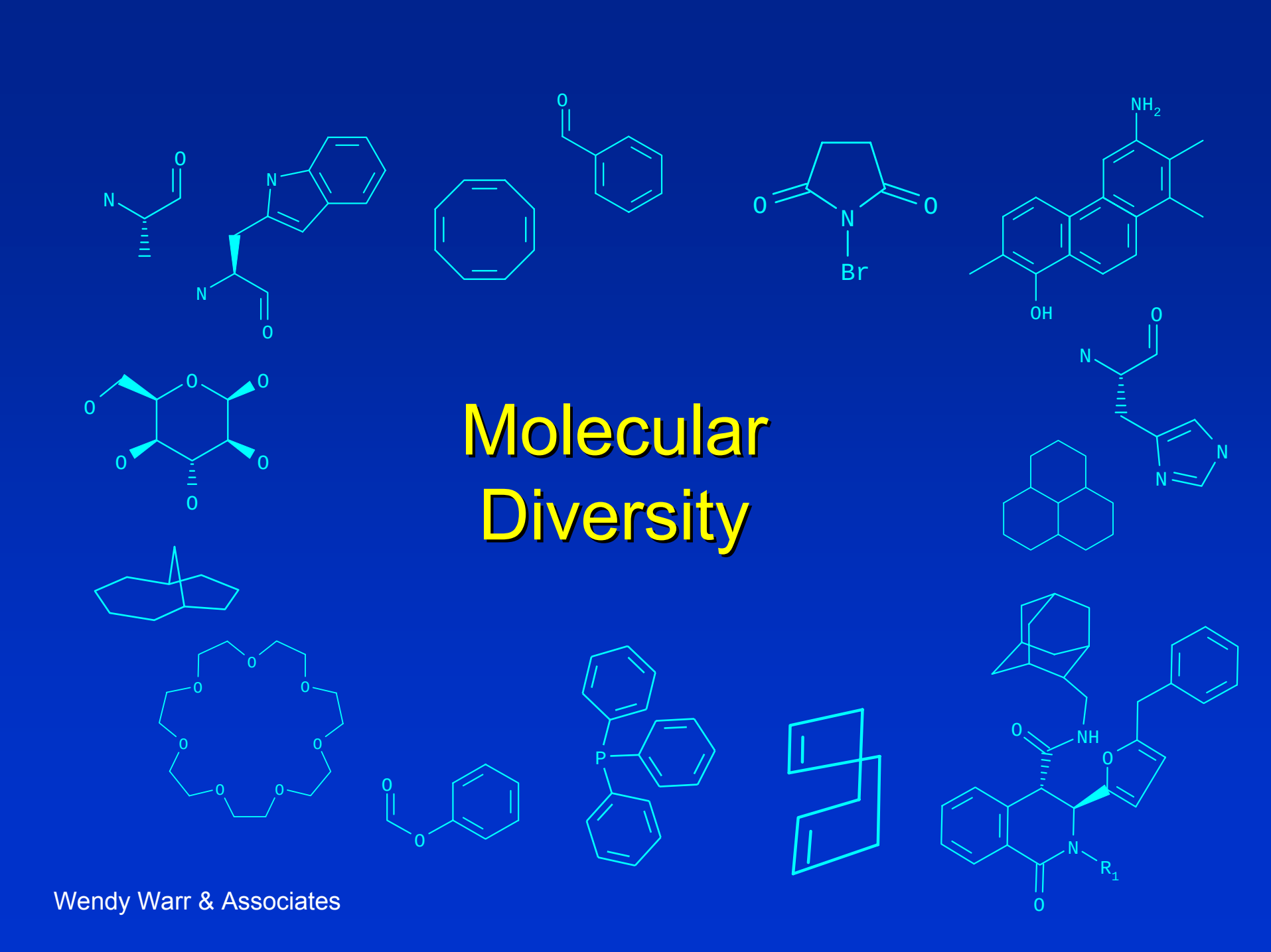
3D Searching

- Jakes, S.E.; Willett, P. Pharmacophoric pattern matching in files of 3D chemical structures: selection of inter-atomic distance screens. *J. Mol. Graph.* **1986**, *4*, 12- 20.
- Cringean, J.K.; Pepperrell, C.A.; Poirrette, A.R.; Willett, P. Selection of screens for three-dimensional substructure searching. *Tetrahedron Computer Methodology* **1990**, *3*, 37-46.
- Brint, A. T.; Willett, P. Pharmacophoric Pattern Matching in Files of 3D Chemical Structures: Comparison of Geometric Searching Algorithms. *J. Mol. Graph.* **1987**, *5*, 49-56.

3D Searching

- Martin, Y. C.; Danaher, E. B.; May, C. S.; Weininger, D. MENTHOR, a database system for the storage and retrieval of three-dimensional molecular structures and associated data searchable by substructural, biologic, physical, or geometric properties. *J. Comput.-Aided Mol. Design* **1988**, *2*(*1*), 15-29.
- Van Drie, J. H.; Weininger, D.; Martin, Y. C. ALADDIN: an integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures. *J. Comput.-Aided Mol. Design* **1989**, *3*(*3*), 225-251.

Molecules Twist



Wendy Warr & Associates

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

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?

Selecting Diverse Subsets

- Clustering
- Dissimilarity-based selection
- Partitioning/cell-based approaches
- Optimization-based methods

Measuring Diversity

Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* **1995**, *38*, 1431-1436.

Diversity

- 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.
- 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.*, **37** (1), 1 -9, 1997.

Product-based or Reagent-based Design

Gillet, V. J.; Willett, P.; Bradshaw, J. The effectiveness of reactant pools for generating structurally diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731-740.

Leads and Drugs

- Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23(1-3)*, 3-25.
- Hann, M. H.; 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.

Leads and Drugs

- 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.
- Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed. Engl.* **1999**, *38(24)*, 3743-3748.

Methods for vHTS

- Identifying drug-like structures
- 2D similarity
- 3D pharmacophores
- Docking

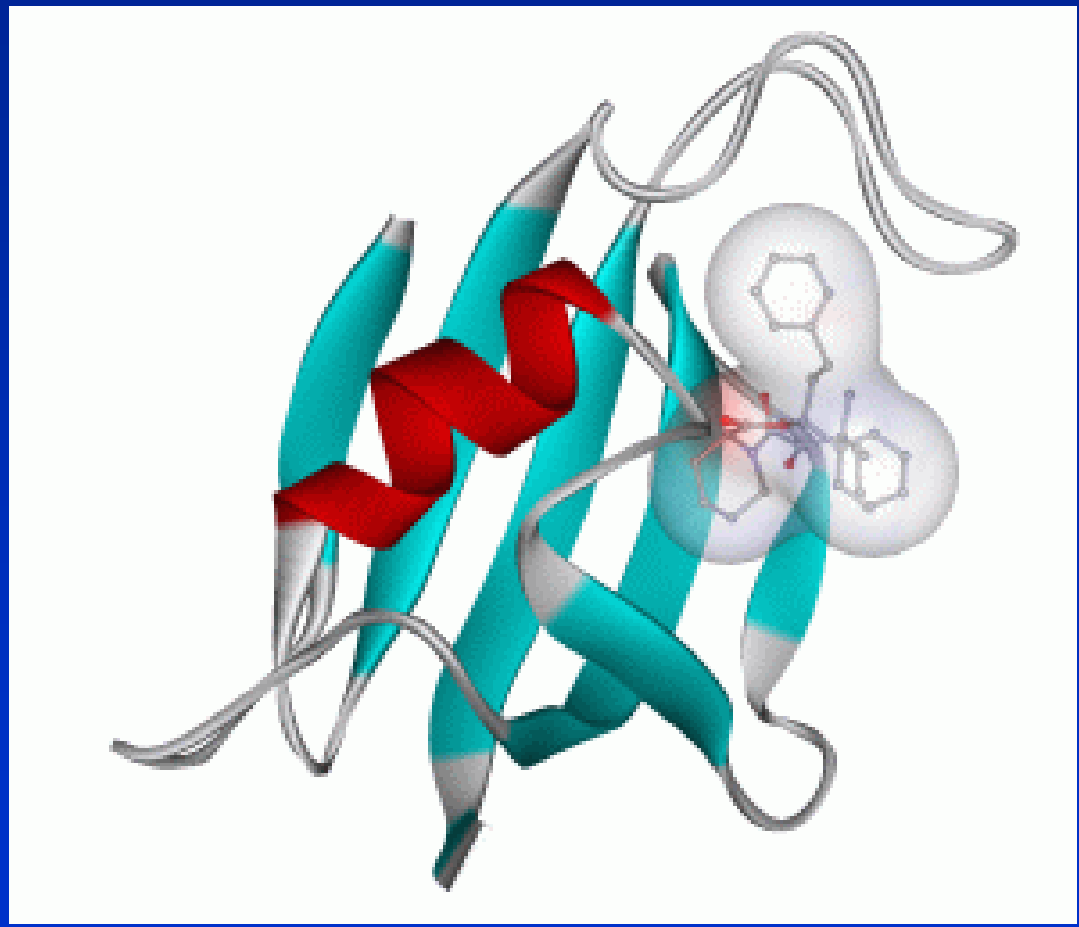
Docking

- Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. *J. Mol. Biol.*, **1982**, *161*, 269.
- Shoichet, B. K.; Kuntz, I.D. Protein docking and complementarity. *J. Mol. Biol.* **1991**, *221*, 327-46.
- Shoichet, B. K.; Stroud, R. M.; Santi, D. V.; Kuntz, I. D.; Perry, K. M. Structure-based discovery of inhibitors of thymidylate synthase. *Science* **1993,** *259*, 1445-1450.

Docking

- Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*(3), 243-56.
- Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727-748.

GOLD docking result for PDB 1FKG



Most Cited. Top Twenty

- 20. Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165-179.
- 19. Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept**.** *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28-35.
- 18. 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.

Most Cited. Top Twenty

- 17. Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731-740.
- 16. Krygowski, T. M. Crystallographic Studies of Inter- and Intramolecular Interactions Reflected in Aaromatic Character of π-Electron Systems. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 70-78.
- 15. 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.

Most Cited

- **14.** Gillet, V. J.; Willett, P.; Bradshaw, J. The effectiveness of reactant pools for generating structurally diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731-740.
- **13.** Hall, L. H.; Mohney, B; Kier, L. B. The electrotopological state: structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31,* 76-82.
- **12.** Hall, L. H.; Kier, L. B. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039-1045.

Most Cited

- **11.** Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.;* **1987**, *27*, 21-35.
- **10.** Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73.
- **9.** 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.

Most Cited

- **8.** 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.
- **7.** Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983-996.
- **6.** Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854-866.

Most Cited

- **5.** Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36.
- **4.** Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163-172.
- **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.

Most Cited

- **2.** Fletcher, D. A.; McMeeking, R. F.; Parkin, D. The United Kingdom chemical database service. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 746-749.
- **1.** 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.

Acknowledgments

- Chemical Abstracts Service
- Eric Shively, CAS
- Michael Lynch

The Next 25 Years?

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."

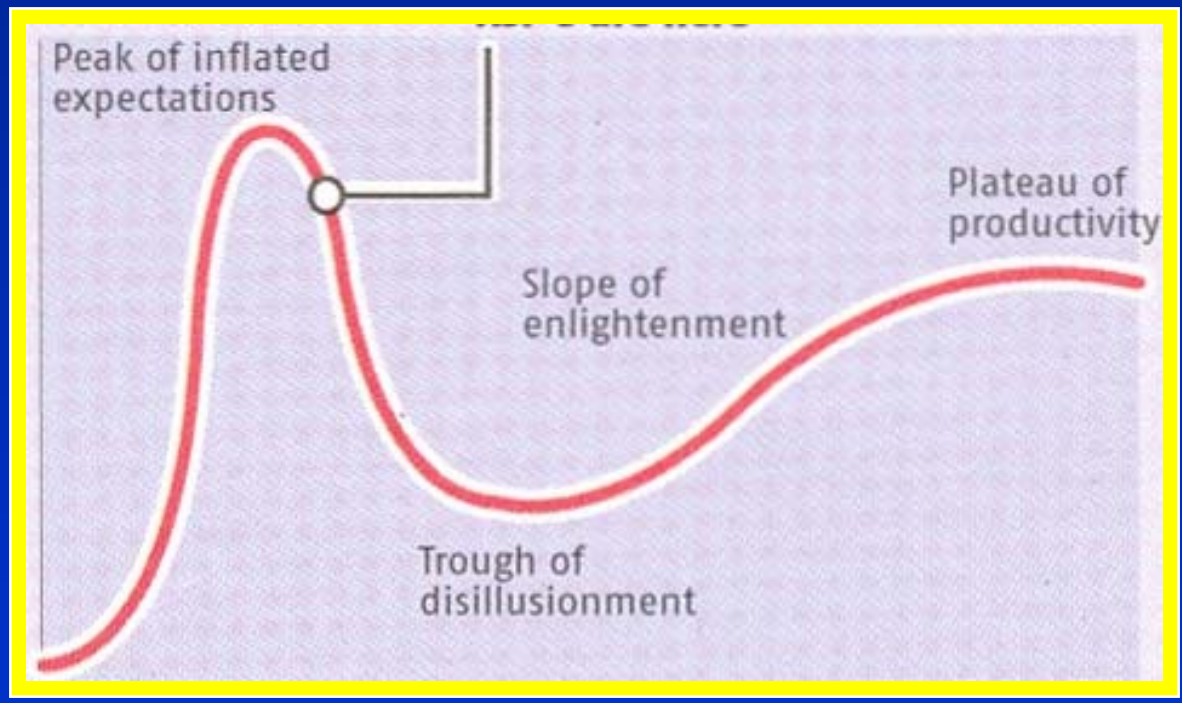
Popular Mechanics, March 1949

Basic Principle of Forecasting

"Give them a number or give them a date, but never both."

Edgar Fiedler

Technology Adoption Cycle



Information Overload

- 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

A Mobile CAS Service

New Infrastructures, Middleware

- e-Science
- Cyberinfrastructure
- eSciDoc

"Here we sit side by side with those on whose shoulders we stand"

Michael Lynch, 2002