# Intrinsic dimensionality of chemical space: Characterization and applications

Subhash C. Basak, Gregory D. Grunwald
and Subhabrata Majumdar
NRRI-UMD, University of Minnesota

NATURAL RESOURCES
RESEARCH INSTITUTE

# Acknowledgements

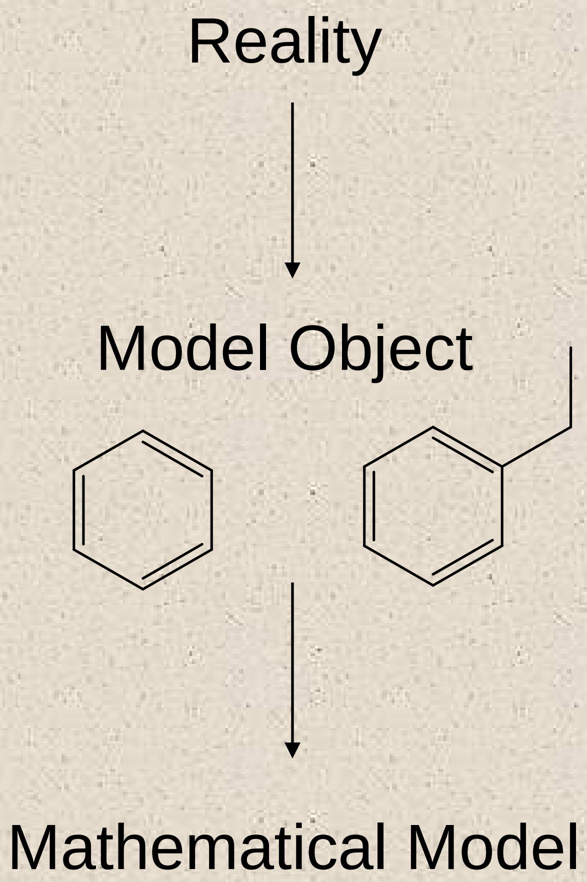
- US Environmental Protection Agency
- United States Air Force, Office of Scientific Research
- Agency for Toxic Substances and Disease Registry, Center for Disease Control and Prevention
- USDA
- ~ US 7.5 Million dollars since 1987

UMD DULUTH NATURAL RESOURCES RESEARCH INSTITUTE

Property/ Activity/ Toxicity = $f$ (S)

# What is structure ?

- The structure of an assembled entity, e. g., a molecule can be looked upon as the relationship among its constituent parts
- A graph, G= [V, E] is an adequate representation of molecules where V is the set of atoms and E is the set of bonds or edges



Reality

Model Object

Mathematical Model

*Method, Model and Matter, by Bunge*

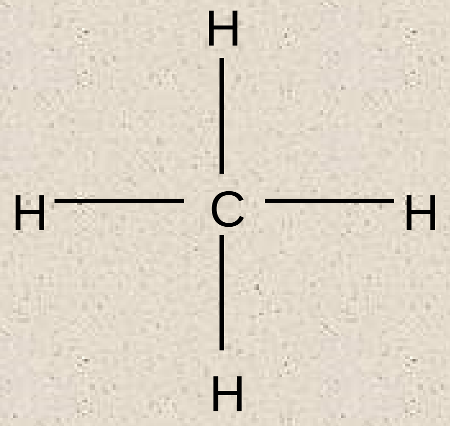
# Representation of Molecular Structures by Graphs

Let V = (1, 2, 3, 4, 5)

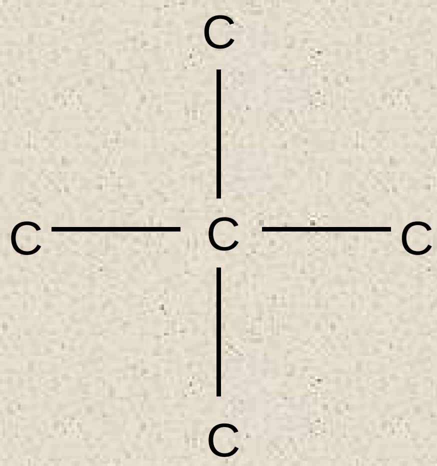
V x V = {(1,1), (1,2), (1,3), (1,4), (1,5) …}

$R_1$ = {(1,5), (5,1), (2,5), (5,2), (3,5), (5,3), (4,5), (5,4)}

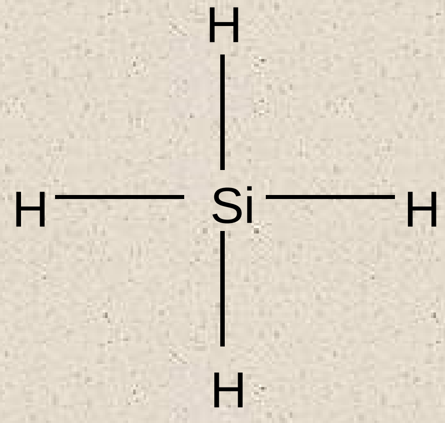
A binary relation on the set V



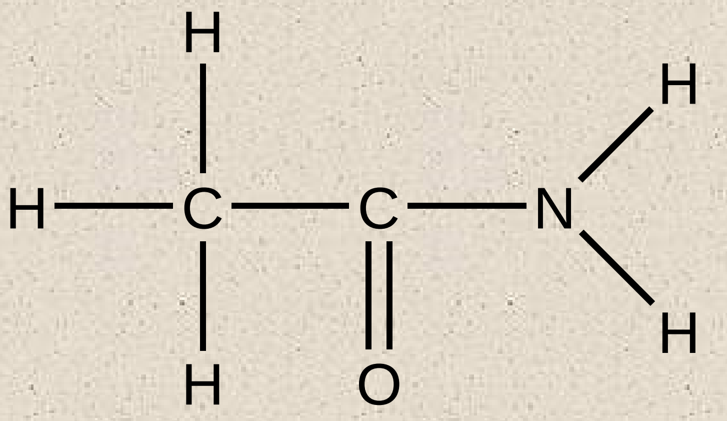
Methane



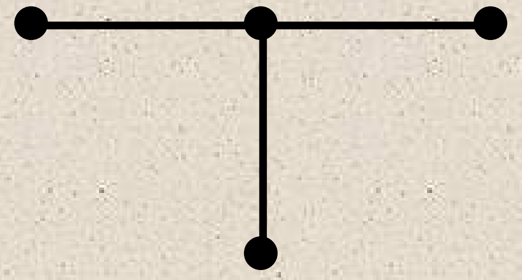
Neopentane



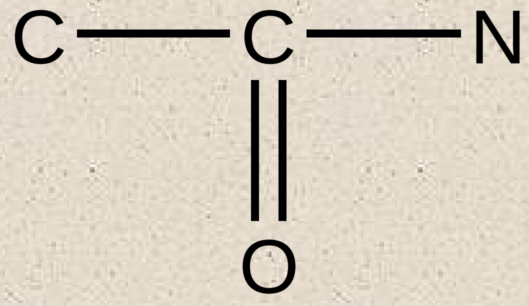
Silicon tetrahydride



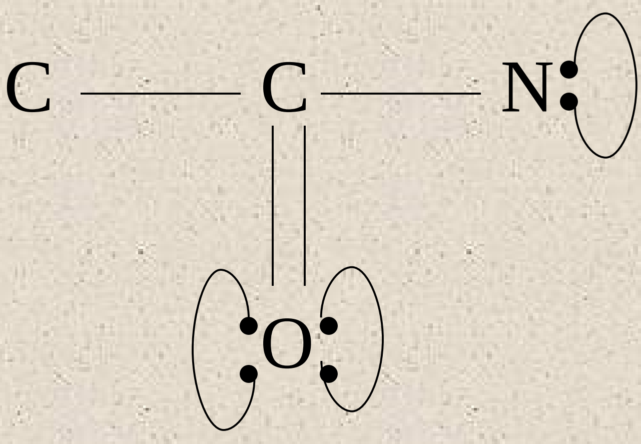
Molecular Structure



Simple Graph



Multigraph



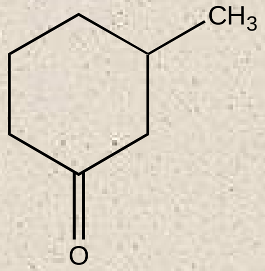
Pseudograph

UMD DULUTH Natural Resources Research Institute

# Structure is a complex idea

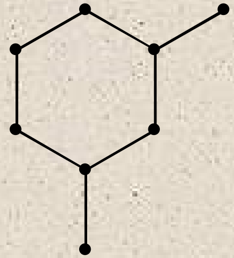
# Hierarchical Approach to Chemical Structure Representation

3-methylcyclohexanone



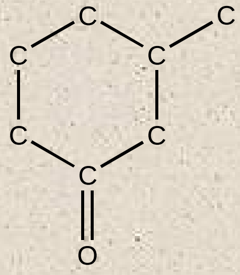
Chemist’s representation of structure

Topostructural Model



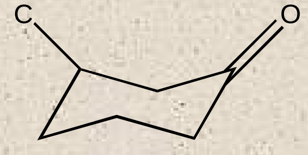
**Simple graph:**
Purely structural representation

Topochemical Model



**Chemical graph:**
Contains chemical and valence information

Geometrical Model



**3-Dimensional:**
Based on chemical graph

Quantum Chemical Model

$$H\Psi = E\Psi$$

Based on principals of quantum mechanics

# Characterization of Molecular Graphs Using TIs

Molecular graphs can be characterized using numerical graph invariants or topological indices (TIs)

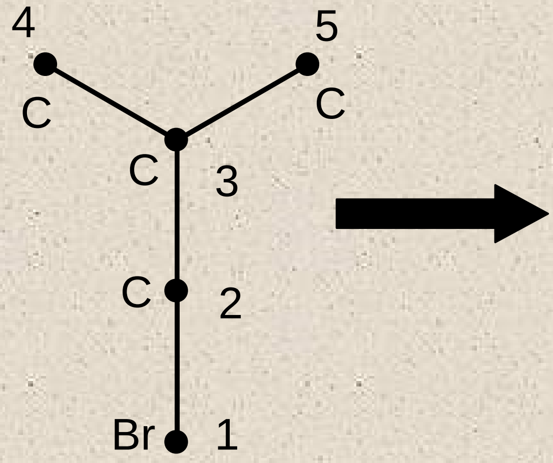
- Simple graph
- Multigraph
- Weighted graphs

Molecular Graph → Molecular descriptor

# Wiener Index, W

$$W = 1/2 \sum_{ij} d_{ij}$$

where $d_{ij}$ is the distance between vertices $v_i$ and $v_j$ in $G_1$

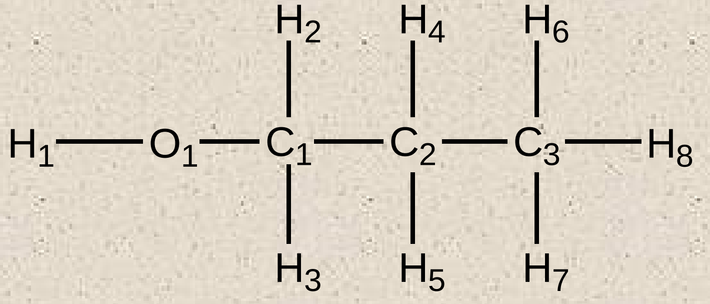


$$\begin{array}{cccccc|c}
 & 1 & 2 & 3 & 4 & 5 & \text{Row Sum} \\
1 & 0 & 1 & 2 & 3 & 3 & 9 \\
2 & 1 & 0 & 1 & 2 & 2 & 6 \\
3 & 2 & 1 & 0 & 1 & 1 & 5 \\
4 & 3 & 2 & 1 & 0 & 2 & 8 \\
5 & 3 & 2 & 1 & 2 & 0 & 8 \\
\hline
 & & & & & & 36
\end{array}$$

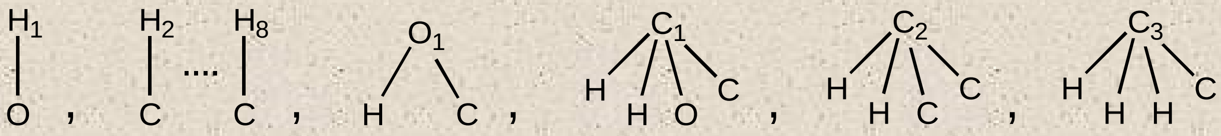
**W = 36 / 2**
**= 18**

# Calculation of IC, SIC & CIC

Labeled graph:



First Order Neighborhoods:



| Subsets: | I | II | III | IV | V | VI |
|---|---|---|---|---|---|---|
| | $(H_1)$ | $(H_2\text{-}H_8)$ | $(O_1)$ | $(C_1)$ | $(C_2)$ | $(C_3)$ |
| Probability: $(p_i)$ | 1/12 | 7/12 | 1/12 | 1/12 | 1/12 | 1/12 |

(Basak, Roy, Magnuson, and Harriss)

# Equivalence Relation

- Can partition the vertex set, *V*(*G*), into disjoint subsets based on topological neighborhoods of vertices up to the *r*th order neighbors and provide indices of neighborhood complexity

- Is reflexive, symmetric, and transitive

# Measures of Complexity and Redundancy

Information Content ($IC_1$)

$$
\begin{aligned}
IC_1 &= -\sum p_i \log_2 p_i \\
&= 5 \times {}^1\!/_{12} \times \log_2 {}^1\!/_{12} + {}^7\!/_{12} \times \log_2 {}^7\!/_{12} \\
&= 1.950 bits
\end{aligned}
$$

$$SIC_1 = IC_1 / \log_2 12 = 0.544$$

$$CIC_1 = \log_2 12 - IC_1 = 1.635 bits$$

Basak, Roy and Ghosh, *Proc. 2nd Intl. Conf. Math.Modelling*, pp. 851-856, **1979.**

Roy, Basak, Harriss and Magnuson, *Mathl. Modelling Sci. Technol.*, pp. 745-750, 1984.

Basak and Magnuson, *Arzneim. Forsch./Drug Res.*,**33**, 501-503, 1983.

Raychaudhury, Ray, Ghosh, Roy and Basak, *J.Comput. Chem.*, **5**, 581-588, 1984.

# QSAR and Molecular Descriptors

# Strategies

- Laboratory experiments
- Property-property correlations [$P_1 = f\,(P_2)$[a]]
- Structure-property correlations [$P = f\,(S)$[b]]
  - QSAR / QSPR
  - Molecular Similarity

[a] Experimentally determined

[b] Calculated

**POLLY**

The Upjohn Company

Glaxo (USA)

US Army

NIH, NINDS

US Environmental Protection Agency

**APProbe**

The Upjohn Company

Glaxo (USA)

# Data Reduction via Principal Components Analysis

## 3,692 chemicals; 90 diverse TIs

(*Basak, Magnusson, Niemi, Regal, and Veith,* ***1987***)

| Principal Component (PC) | Eigenvalue | Percent of variance | Cumulative percent |
|---|---|---|---|
| 1 | 39.6 | 44.0 | 44.0 |
| 2 | 14.6 | 16.2 | 60.2 |
| 3 | 9.9 | 11.0 | 71.2 |
| 4 | 6.4 | 7.1 | 78.3 |
| 5 | 3.3 | 3.7 | 82.0 |
| 6 | 3.2 | 3.5 | 85.5 |
| 7 | 1.9 | 2.1 | 87.6 |
| 8 | 1.8 | 1.9 | 89.5 |
| 9 | 1.5 | 1.7 | 91.2 |
| 10 | 1.2 | 1.3 | 92.5 |

- 10 PCs with Eigenvalues greater than 1
- First 10 PCs explain 92% of the variance within the data
- First 4 PCs account for 78% of the variance within the data

PC1 → Size

PC2 → Symmetry

PC3 → Branching

PC4 → Cyclicity

*S.C. Basak, V.R. Magnuson, G.J. Niemi, R.R. Regal*
*Discrete Applied Mathematics 19 (1988) 17-44*

# Topological Indices: Their Nature and Mutual Relatedness

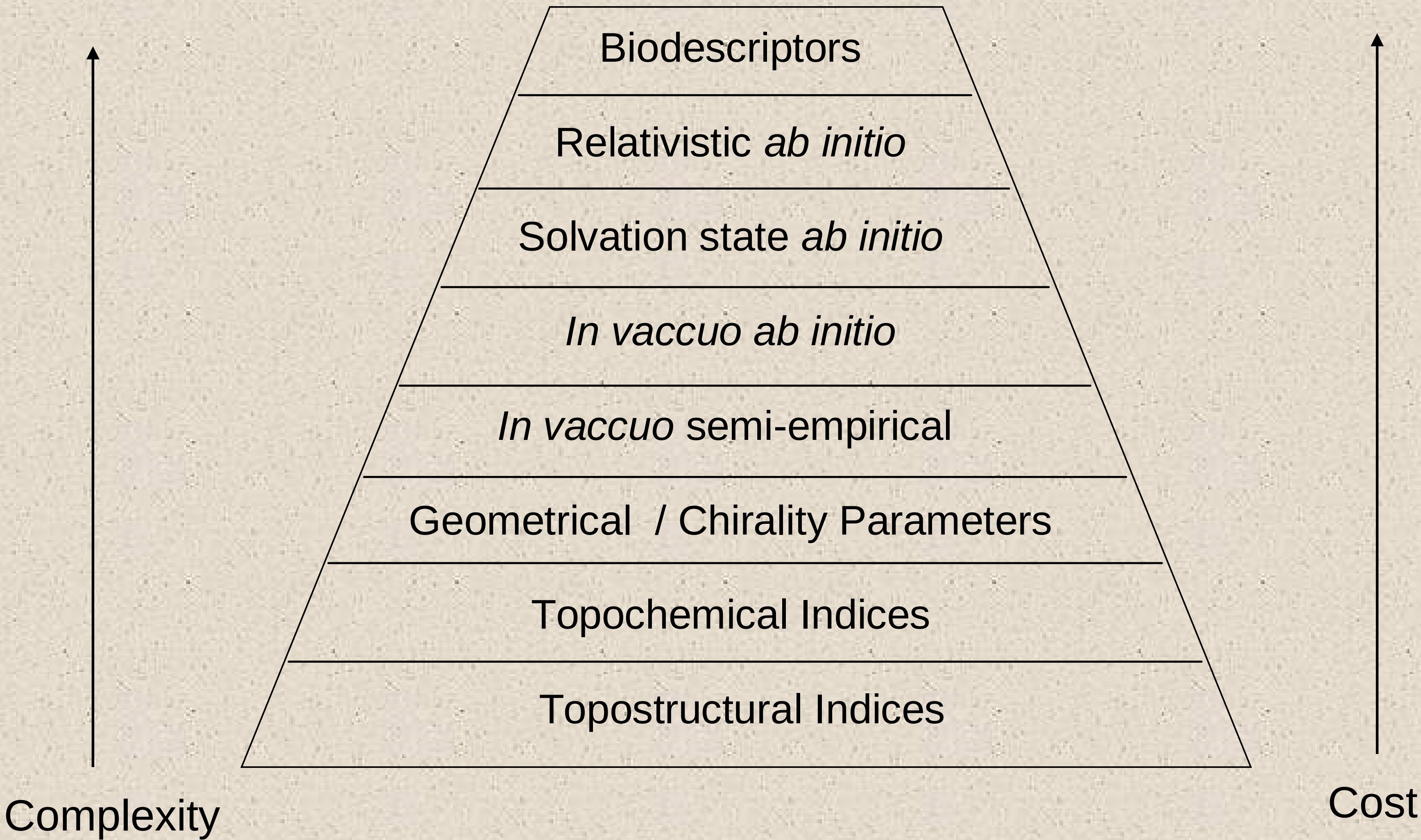
**Subhash C. Basak, Alexandru T. Balaban,**
**Gregory D. Grunwald, and Brian D. Gute**
Natural Resources Research Institute, University of Minnesota--Duluth, Duluth, Minnesota 55811, and Organic Chemistry Department, Polytechnic University Bucharest, Splaiul Independentei 313, 77206 Bucharest, Romania

**Journal of**
**Chemical**
**Information and**
**Computer Sciences**®

Reprinted from
Volume 40, Number 4, Pages 891-898

# Hierarchical QSAR

# Hierarchical QSAR



Biodescriptors

Relativistic *ab initio*

Solvation state *ab initio*

*In vaccuo ab initio*

*In vaccuo* semi-empirical

Geometrical / Chirality Parameters

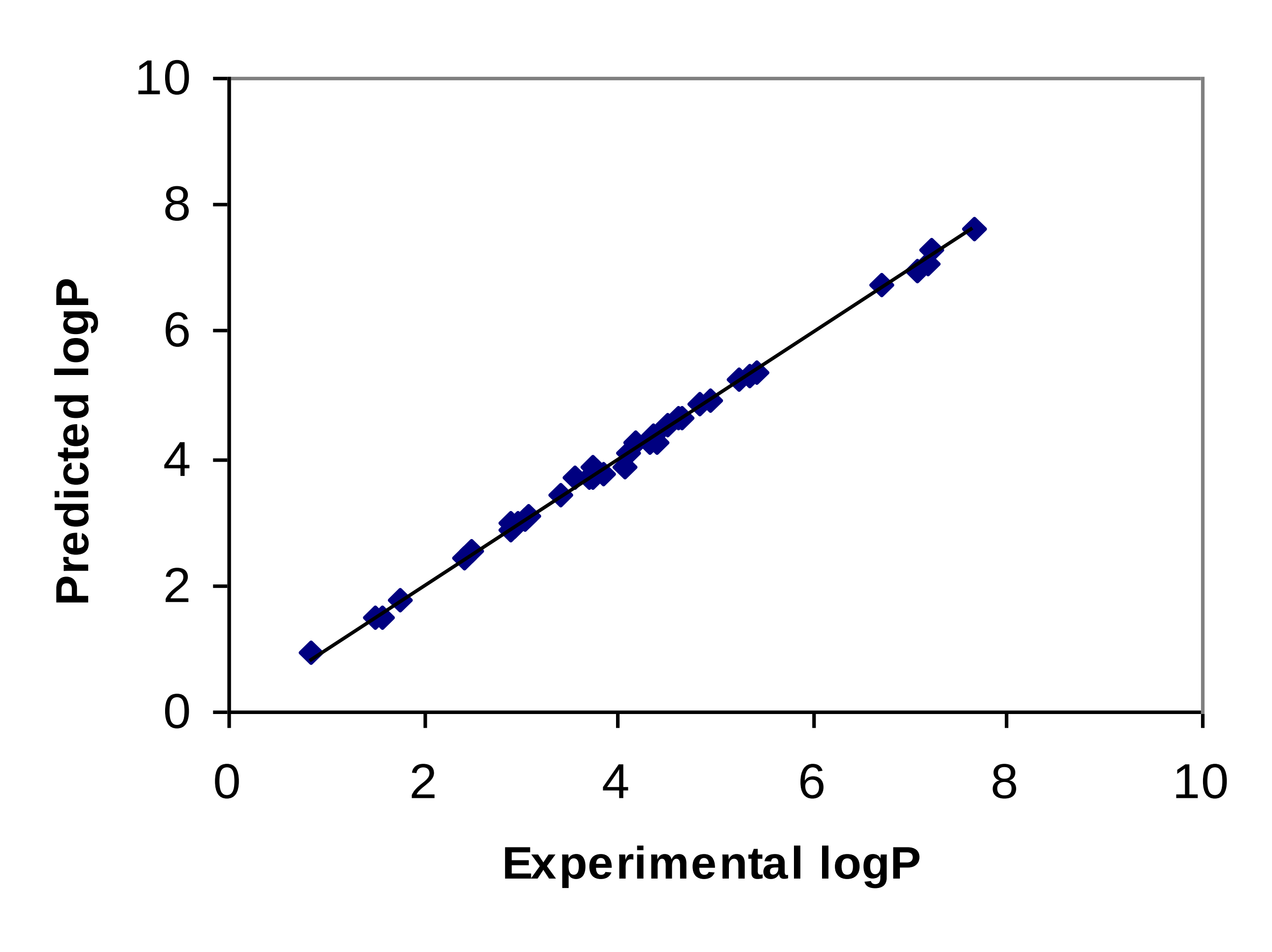
Topochemical Indices

Topostructural Indices

Complexity

Cost

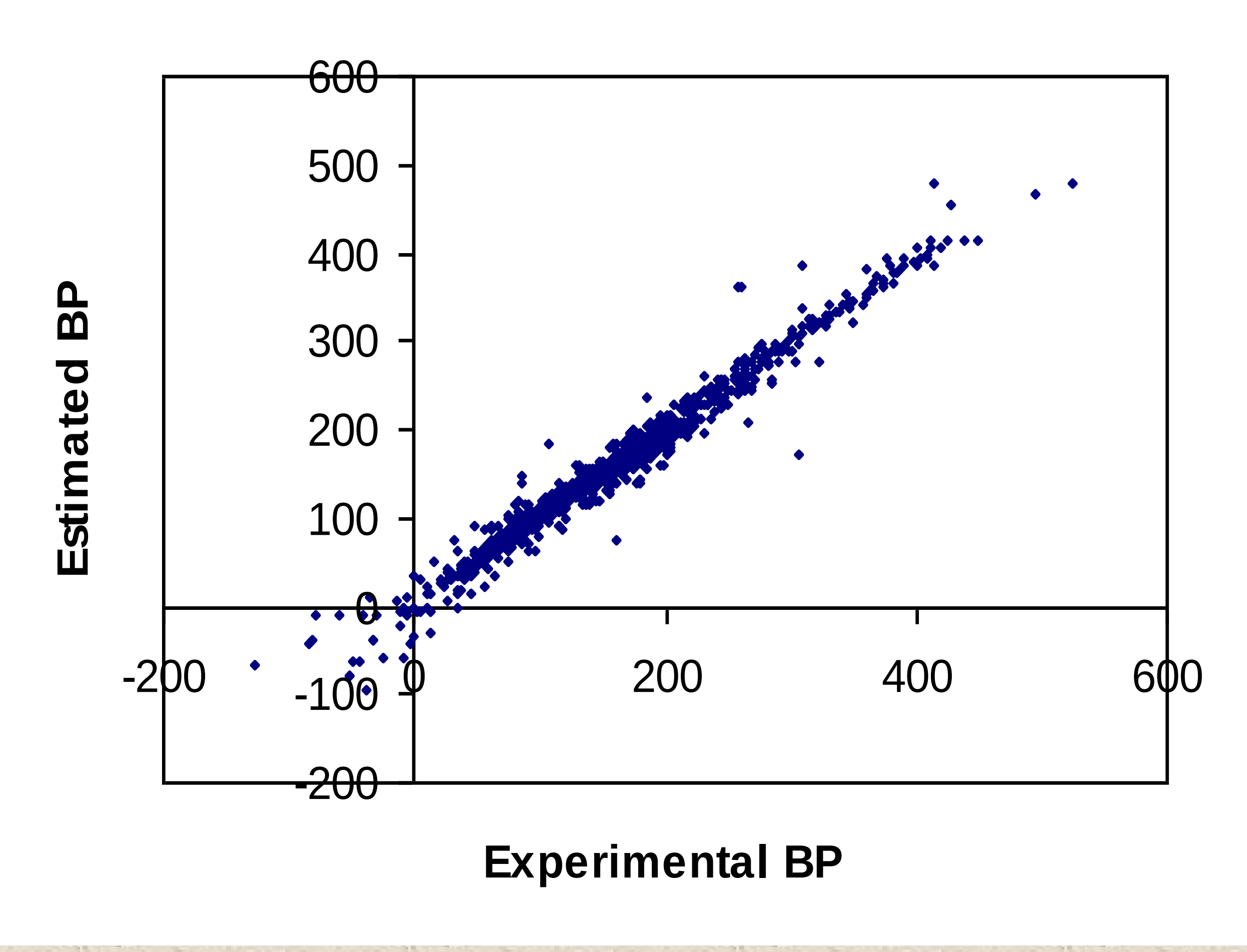
# Blood:Air Partition Coefficient Model (TC) Developed on 39 Diverse Chemicals



S. C. Basak, D. Mills, H. A. El-Masri, M. M. Mumtzaz, and D. M. Hawkins *Environ. Toxicol. Pharmacol.*, 16, 45–55 (2004).

# Normal Boiling Point for 1015 Diverse Chemicals

*n* = 1015, $R^2$ = 0.97, *s* = 15.7 , *F* = 4014



Basak, S. C. and Mills, D. *MATCH*, 2001, 44, 15-30.

# Graph Theoretic vs Quantum Chemical Descriptors for the Prediction of Vapor Pressure

- 121 chlorinated chemicals
- Supercooled liquid VP at 298K
- Graph theoretic descriptors: $q^2 = 0.988$
- Polarizability (DFT, B3LYP): $q^2 = 0.974$

Basak, S. C.; Mills, D. SAR QSAR Environ. Res., in press.

# Improvement in Predictive Models Upon Inclusion of Quantum Chemical Descriptors?

| Description of Data Set and Property/Activity | Improvement |
|---|---|
| Acute toxicity of benzene derivatives | Minimal |
| Dermal penetration of PAHs | None |
| Mutagenicity of aromatic and heteroaromatic amines | None |
| Mutagenicity of 508 diverse compounds | None |
| Vapor pressure of 469 diverse compounds | None |
| Cellular toxicity of halocarbons | Minimal |
| Mosquito repellency of aminoamides | None |
| Mosquito repellency of DEET-related compounds | None |
| Blood and tissue:air partition coefficient for rat and human (blood, fat, brain, liver, muscle, and kidney) | None |
| Aryl hydrocarbon receptor binding affinity of dibenzofurans | None |

Basak, S. C.; Mills, D.; Mumtaz, M. M.; Balasubramanian, K. Use of topological indices in predicting **aryl hydrocarbon receptor binding potency of dibenzofurans**: A hierarchical QSAR approach. *Indian J. Chem.*, **2003,** 42A, 1385-1391.

- Gute, B. D.; Grunwald, G. D.; Basak, S. C. Prediction of the **dermal penetration of polycyclic aromatic hydrocarbons** (PAHs): A hierarchical QSAR approach. *SAR QSAR Environ. Res.*, **1999,** 10, 1-15.

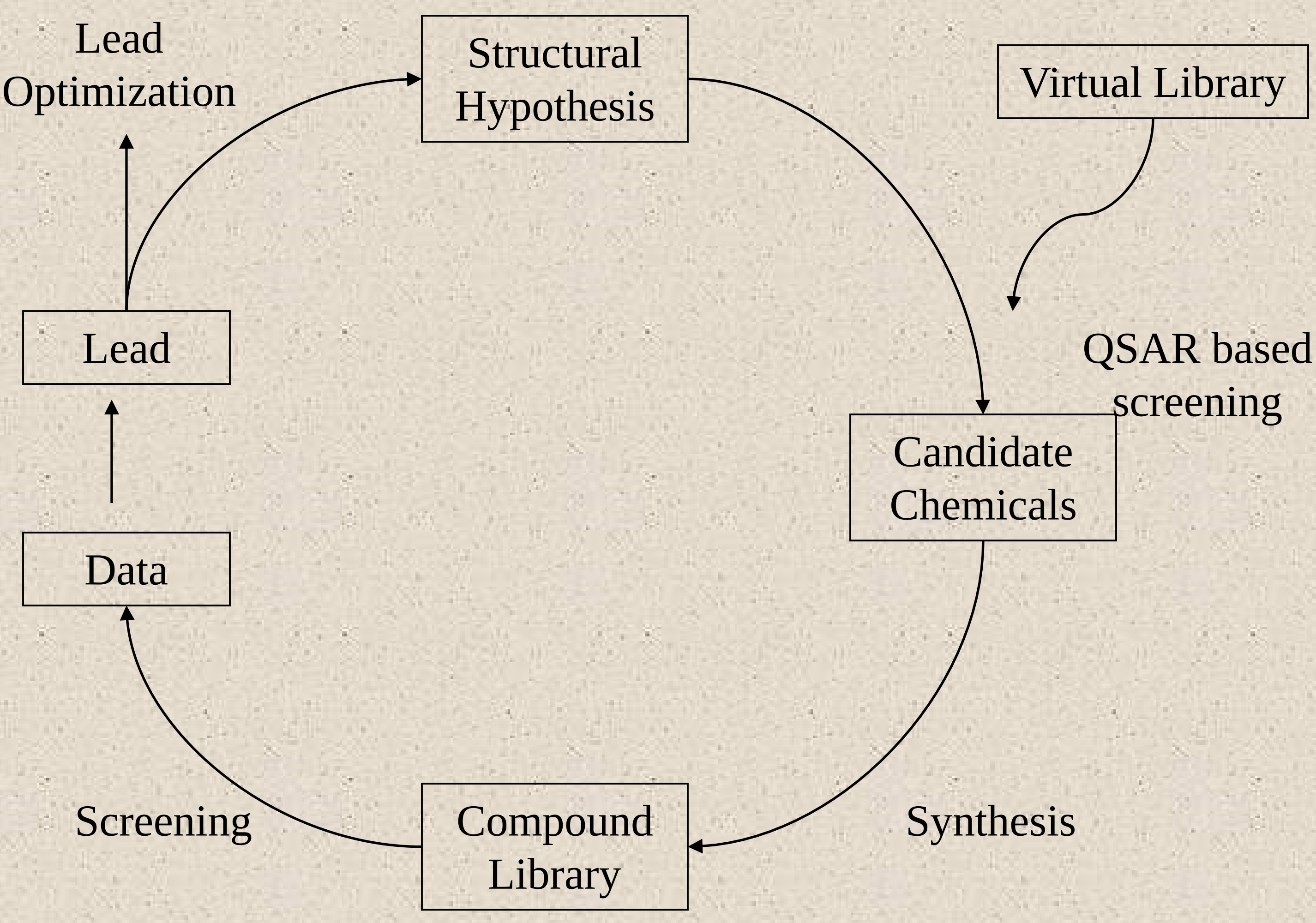
Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of topostructural, topochemical, and geometric parameters in the **prediction of vapor pressure: A hierarchical approach**. *J. Chem. Inf. Comput. Sci.*, **1997,** 37, 651-655.

Basak, S. C.; Grunwald, G. D.; Host, G. E.; Niemi, G. J.; Bradbury, S. P. A comparative study of molecular similarity, statistical, and neural network methods **for predicting toxic modes of action of chemicals**. *Environ. Toxicol. Chem.*, **1998,** 17, 1056-1064.

Basak, S. C.; Gute, B. D.; Drewes, L. R. **Predicting blood-brain transport of drugs:** A computational approach. *Pharm. Res.*, **1996,** 13, 775-778.

Mushrush, G. W.; Basak, S. C.; Slone, J. E.; Beal, E. J.; Basu, S.; Stalick, W. M.; Hardy, D. R. Computational study of the environmental fate of selected aircraft fuel system **deicing compou**nds. *J. Environ. Sci. Health*, **1997,** A32, 2201-2211.

# Combinatorial Chemistry & QSAR



Lead Optimization

Structural Hypothesis

Virtual Library

Lead

QSAR based screening

Candidate Chemicals

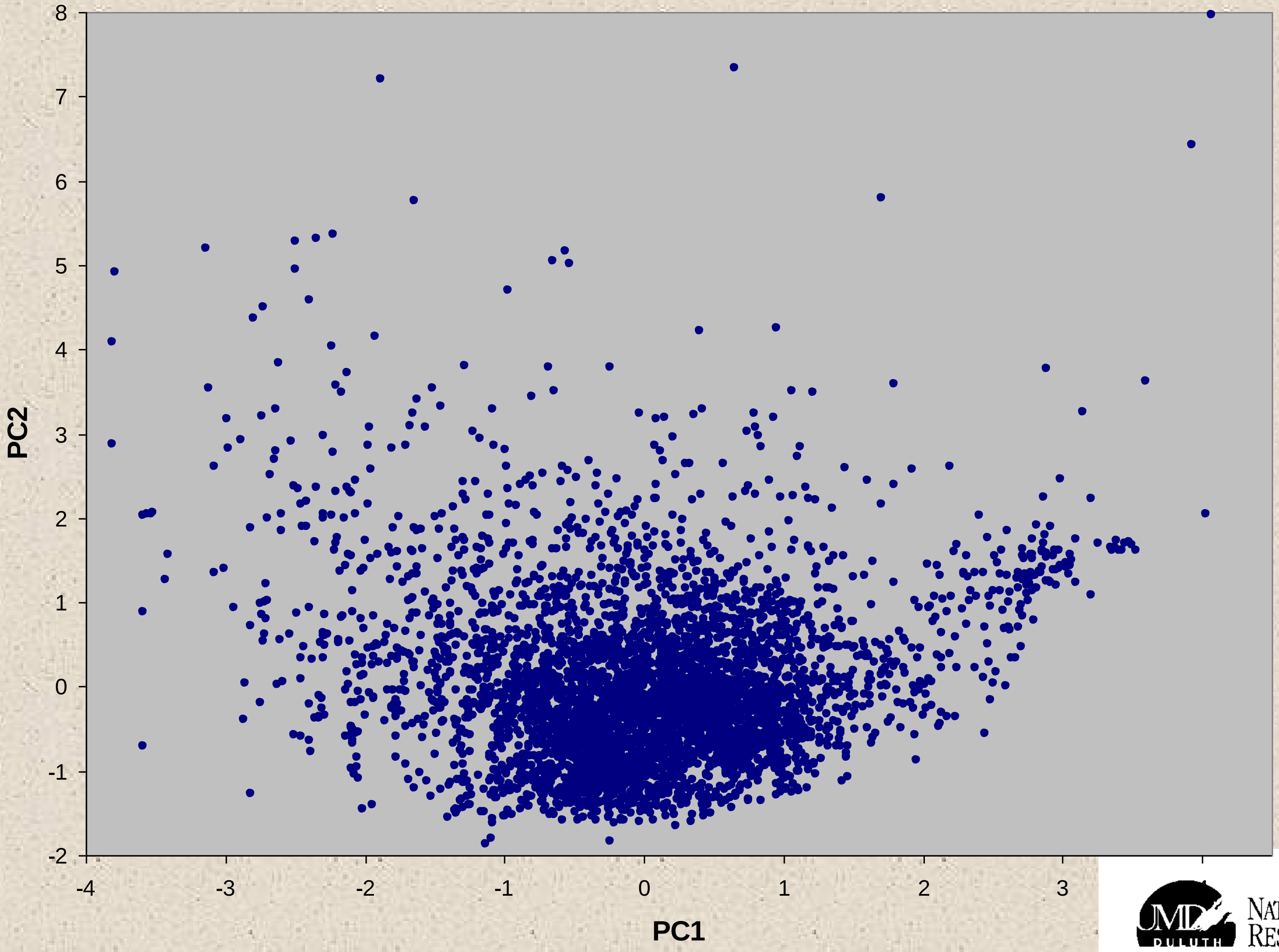
Data

Screening

Compound Library

Synthesis

# $PC_1$ *vs.* $PC_2$ for **4,453** chemicals based on the correlation matrix of 98 variables

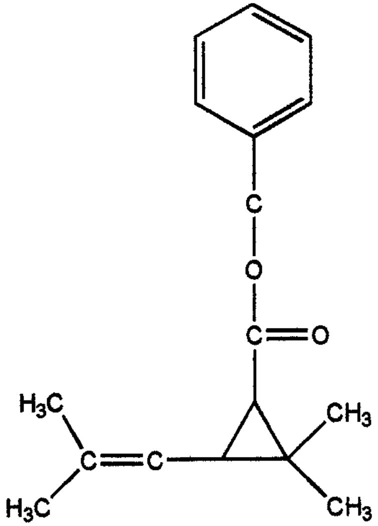


NATURAL RESOURCES RESEARCH INSTITUTE

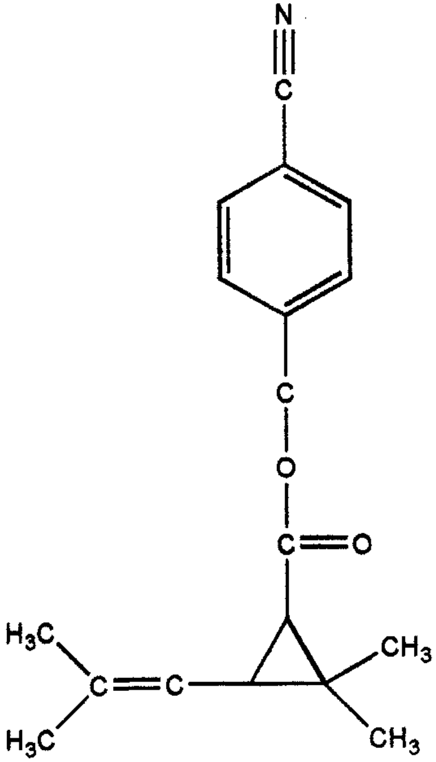
# Euclidean Distance

$$ED_{ij} = \left[\sum_{k=1}^{n} (D_{ik} - D_{jk})^2\right]^{1/2}$$

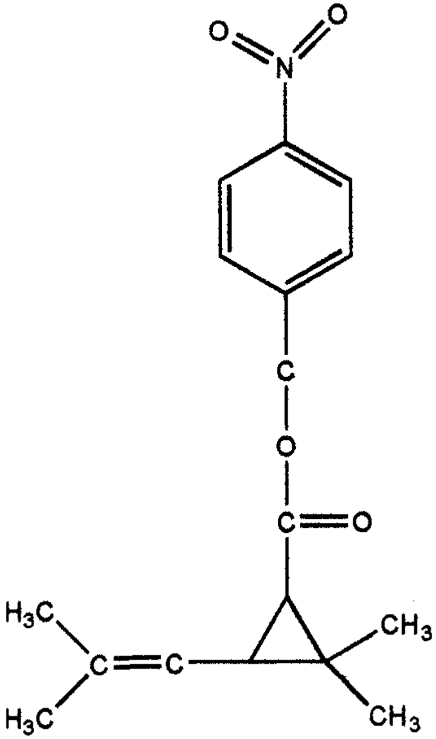
where n = the number of dimensions and $D_{ik}$ and $D_{jk}$ equal the data values of the kth dimension for chemicals i and j, respectively.



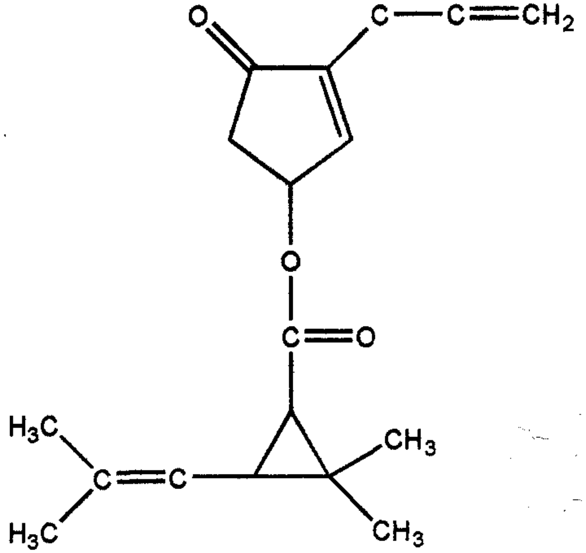
Probe



ED = 0.478



ED = 0.700



ED = 1.159

# *K*-Neighbor Selection and Property Estimation

Intermolecular similarity of chemicals in each set was quantified using 3 to 5 distinct similarity methods.

For each chemical, *K*-nearest neighbors were determined for *K* = 1, 2, …, 10, 15, 20, 25.

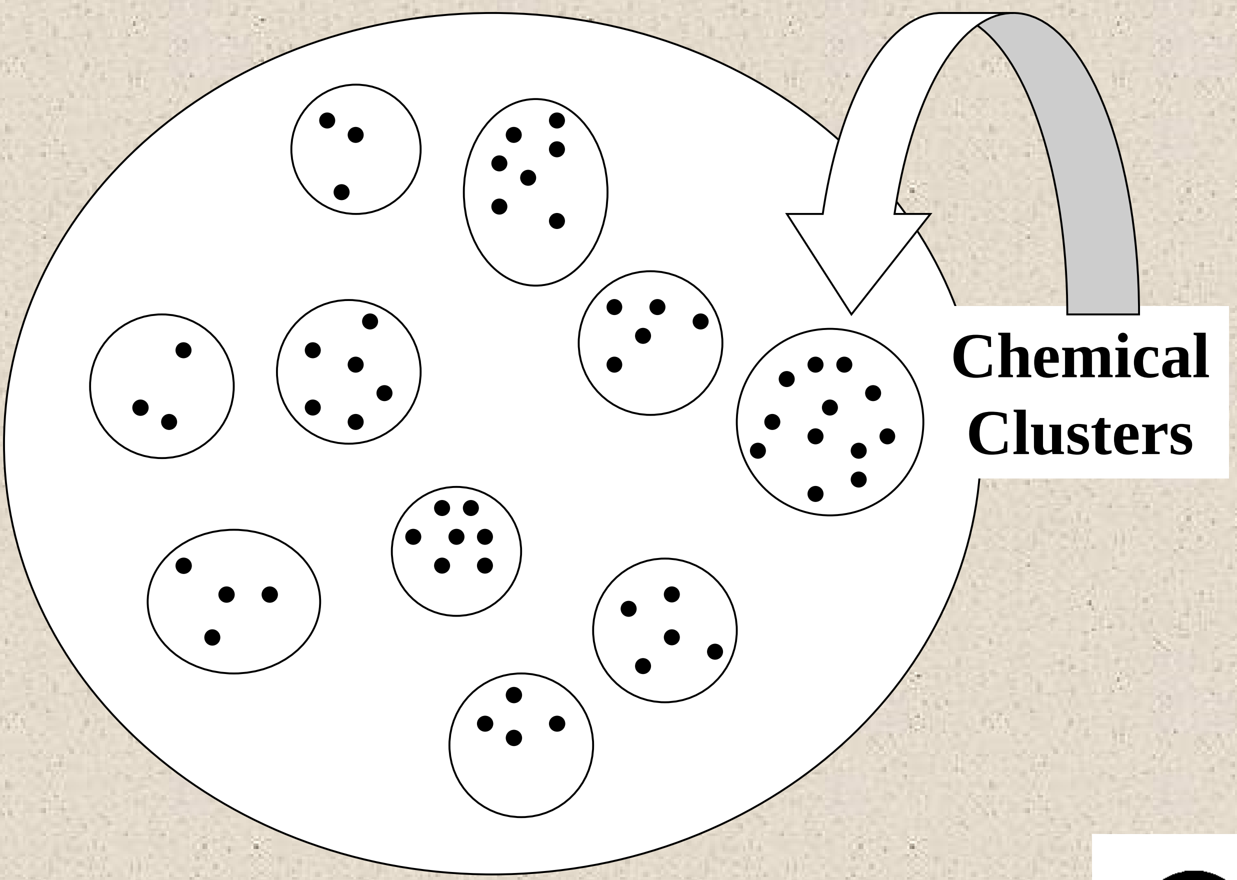
Estimated property values are determined as the mean observed value of the *K*-nearest neighbors.

KNN Estimation of Boiling Points for **1037** Diverse Chemicals

| ***k*** | ***r*** | ***se*** |
|---|---|---|
| 5 | 0.958 | 27.6 |
| 10 | 0.956 | 28.5 |
| 15 | 0.953 | 30.0 |
| 20 | 0.949 | 31.5 |
| 25 | 0.946 | 32.9 |
| 50 | 0.927 | 37.4 |

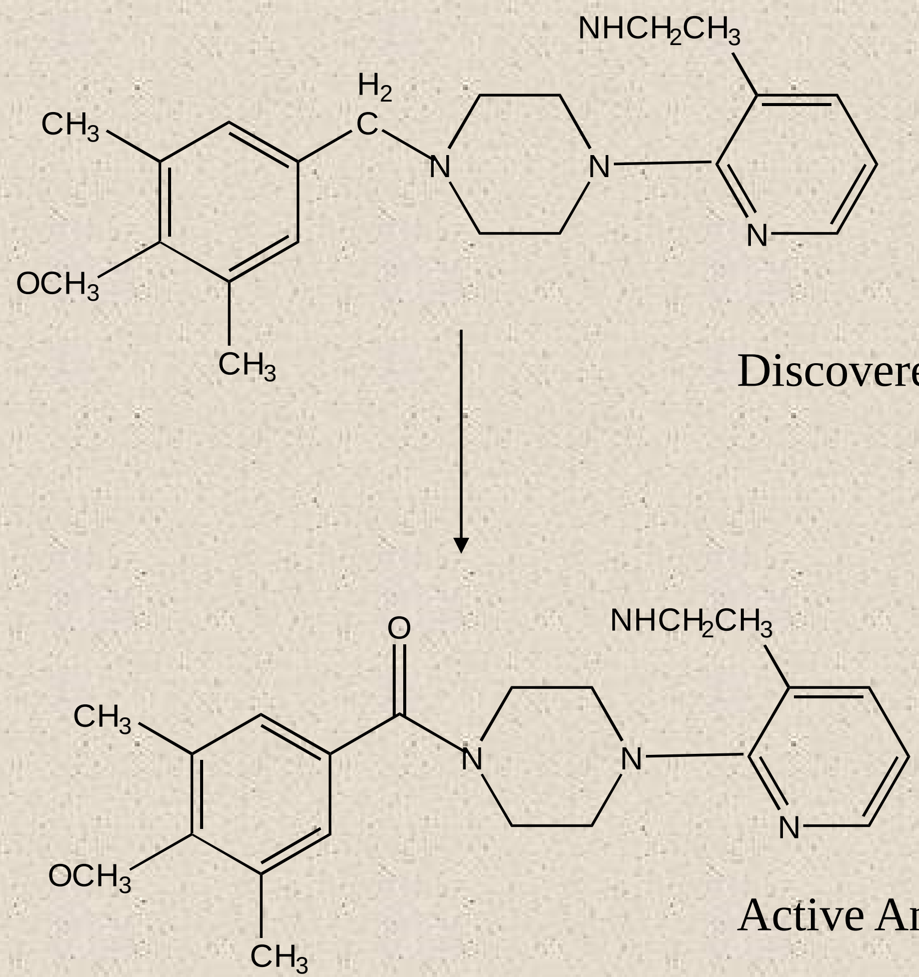
# Structure Space

Chemical Space



Chemical Clusters

# HIV-I RT Inhibitor Discovered by Similarity Search



Discovered Lead

Active Analog

Discovered by Upjohn-Pharmacia

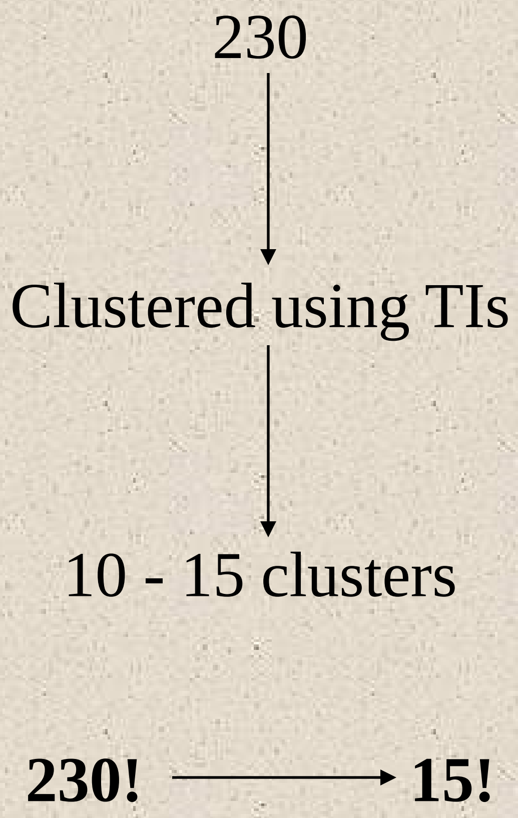
# JP-8

(~230 chemicals, ~2,000 chemicals)

- Skin toxicity
- Immunosuppression
- Systemic toxicity

230! or 2,000! mixtures to be tested

230! or 2,000! mixtures to be tested



230

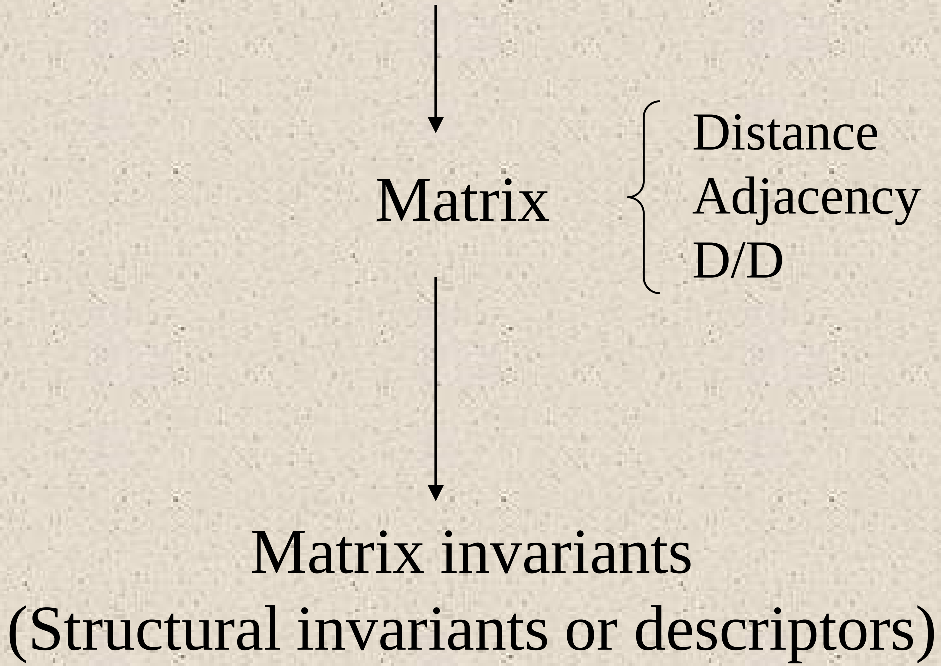
Clustered using TIs

10 - 15 clusters

**230!** → **15!**

# Quo Vadimus?

- Chemical structure
- DNA sequence, genomics
- Proteomics pattern

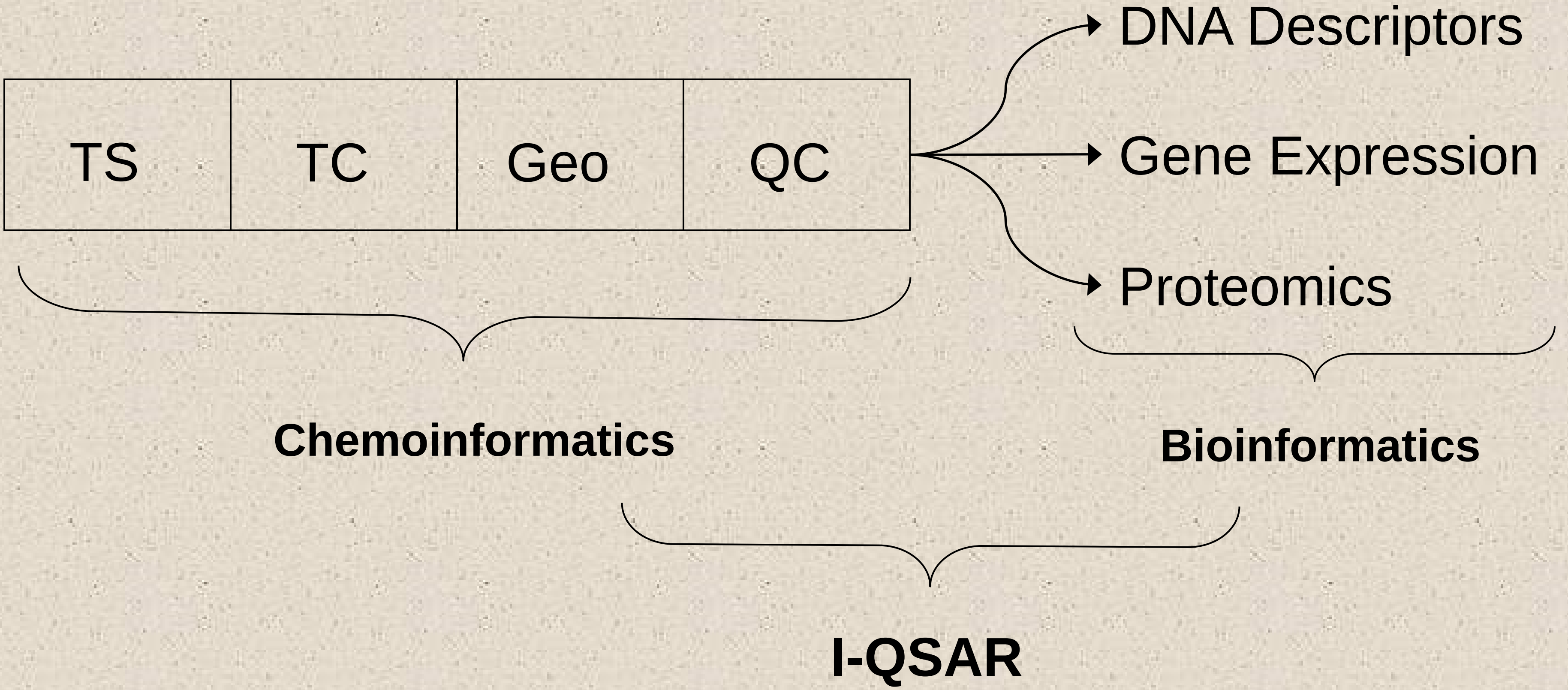


Matrix

Distance
Adjacency
D/D

Matrix invariants
(Structural invariants or descriptors)

# Integrated QSAR (I-QSAR)



TS

TC

Geo

QC

DNA Descriptors

Gene Expression

Proteomics

**Chemoinformatics**

**Bioinformatics**

**I-QSAR**

# Chemo-bioinformatics Guest editorial JCIM, 4 6, 1, 2006

Discrete mathematical chemistry has made important advances in the past twenty five years.  This has been fueled primarily by two factors: a) formulation of new concepts and b) easy access to high speed computers. Methods developed in this field have found applications in pharmaceutical drug design and hazard assessment of environmental pollutants.

# Chemo-bioinformatics Guest editorial JCIM, 4 6, 1, 2006

Interestingly, discrete mathematical concepts, originally developed for the characterization of chemical systems, are being extended to deal with explosion of data in the "omics" science, viz., genomics, proteomics, etc.  A few of the papers from the Fourth Indo-US Workshop published in this issue of JCIM are outstanding examples of this expanding chemo-bioinformatics continuum.

# The enormous landscape

Even the same atoms of the same element, when they exist in different molecules, exhibit different behaviours. The chemical symbol H even seems to signify atoms of a completely different nature. I***n chemistry, this terrible individuality should never be avoided by “averaging,***” and, moreover, **innumerable combinations** of such atoms form the subject of chemical research”

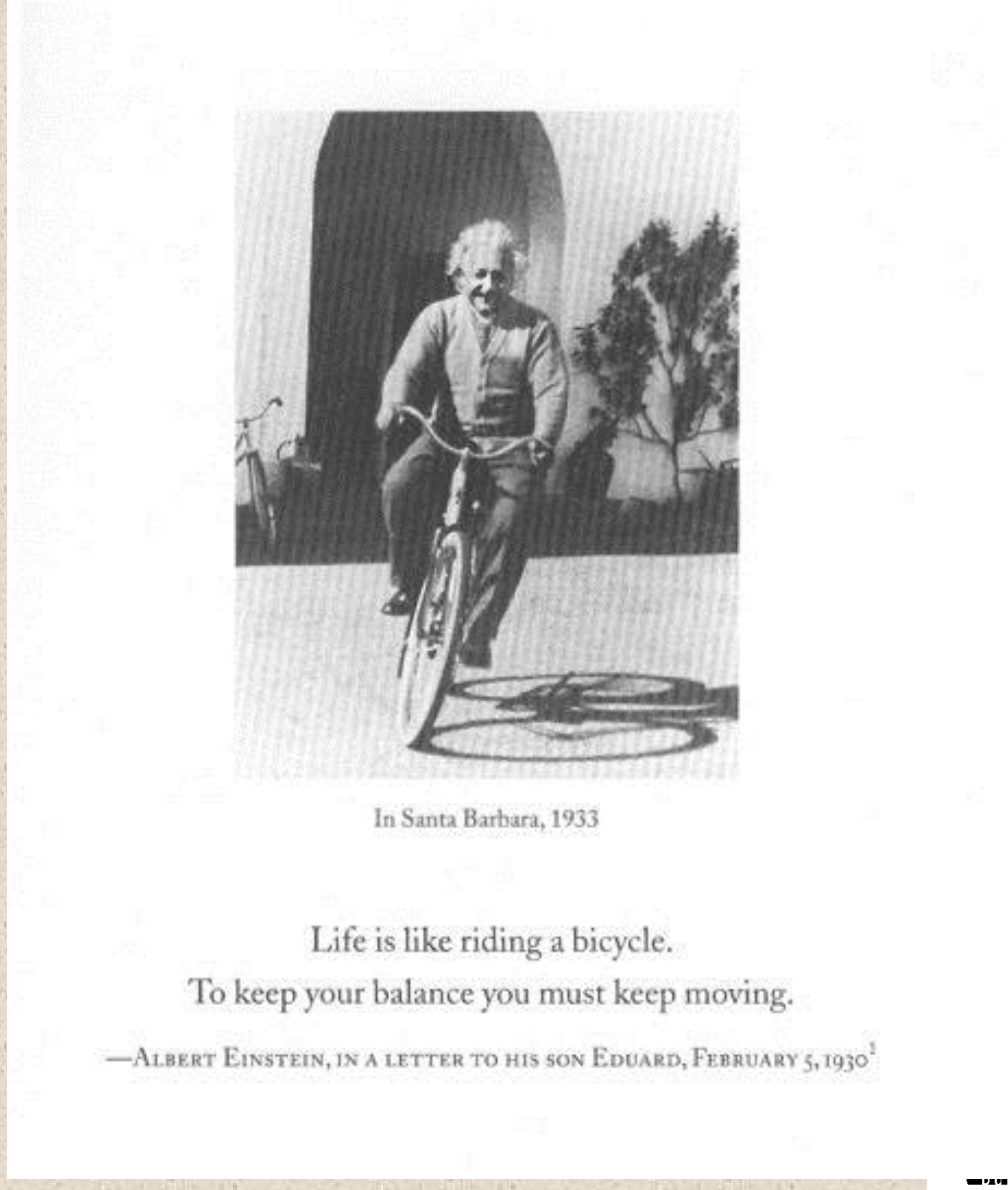
K. Fukui, Nobel Lecture, 1981

# Isomorphic laws in Science

Not only are general aspects and viewpoints are alike in different fields of science; we find also formally identical or isomorphic laws in completely different fields

L. von Bertalanffy, ***British Journal for the philosophy of science***, 1950

NATURAL RESOURCES
RESEARCH INSTITUTE



In Santa Barbara, 1933

Life is like riding a bicycle.
To keep your balance you must keep moving.

—Albert Einstein, in a letter to his son Eduard, February 5, 1930[1]