Sociedade Brasileira de Cristalografia

XVI Reunião da SBCr

IFSC/USP, São Carlos, SP, 16-19/03/03

Use of the Cambridge Structural Database in Study of Single and Partial Double C-X (X=C,N,O) Bonds in Organic Molecules in Crystalline State

Rudolf Kiralj and Márcia M. C. Ferreira

**Laboratório de Quimiometrica Teórica e Aplicada,
Instituto de Química,
Universidade Estadual de Campinas (UNICAMP),
Campinas, SP, 13084-971, Brazil**

***E-mails: marcia@iqm.unicamp.br,
rudolf@iqm.unicamp.br,
URL: http://lqta.iqm.unicamp.br***

UNICAMP

Contents

Introduction to the CSD – general comments

What is the Cambridge Structural Database (CSD)?

Cambridge Crystallographic Data Centre (CCDC), University of Cambridge, UK
http://www.ccdc.cam.ac.uk/

-a database containing structural information (atomic parameters for a crystal of known cell dimensions and space group)

-includes detailed information from X-ray, neutron and synchrotron diffraction studies

-covers organic compounds, organo-metallic compounds, and metal-organic coplexes

-does not include: proteins, high polymers, inorganic compounds, purely inorganic carbon compounds (carbides, carbonates, carbonyls and cyanates)

How many crystal structures does it contain?

-established in 1965: a few thousand structures
-current annual increase is > 20 000 structure
-the last version (November 2002) has **272066** structures

The number of chemical compounds in the CAS (Chemical Abstract Service) Registry Database, >50% of which are peptides and proteins, is growing superexponentially. Exponential trend is observed for the CSD (Figure 1).

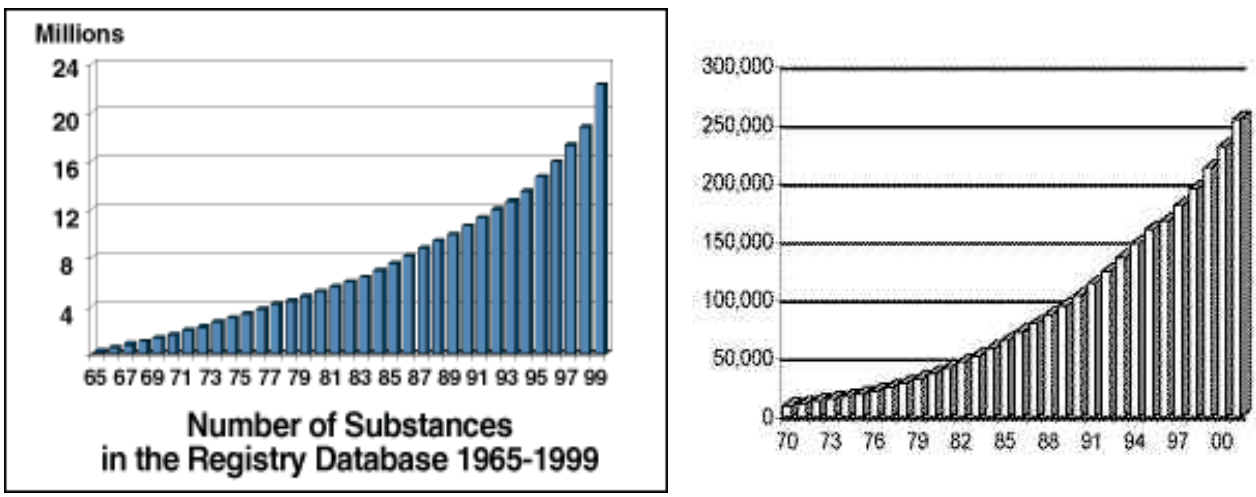


Figure 1. Left: Cumulative growth of the CAS-Registry Database. http://www.cas.org/casdb.html#regdb
Contents on 14-March-2003: 47 120 998 chem. compounds.
Right: Cumulative growth of the CSD from 1970 to 2001. CCDC Annual Report for 2001.
http://www.ccdc.cam.ac.uk/annrep2001/report.html

How and how much scientists use the CSD?

3 basic modes: -1D, 2D, 3D information (Figure 2)

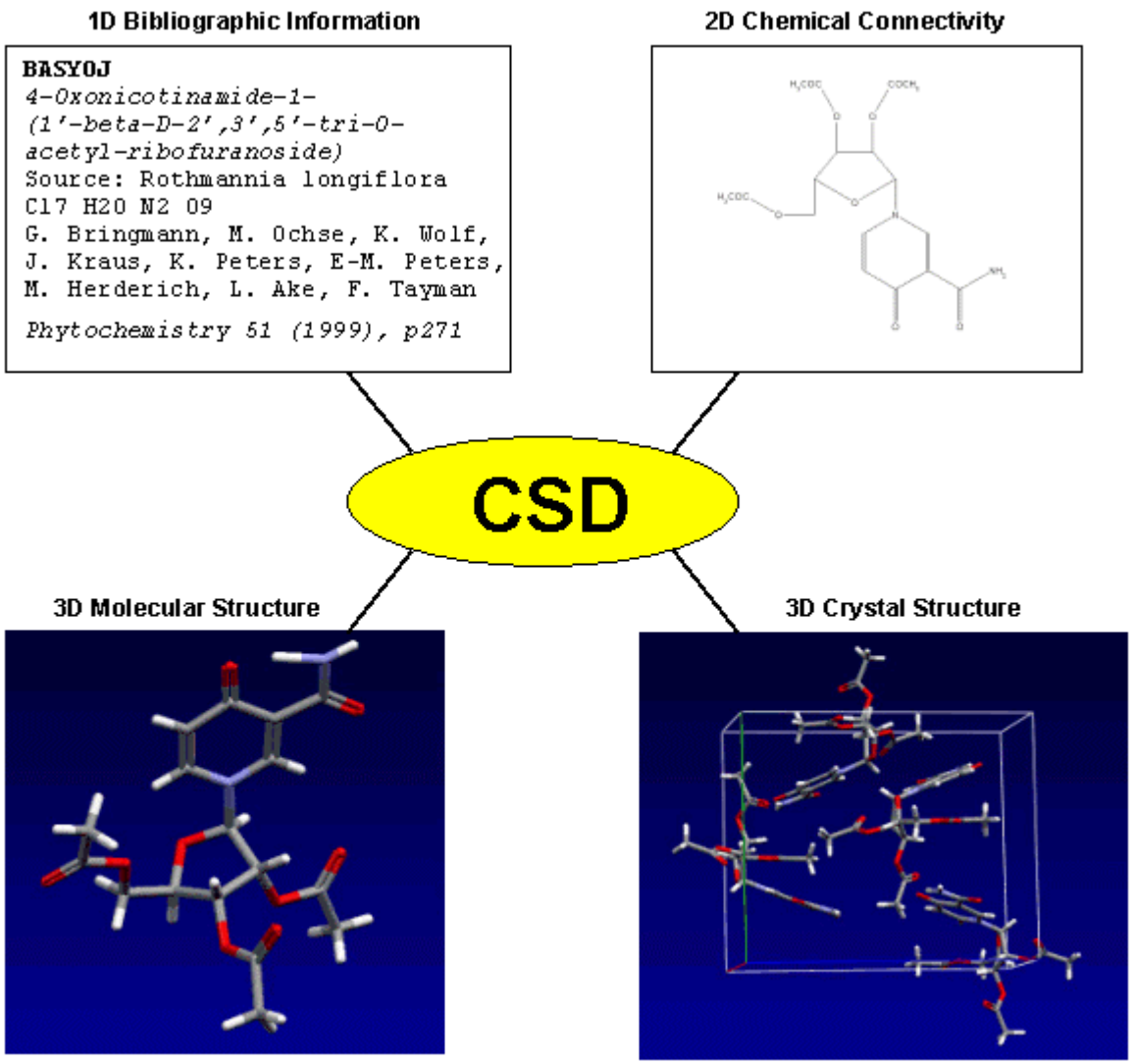
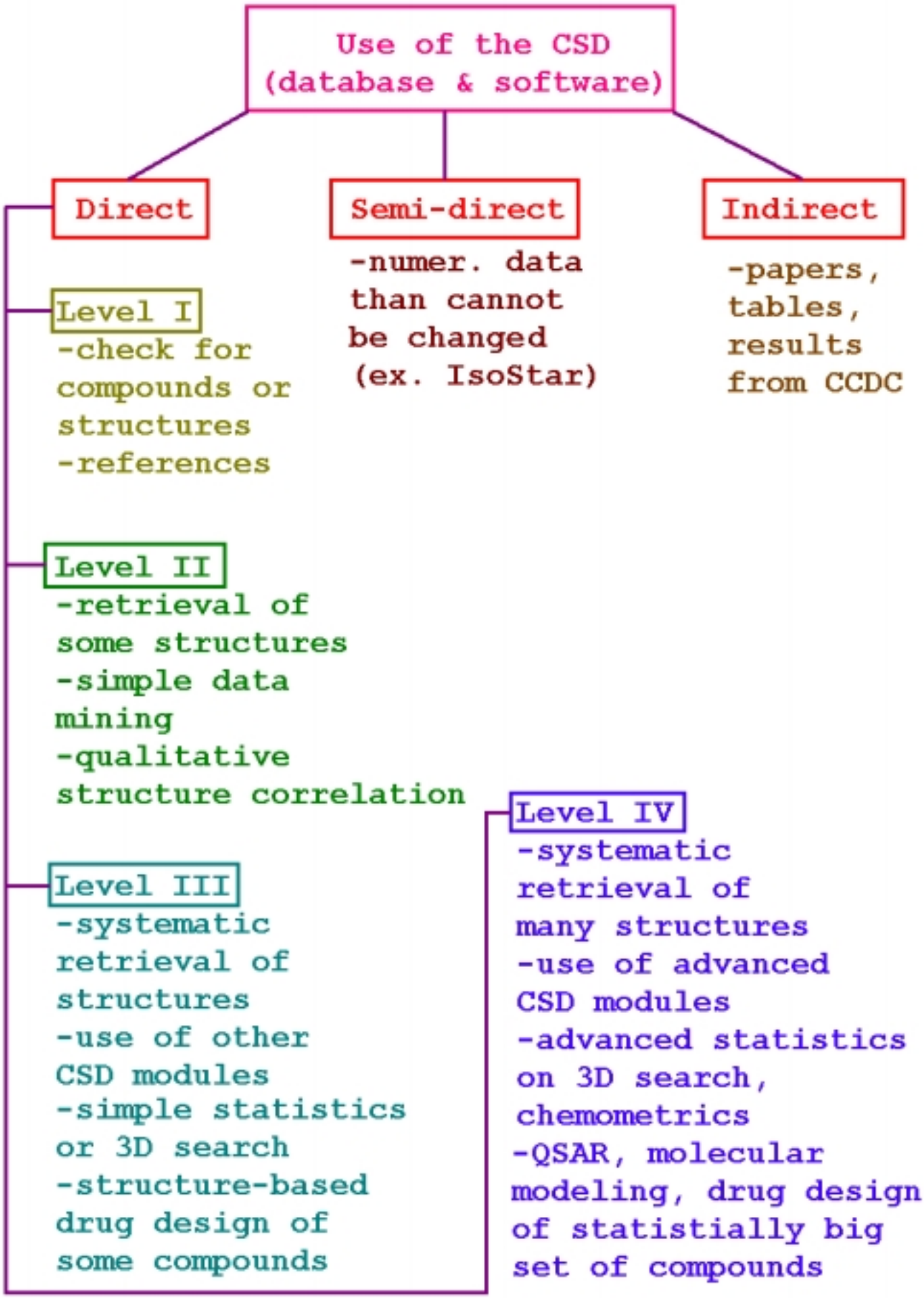


Figure 2. Types of information in the CSD.
Source: the CSD home page.


Use of the CSD
(database & software)
Direct
Semi-direct
Indirect
-numer. data than cannot be changed (ex. IsoStar)
-papers, tables, results from CCDC
Level I
-check for compounds or structures
-references
Level II
-retrieval of some structures
-simple data mining
-qualitative structure correlation
Level III
-systematic retrieval of structures
-use of other CSD modules
-simple statistics or 3D search
-structure-based drug design of some compounds
Level IV
-systematic retrieval of many structures
-use of advanced CSD modules
-advanced statistics on 3D search, chemometrics
-QSAR, molecular modeling, drug design of statistially big set of compounds

The CSD use in Brazil – some statistical parameters

What are the CSD products free for Brazilian scientist?

http://www.ccdc.cam.ac.uk/prods/

CCDC Products

CSD
Cambridge Structural Database

ConQuest
New Interface to the CSD

QUEST
Search and Retrieval Program for the CSD

VISTA
Statistical Analysis of Geometric and Other Data

PreQuest
Creation of In-house Databases

Mercury
Crystal Structure Visualisation
Available for free download

RPluto
Graphical Display of Molecular and Crystal Structures

DBUse
Database of Publications using the CSD and Other CCDC Products

IsoStar
Knowledge Base of Intermolecular Interactions

SuperStar
Predicting Protein-Ligand Interactions

GOLD
Protein-Ligand Docking

Relibase +
Easy Searching of Protein-Ligand Complexes

DASH
Structure Solution from Powder Diffraction Data

- -free for Latin America (Affiliation centre: Instituto "Rocasolano" -CSIC, Madrid, Spain)
- -free to download
- -public
- -commercial

How many CSD licenses are in Brazil today?

The number of the licenses follows curvilinear growth (Figure 3). There is a characteristic geographical distribution of the licenses (Figure 4).

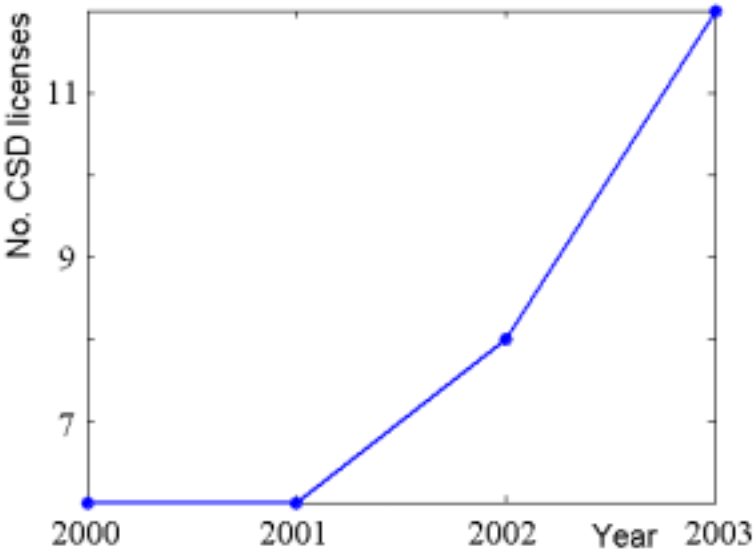


Figure 3. The growth of the CSD potential use in Brazil.

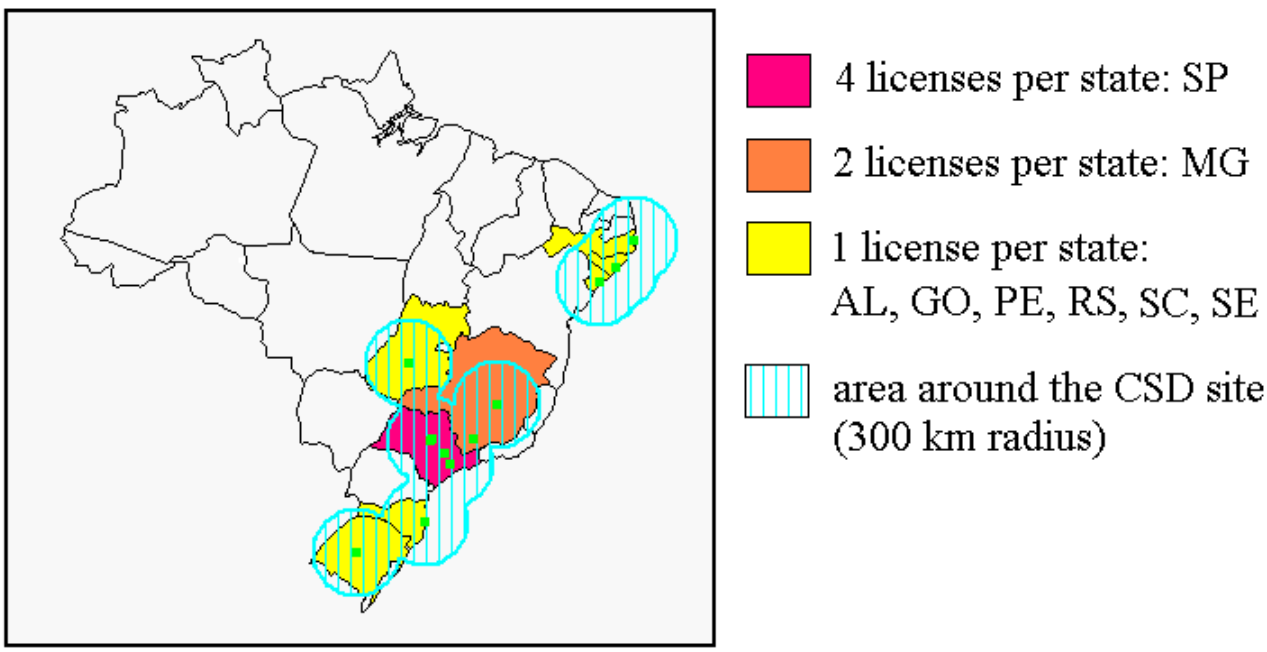


Figure 4. Geographical distribution of the 2003 CSD licenses in Brazil.

How much the CSD is used in Brazil?

-no reliable parameters

-probably the CSD is not much used in Brazil

-a tentative for the CSD use statistics for Brazil:

Country	Licenses 1992-1993	WDC10 (1997 ed.) Keyword: "database"	ISI Web of Sc. "Cambridge Structural Database" in title, keywords or abstract – 14/03/03
Argentina	7	0	2
Brazil	12	0	1
Chile	2	0	0
Colombia	4	0	0
Costa Rica	1	0	0
Cuba	3	0	0
Mexico	11	0	2
Peru	1	0	0
Uruguay	1	0	2
Venezuela	3	0	1
F. YU Countries	5	3	11
Hungary	3	0	14
Japan	75	3	19
Germany	59	7	43
UK	38	14	120
Russia	53	6	18
USA	228	5	91

The CSD in study of bond lengths: our study

Definition of the problem:

-the CSD enables data mining for many high-quality crystal structures of many classes of compounds

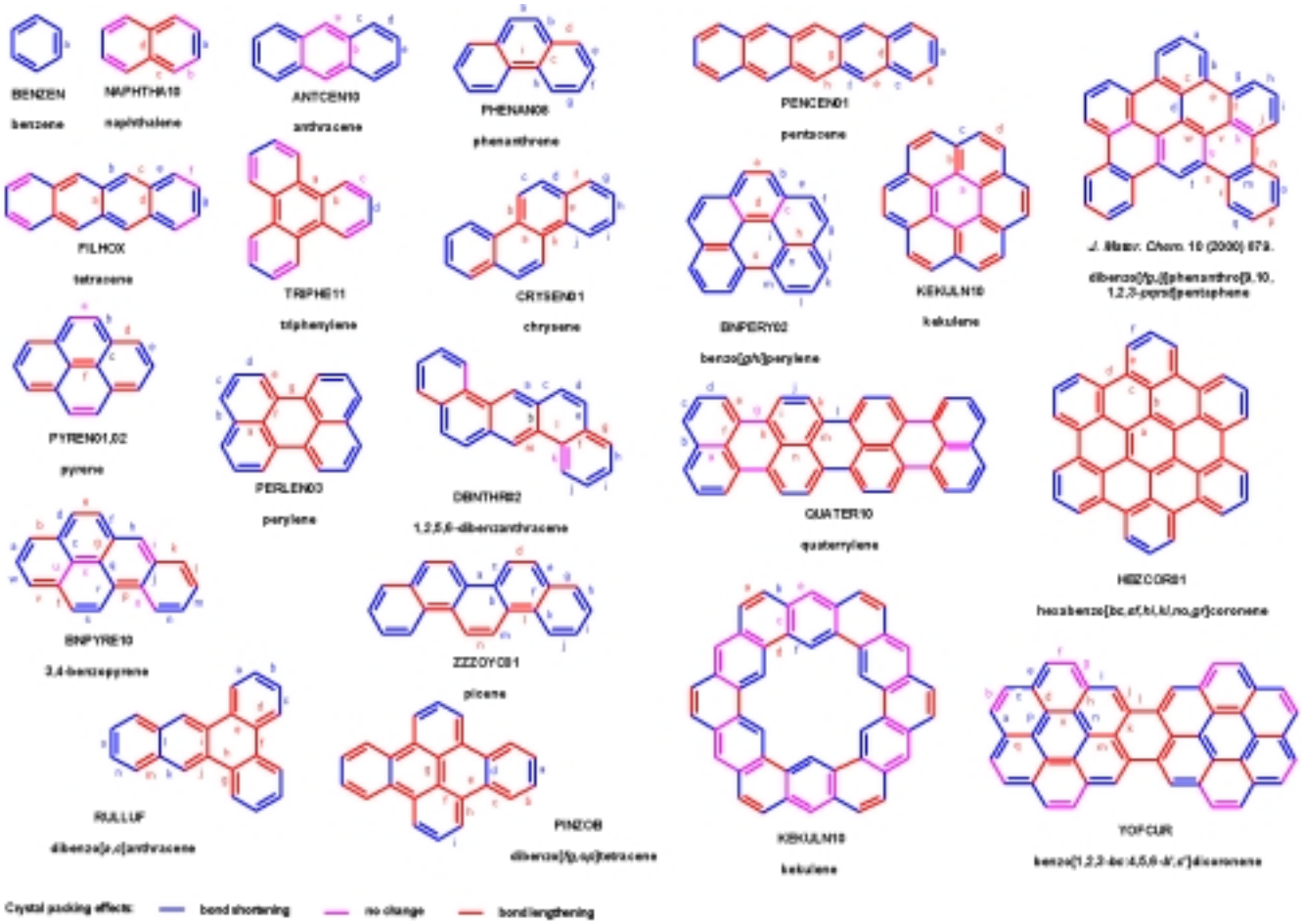
-questions: What is a chemical, especially covalent bond? In organic molecules? What is a C-C bond, especially partial double, aromatic? What does it depend on? Can it be predicted in various chemical problems?

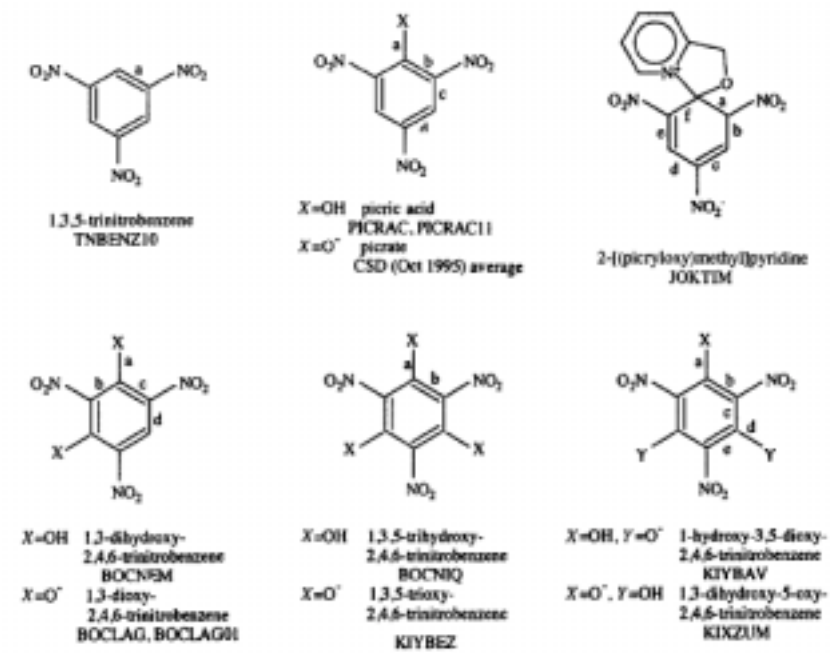
Typical classes of organic (C-H-N-O) compounds:

-simple organic compounds mainly with single C-C bonds
-planar benzenoid polycyclic aromatic hydrocarbons (PB-PAHs) and their aza-, diaza- and polyaza-derivatives
-picrate-like systems (picrate derivatives, picrates)
-nucleobases (nucleic acid bases)

Search for quantitative relationships (equations):

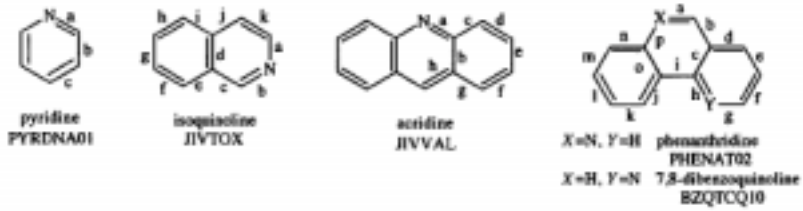
-quantitative relationships between experimental bond lengths in crystalline state and bond descriptors (the Pauling π-bond orders, other bond orders, topological descriptors, crystal packing parameters)



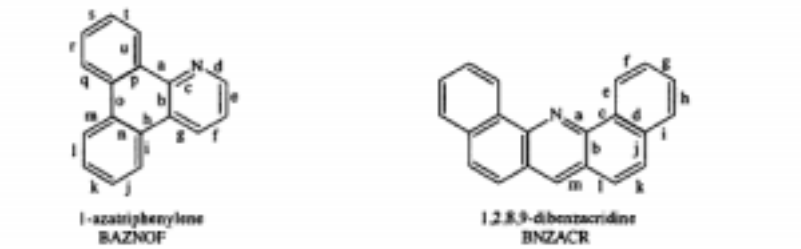


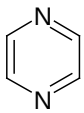
PB-PAHs 2002:
R. Kiralj, M. M. C. Ferreira, *J. Chem. Inf. Comput. Sci.*, **42** (2002) 508-523.

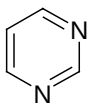
Picrates:
R. Kiralj *et al.*, *Acta Cryst.*, **B52** (1996) 823-837.

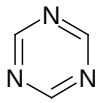


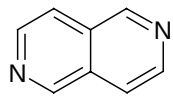
Aza-PAHs:
R. Kiralj *et al.*, *Acta Cryst.*, **B52** (1996) 823-837.

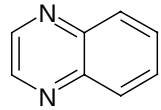


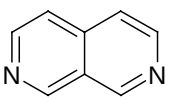

pyrazine
PYRAZ01

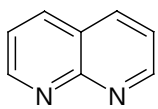

pyrimidine
PRMDIN01

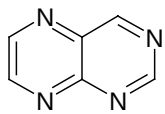

s-triazin
TRIZIN

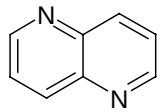

2,6-diazanaphthalene
DIAZNP

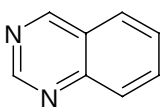

quinoxaline
YAWJIP

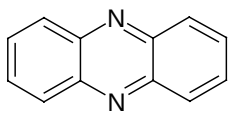

2,7-diazanaphthalene
DAZNAQ

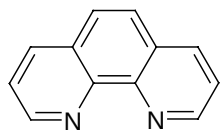

1,8-naphthyridine
NAPTYR, NAPTYR11

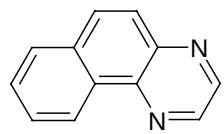

pteridine
PTERID11

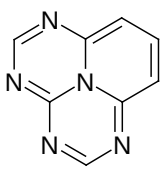

1,5-diazanaphthalene
DAZNAR

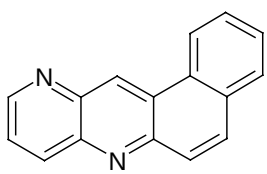

quinoxaline
QUINAZ

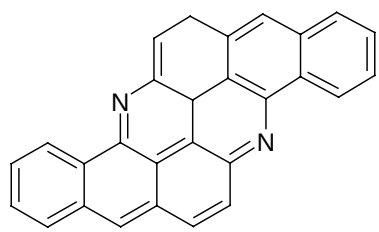

phenazine
PHENAZ04


1,10-phenanthroline
AMPCLP

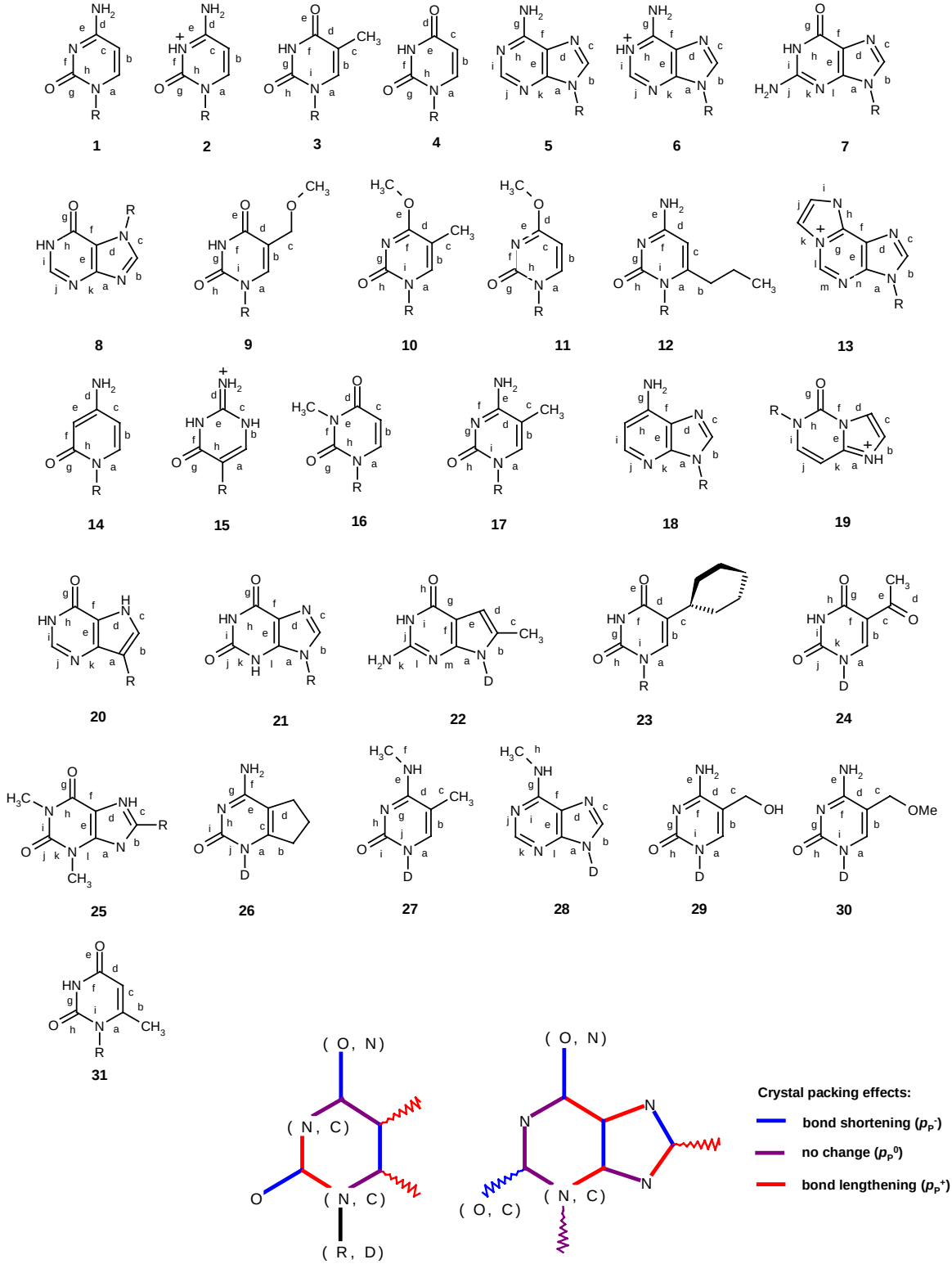

1,4-diazaphenanthrene
BEZSAA


1,3,4,6-tetra-
azacycl(3.3.3.)azine
TAZCAZ


naphtho(2,1-b)(1,5)
naphthyridine
VERMOU


flavathrene
BEFTAH

Diaza- and polyaza-PAHs:
R. Kiralj *et al.*, *J. Mol. Struct. -Theochem.*, **427** (1998) 25-37.



Nucleobases 2003:
R. Kiralj, M. M. C. Ferreira, *J. Chem. Inf. Comput. Sci.*, in press.

Simple organic compounds, carbon allotropes and molecular complexes from the CSD, spectroscopy, proper *ab initio* calculations, and other literature sources: M. M. C. Ferreira, R. Kiralj, in preparation.

Hydrocarbons:

-ethane
-ethylene
-acethylene
-egzohydrogenated (zig-zag) nanotubes (7,0) and (12,0)
-*n*-nonane
-allene
-dodecahedrane
-cyclopentadienyl anion
-cyclobutane
-cyclohexane
-hexabenzocoronene (its central ring)
-benzene
-(PB-PAHs)

Carbon allotropes:

-fullerenes C_{60} and C_{70}
-carbyne
-hexagonal and rhombohedral graphite
-cubic diamond
-FCC metallic carbon
-acetylide anion

Complex species:

-hexachlorobenzene
-$(pyrene)_2$, pyrene-TCNQ, pyrene-$(DOP)_2$
-$(TCNQ)_2$, $(TCNE)_2$, $(DDQ)_2$

Data mining strategy

	PB-PAHs	Aza-PAHs	Diaza-, Polyaza-PAHs	Picrates	Nucleo-bases	Simple organic comp. etc.
R	≤ 7%	≤ 7%	≤ 7%	≤ 7%	≤ 6%	≤ 7%
σ/Å	≤0.015	≤0.015	≤0.015	≤0.015	≤0.005	≤0.005
Year	any	any	any	any	≤1975	any
disorder	no	no	no	no	no	no
errors	no	no	no	no	no	no
atm type	CH	CHN	CHN	CHNO	CHNO	CHNOX
org.met.	no	no	no	no	no	no
met. com	no	no	no	no	no	no
planar	yes	yes	yes	any	any	any
excluded bonds	-	-	N-N	-	N-N N-O O-O	N-N O-O
retrieved species	PB-PAHs	aza-PAHs	diaza-, polyaza-PAHs	picrates	nucleo-sides	hydro-carbons C allotr. mol.com.

General C-C bond

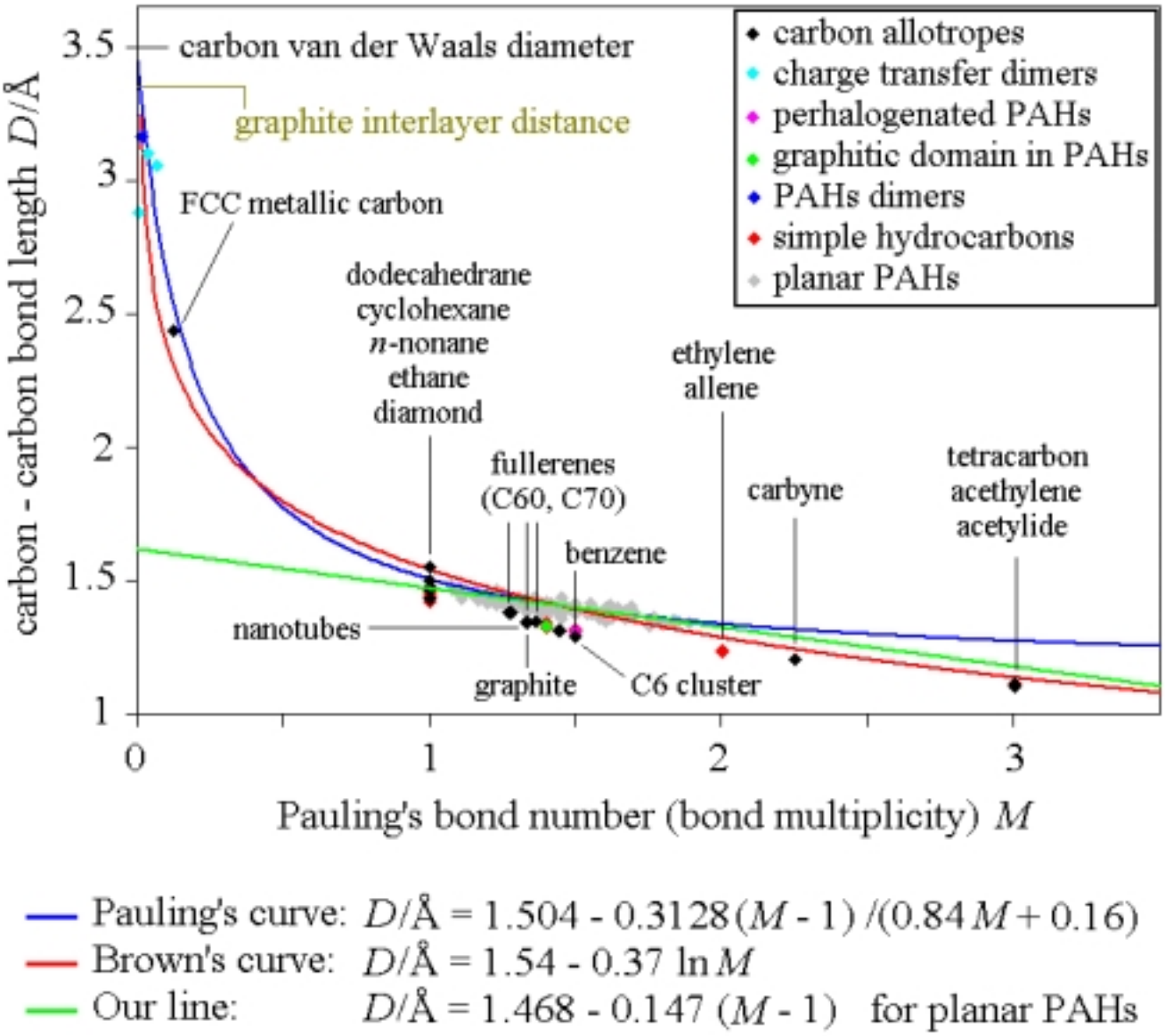


Figure 5. C-C bond length as a function of the Pauling's bond multiplicity, from acethylene and acetylide anion to saturated hydrocarbons and diamond, and even more to intermolecular complexes bound by weak C-C bonds. Bond orders of other types give very similar results.

Novoa *et al*. discovered the longest C-C bond even known: an electron defficient bond in donor-acceptor complexes, with very small bond number, and bond lengths up to the graphite interlayer distance.

Various carbon compounds, allotropes and species are spread over the whole range of the bond number, along the Pauling's and Brown's curves, supporting the new findings on the longest C-C bonds.

This discovery shakes the old concept of the C-C bond: "nonbonding interatomic" or "intermolecular interactions" should be considered as intemolecular chemical bonds.

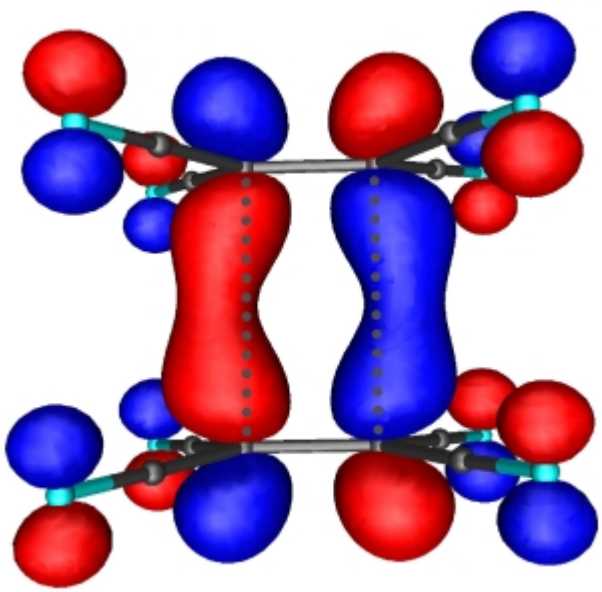


Figure 6. J. J. Novoa *et al*., *Angew. Chem. Int. Ed*., **40** (2001) 2540. $[TCNE]_2^{2-}$ complex is an example of the longest C-C bonds with lengths 2.8 – 3.5 **Å**. Four C atoms share two electrons (*M* < 0.5). Del Sesto *et al*., *Chem.-Eur. J*., **8** (2002) 4894; J. J. Novoa *et al*., *Cryst. Eng. Comm*., (2002) 373.

Univariate Structure Corrrelation for Bond Lengths in π-Systems: PB-PAHs, aza-PAHs, diaza-PAHS, polyaza-PAHs, picrates, nucleobases

Relationships between structural parameters:

-Structure correlation (SC)

-(Quantitative) Structure-Structure Relationships ((Q)SSR)

-Bond Length-Bond Order Relationships (BLBOR)

-Bond Length-Bond Descriptor Relationships (BLBDR)

Our study: Univariate relationships (linear regression) between bond lengths D and Pauling π-bond orders P ($P = M - 1$): $D/\text{Å} = a + b\,P$

a and b exhibit expected similarities and differences:

		a	b
C-C	**PB-PAHs**	**1.468(2)**	**-0.147(5)**
	Aza-PAHs	**1.462(6)**	**-0.143(13)**
	Diaza-PAHs	**1.458(3)**	**-0.143(8)**
	Polyaza-PAHs	**1.421(30)**	**-0.087(81)**
	Picrates	**1.497(11)**	**-0.212(25)**
	Nucleobases	**1.487(7)**	**-0.202(16)**
C-N	**Aza-PAHs**	**1.444(10)**	**-0.184(18)**
	Diaza-PAHs	**1.415(4)**	**-0.152(8)**
	Polyaza-PAHs	**1.431(20)**	**-0.128(42)**
	Nucleobases	**1.398(3)**	**-0.127(7)**
C-O	Picrates	1.326(5)	-0.198(18)
	Nucleobases	1.295(16)	-0.101(26)
All	Nucleobases	1.429(5)	-0.199(13)

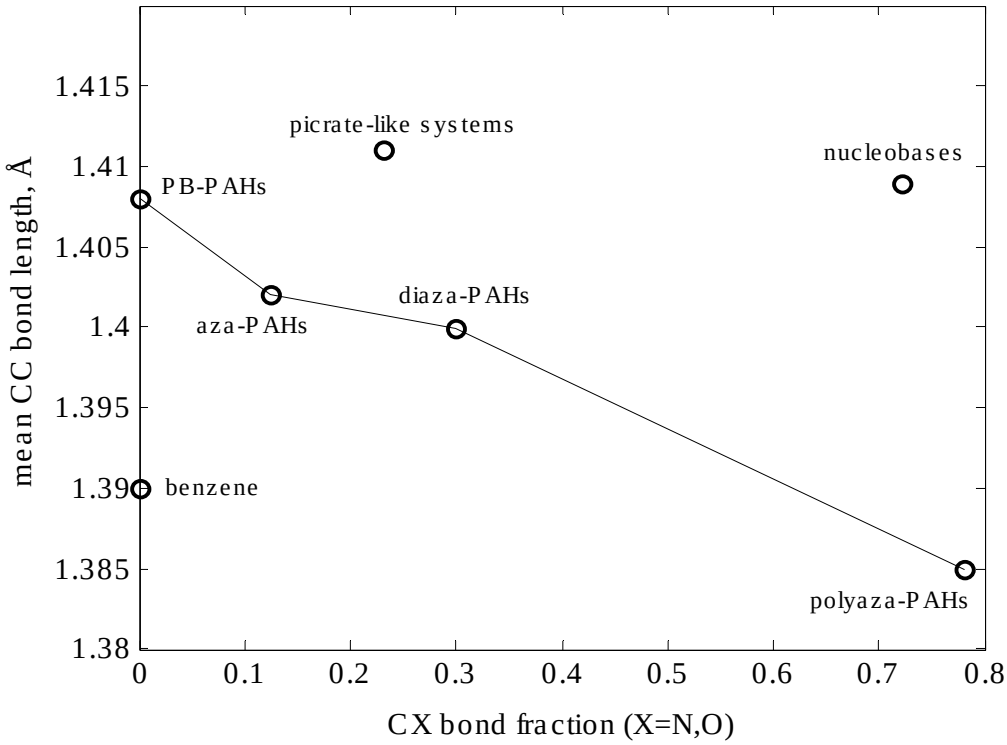


Figure 7. The mean C-C bond length in various π-systems.

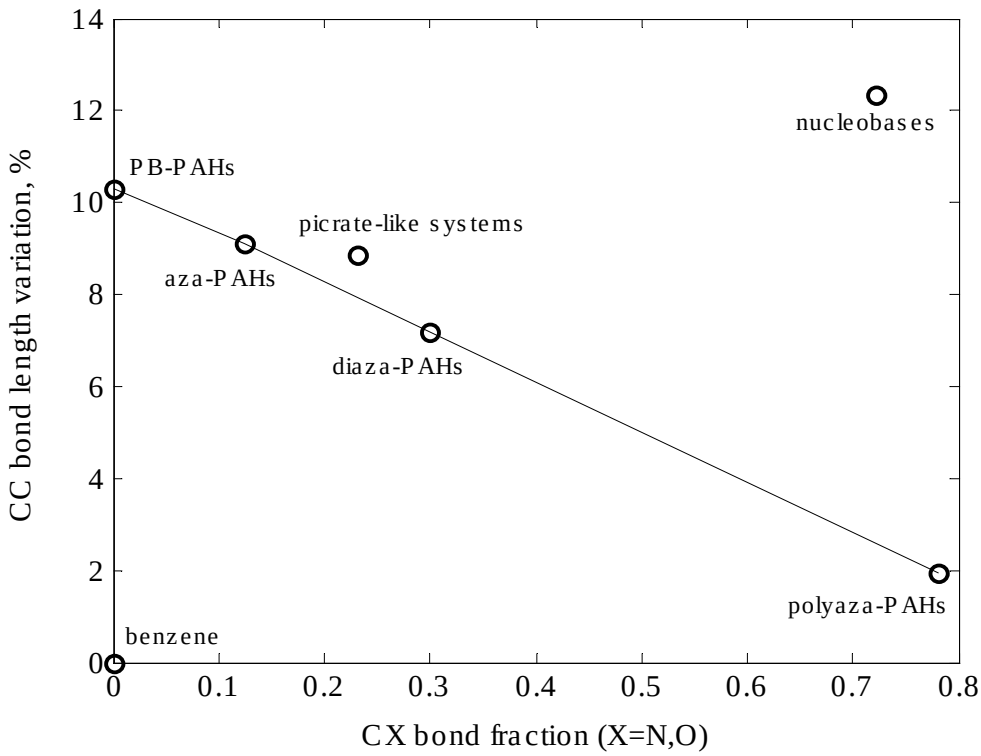


Figure 8. The C-C bond length variation in various π- systems.

Multivariate Approach to Bond Length Prediction in π-Systems: PB-PAHs, nucleobases

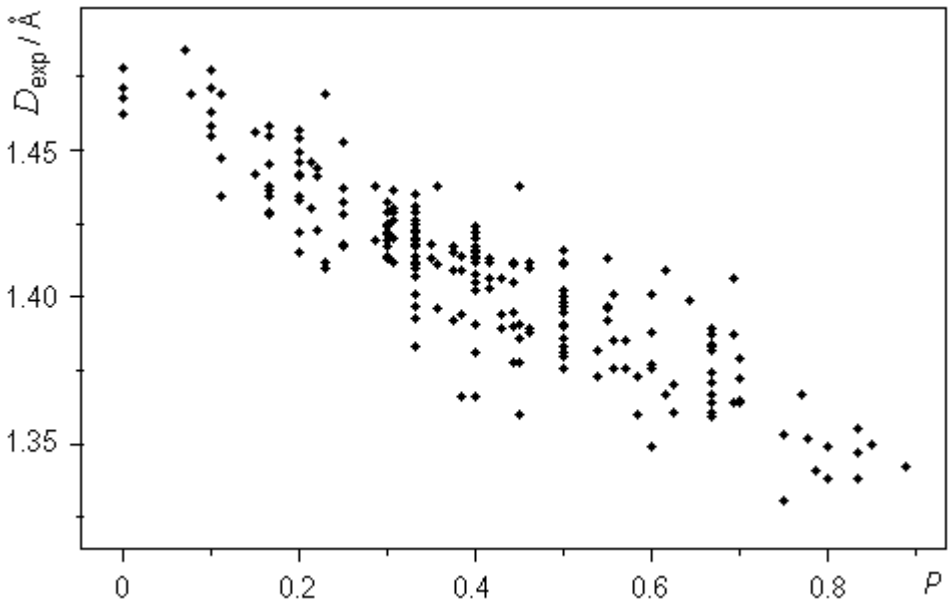


Figure 9. Degeneration of the $D - P$ graph for PB-PAHs.

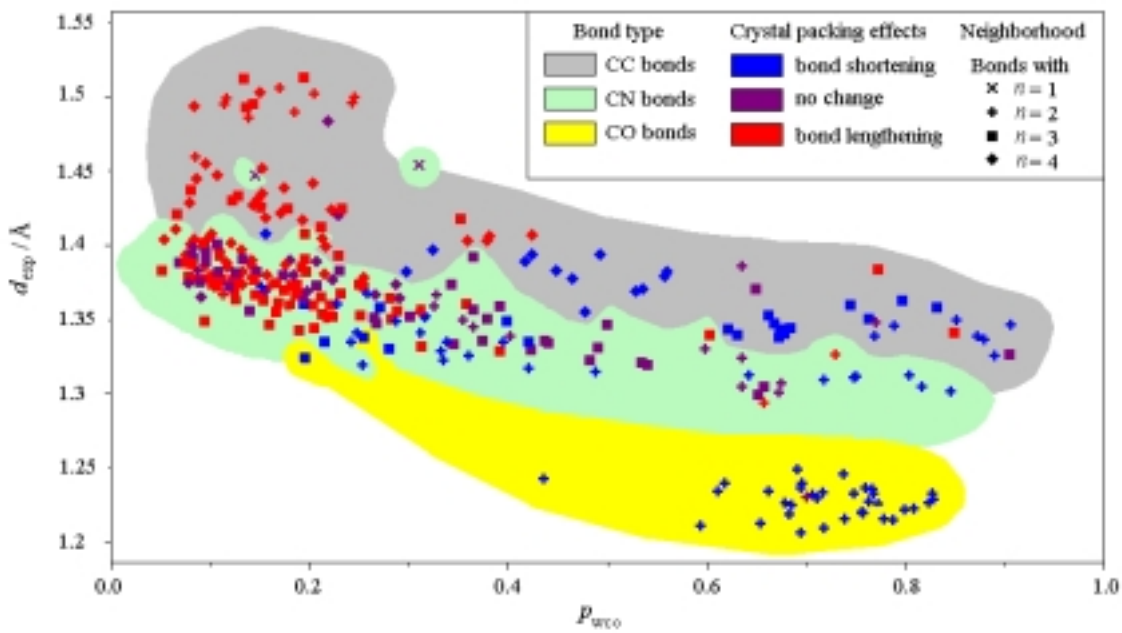


Figure 10. Degeneration of the $D - P_w$ graph for nucleobases.

Bond descriptors to decrease the data degeneration:

-Pauling π-bond order P

-weighted Pauling π-bond order P_w

-Pauling π-bond orders including crystal packing effects (H-bonds, vdW interactions, etc.) P_x

-electrotopological index: the sum of atomic numbers Q

-topological indices:

-the number of neighboring bonds n

-the number of neighboring rings m

-the number of neighboring bonds l around bonds already counted for n

PB-PAHs: P, various P_x, n, m, l

Nucleobases: P, P_w, various P_x, n, Q

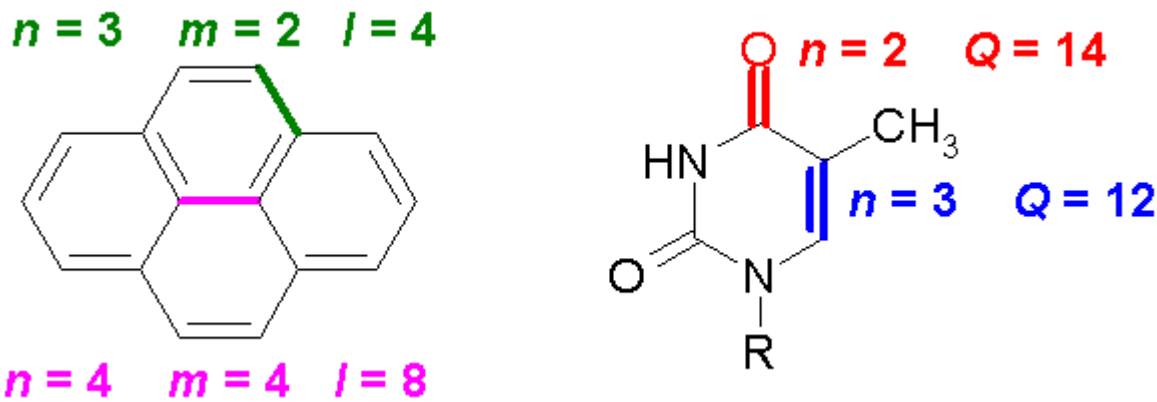


Figure 11. Some examples for (electro)topological indices

Results of the Principal Component Analysis (PCA)

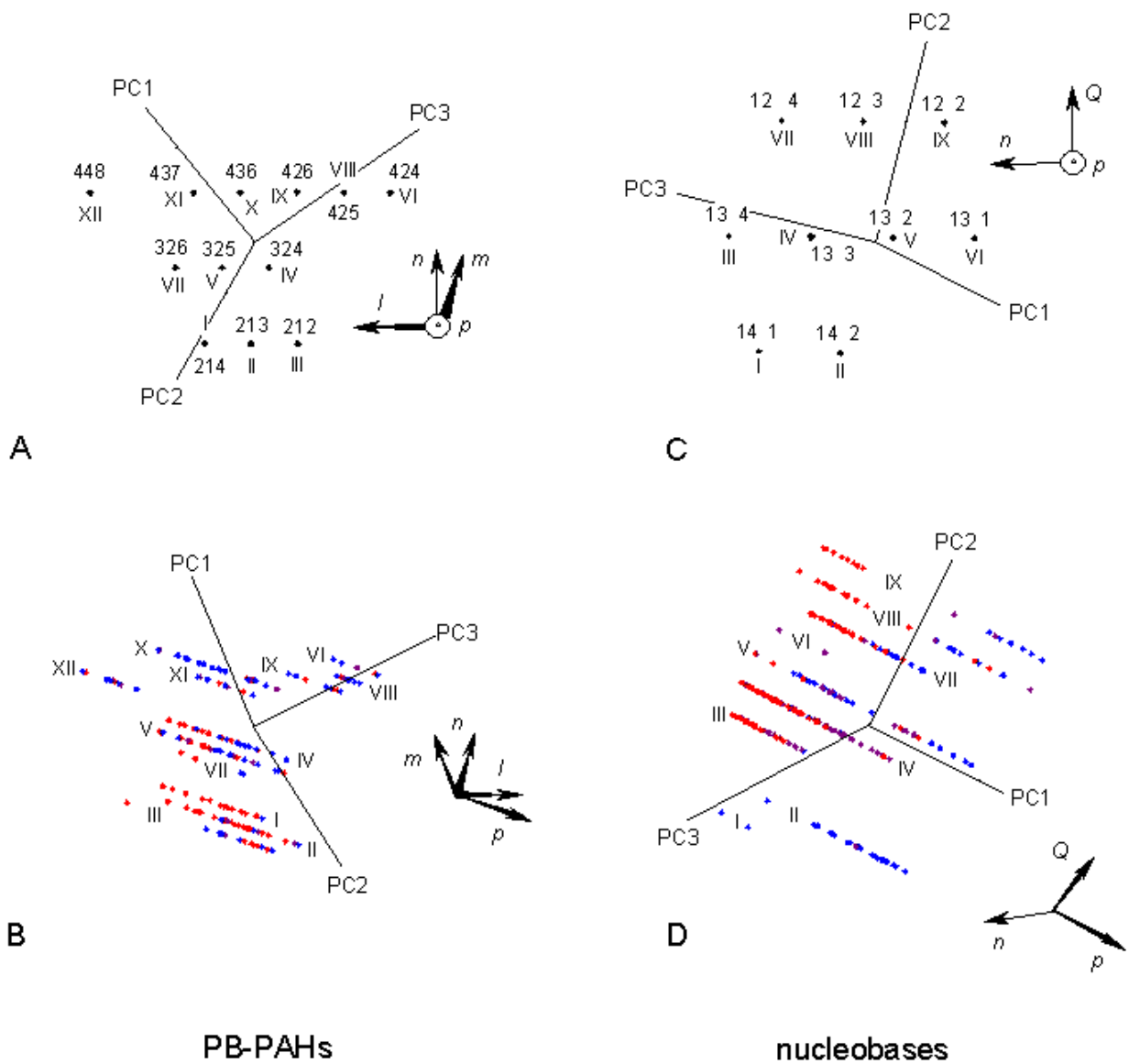


Figure 12. 3D scores plots from PCA show similarity and dissimilarity between PB-PAHs and nucleobases. Roman numerals denote various classes of bonds defined by topological indices: (*nml*) for PB-PAHs, (*Qn*) for nucleobases.

Results of the Hierachical Cluster Analysis (HCA)

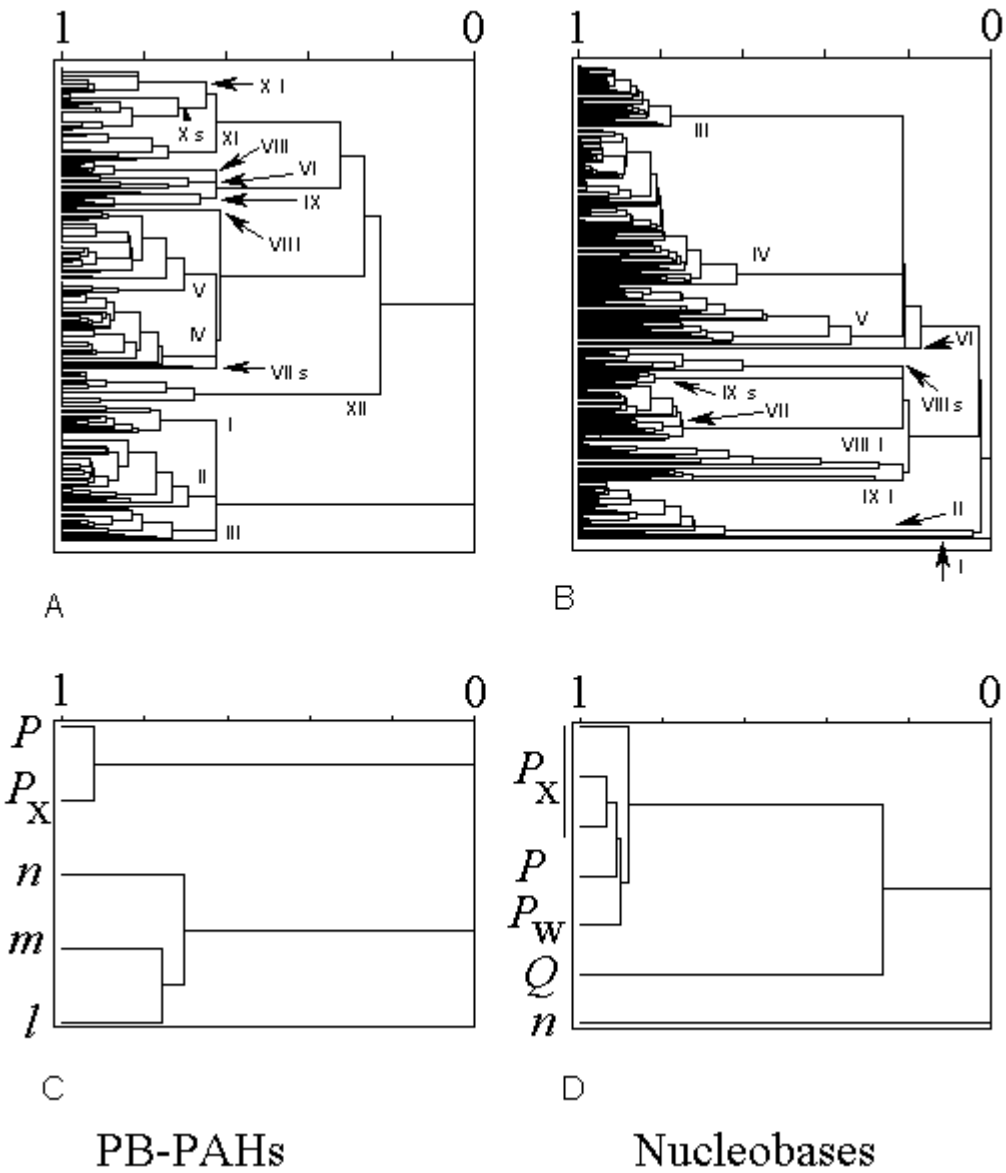


Figure 13. HCA dendograms for samples (A, B) and variables (C, D) for PB-PAHs and nucleobases. The groups from the 3D PCA scores plots are marked in the same way, adding **s** (for short bonds) and **l** (for long bonds) for subgroups of VII and X (PB-PAHs), and VIII and IX (nucleobases).

Multivariate Bond Length-Bond Descriptor Relationships

Partial Least Squares (PLS) regression showed to be more adequate than linear regression for the study of C-C, C-N and C-O bond lengths in PB-PAHs and nucleobases.

Much better PLS regression models were obtained using more bond variables than only Pauling´s bond order *P*.

PB-PAHs:

$$D/\text{Å} = 1.431 - 0.060\,P - 0.063\,P_x + 0.006\,n + 0.004\,m + 0.001\,l$$

R= 0.94, Δ = 0.007 Å → univariate: R = 0.90, Δ = 0.010 Å
PLS: 2 principal components (>96% original data)

Nucleobases:

$$D/\text{Å} = 2.304 - 0.080\,P - 0.078\,P_x - 0.068\,Q - 0.006\,n$$

R= 0.93, Δ = 0.017 Å → univariate: R = 0.65, Δ = 0.037 Å
PLS: 3 principal components (>98% original data)

Future Perspectives: CSD Use With Other Methods

DATA MINING +

CHEMOMETRIC ANALYSIS +

STRUCTURAL & COMPUTATIONAL METHODS

=

A VERY POWERFUL MEANS TO STUDY BOND LENGTHS IN ORGANIC CRYSTALS +

INTERESTING AND USEFUL RESULTS ON BASIC CHEMICAL CONCEPTS (WHAT IS A BOND?? C-C BOND??) +

DIRECTIONS FOR FUTURE STUDIES (intrinsic π-system properties, substitution effects, crystal packing effects, (hetero)aromaticity, crystal packing, etc.)

THIS IS VALID NOT ONLY FOR BOND LENGTH BUT FOR OTHER STUDIES WHICH USE THE CSD EXTENSIVELY.