Determination of the charge of molecular fragments by machine learning methods

A. P. Shevchenko[1,2], A. V. Chuvakov[1], N. A. Latyshev[1]

[1] Samara State Technical University, Samara, Russian Federation

[2] Samara Branch, P.N. Lebedev Physical Institute of the Russian Academy of Sciences, Samara, Russian Federation

Supported by the RSF № 23-23-00387

Content

"Our intelligence is what makes us human, and AI is an extension of that quality." – Yann Le Cun

- Introduction
- History of the problem
- Molecular Fragment Descriptors
- Scheme for solving a prognostic problem
- Data sampling and software tools
- Results and their use
- Conclusion

Introduction

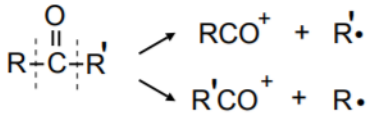
Rule-Based Systems

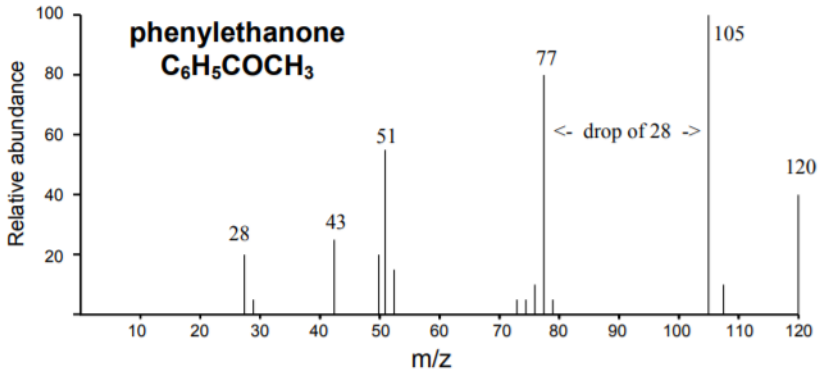
Rules based systems are *deterministic* in nature. Rule-based systems can often *offer quicker, tactical solutions and workarounds*. The requirement for business expert input can help wider business buy-in and make *it easier to explain* how decisions were made. Many projects begin with an expert or rule-based system to *explore* and *understand* the *system*.

Aldehydes & Ketones

Cleavage of bonds next to the carbonyl group (C=O) is a characteristic fragmentation of aldehydes and ketones. A **common fragment is carbon monoxide (CO)** but as it is a molecule and thus uncharged it will not produce a peak of its own. However, it will produce an **m/z drop of 28** somewhere in the spectrum.

The position of the carbonyl group influences the fragmentation pattern because the molecular ion fragments either side of the carbonyl group.





Peaks	
120	molecular ion
105	$C_6H_5CO^+$
77	$C_6H_5^+$
51	$C_4H_3^+$
43	CH_3CO^+
28	CO

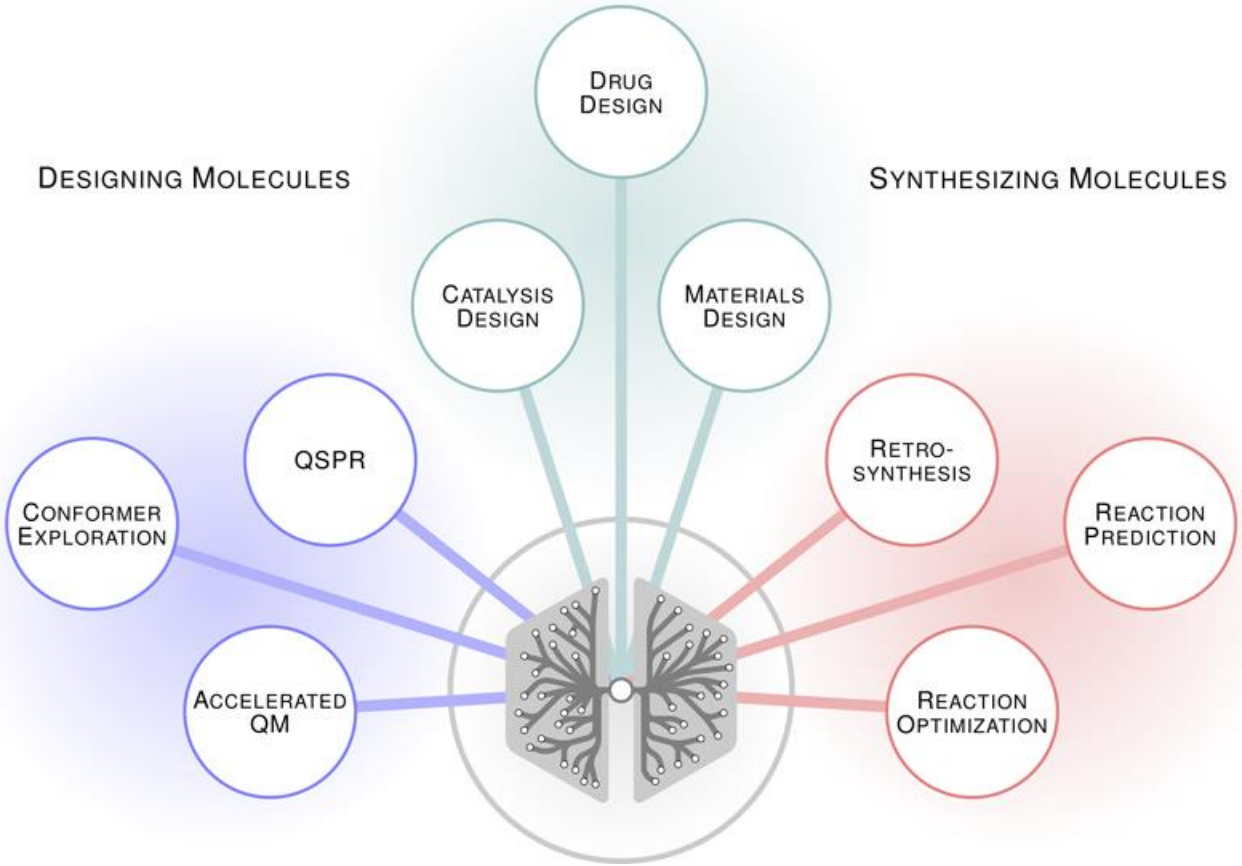
https://en.wikipedia.org/wiki/Dendral

DENDRAL: One of the *first* expert systems (**the 1960s**). The task was to help chemists in determining the molecular structure of an organic compound, from mass spectrometer data.

If there are two peaks in the spectrum at masses X1 and X2
and *X1 + X2 = M + 28*
and *X1 - 28* and *X2 - 28* are high
and at least one of X1 and X2 is high
then *a Ketone group is present*

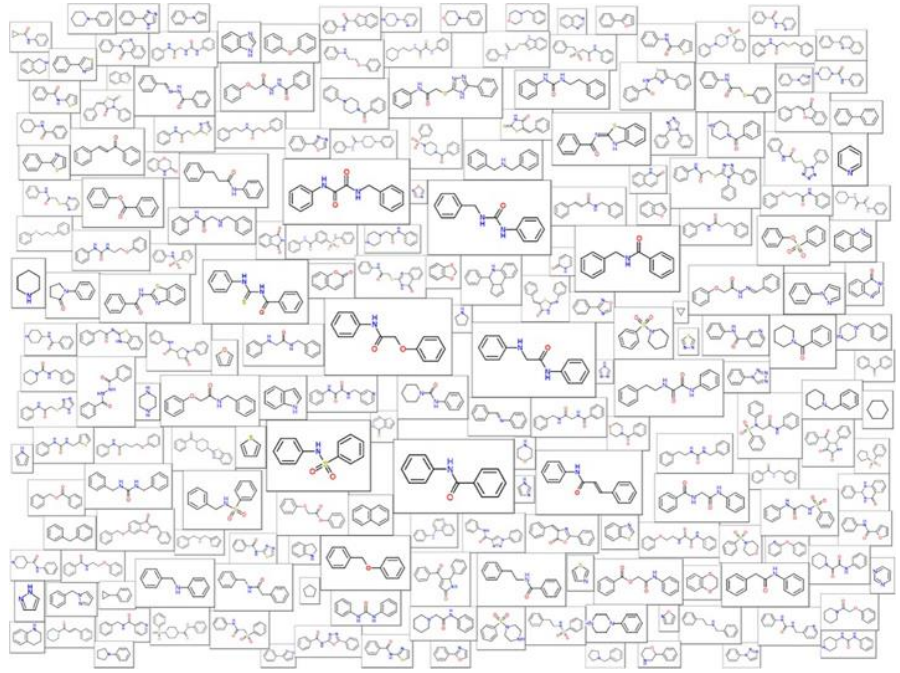
Introduction

Deep Learning and Chemistry

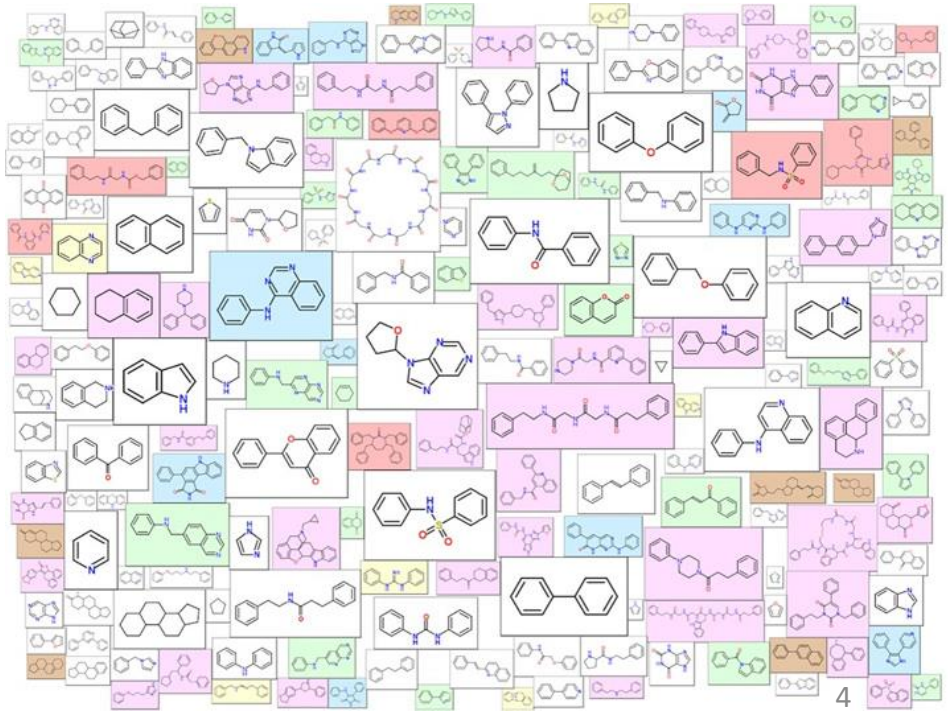


J. Chem. Inf. Model. 2019, 59, 6, 2545–2559

ZINC database



ChEMBL database



Journal of Cheminformatics 2012, 4:12

History of the problem

Statement of the problem and why it is necessary to solve it

Determination of the charge of molecular fragments (MF) in the crystal structure

- An important descriptor for MF being the structural building unit of a crystal;
- Determination of the degree of oxidation of the complexing atom;
- Intelligent assembly of crystals from structural building units.

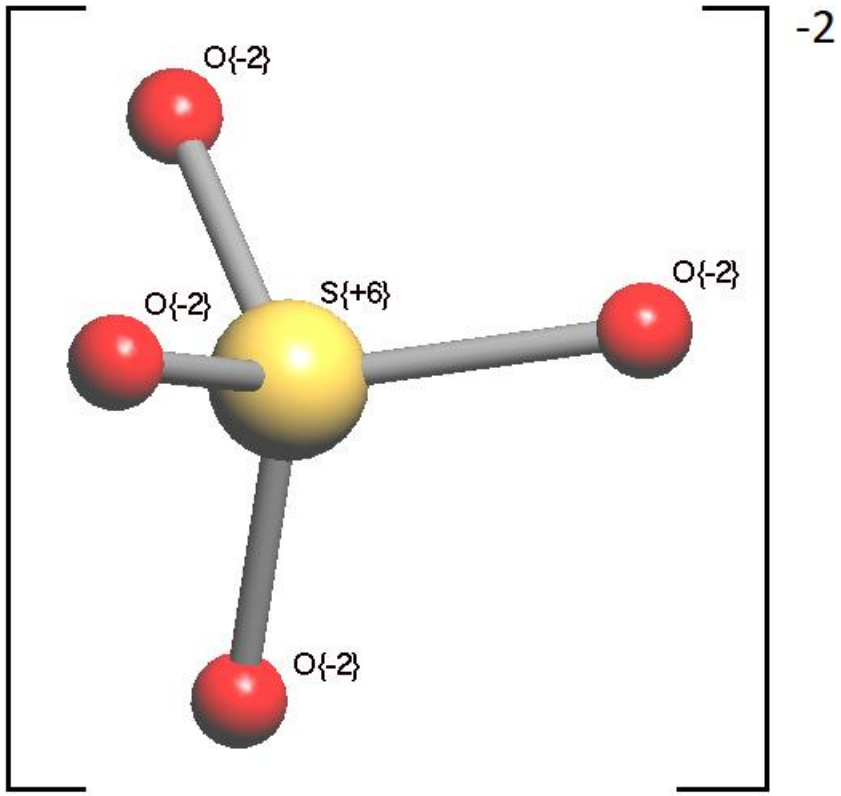
History of the problem

How is the charge of MF determined?

- According to the structural formula MF
- According to the sum of the oxidation states of the atoms that make up the MF
- According to the balance of charges of ligands, outer-sphere particles and metal atoms, that is the oxidation state, in the crystal structure
- According to the chemical name of the MF

History of the problem

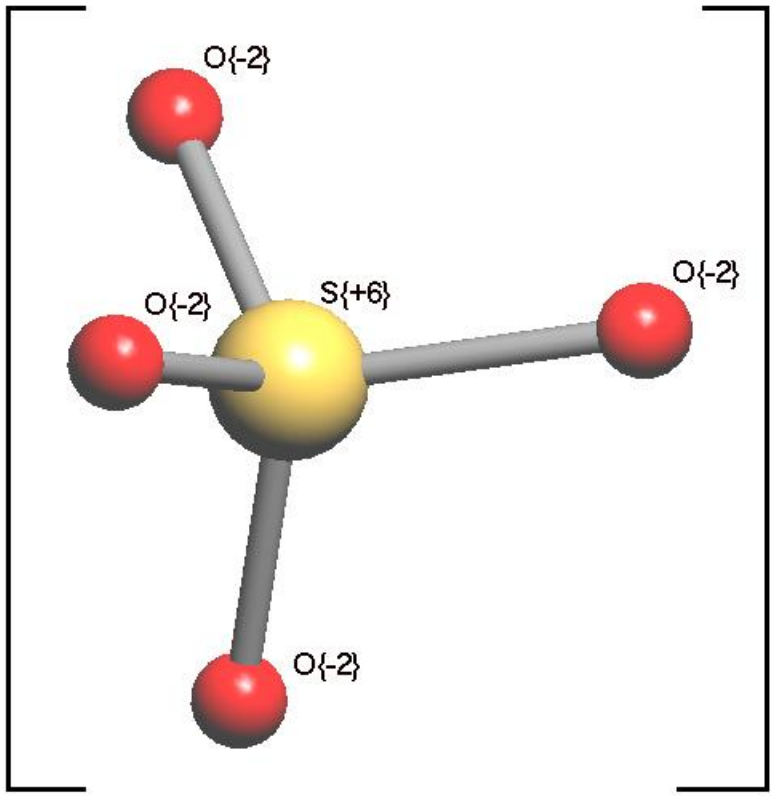
How is the charge of MF determined?



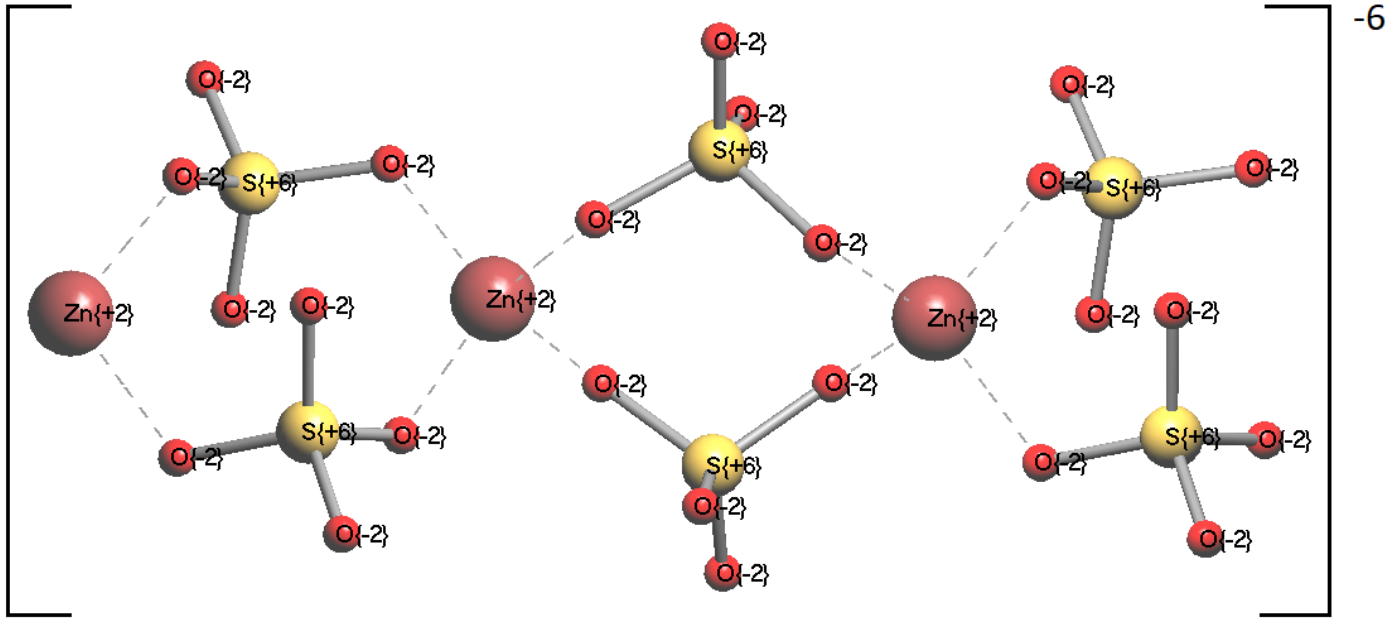
Ligand

History of the problem

How is the charge of MF determined?



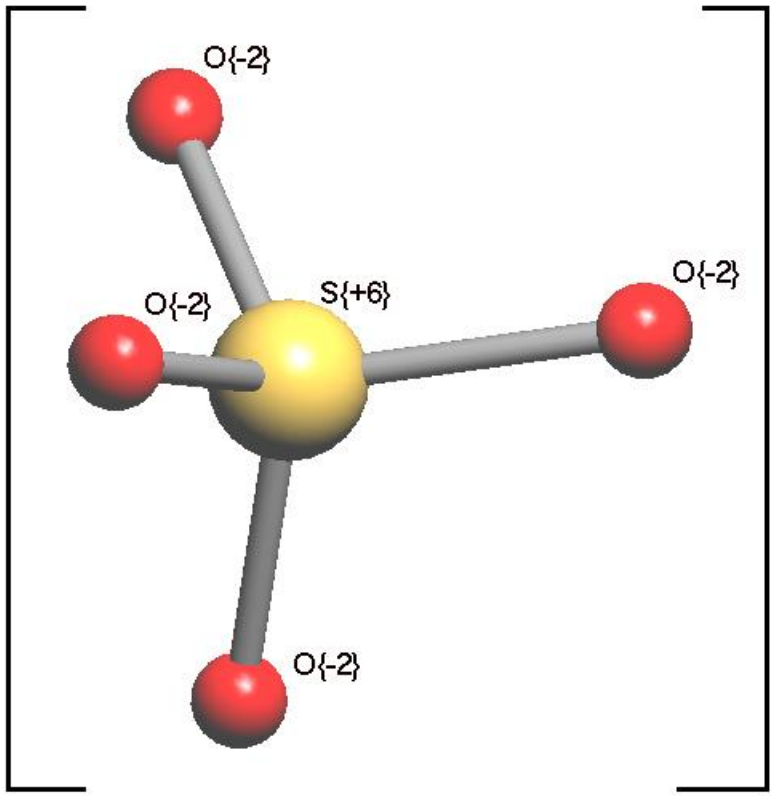
Ligand



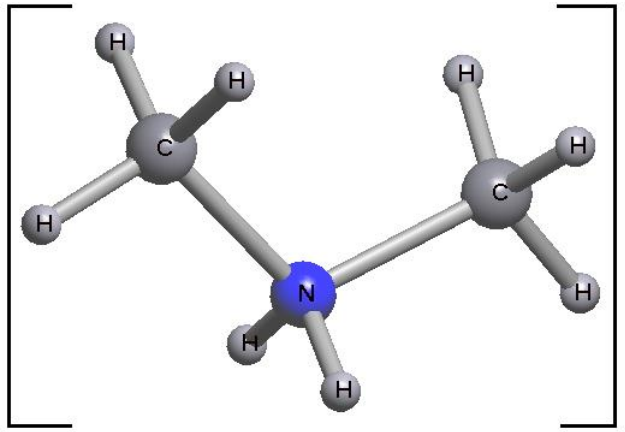
Chain

History of the problem

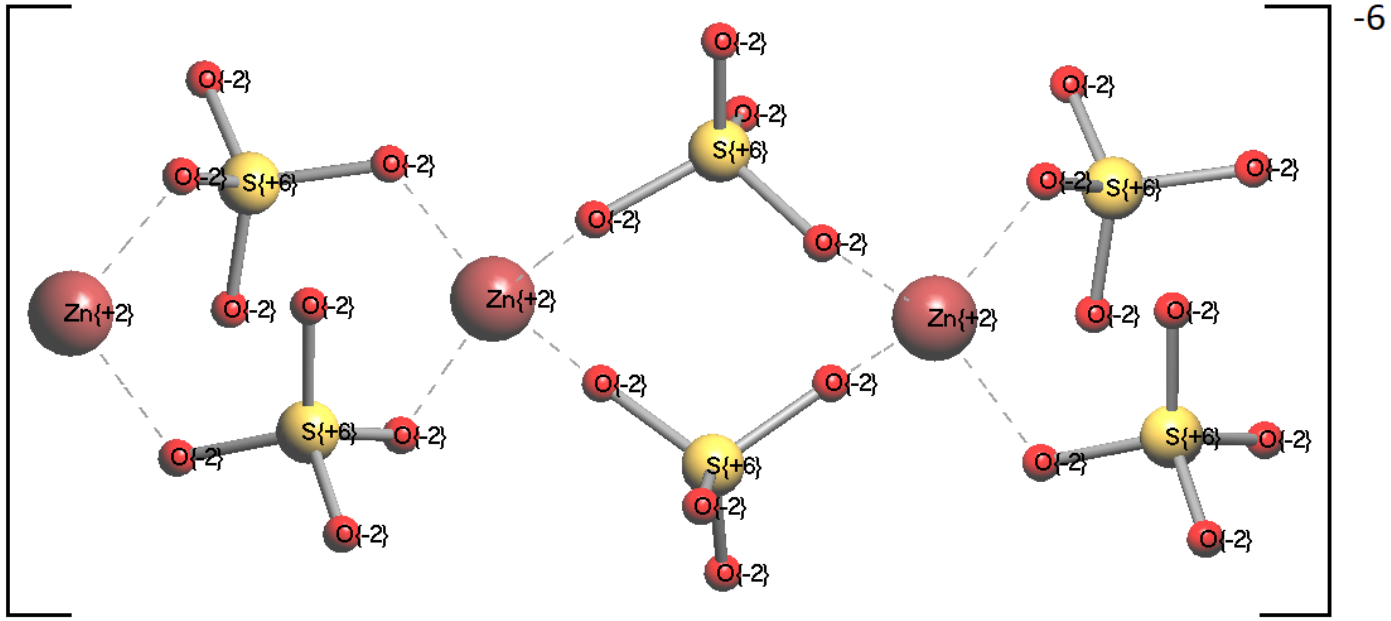
How is the charge of MF determined?



Ligand



Counterion

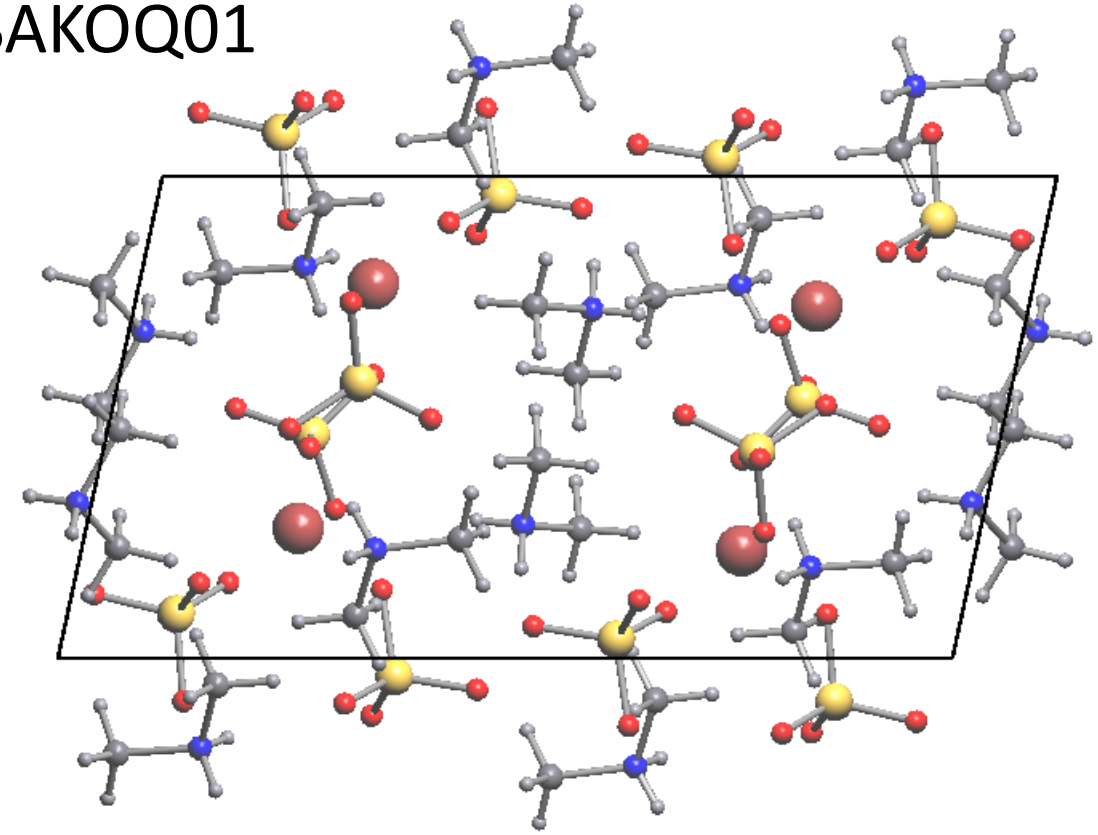


Chain

History of the problem

How is the charge of MF determined?

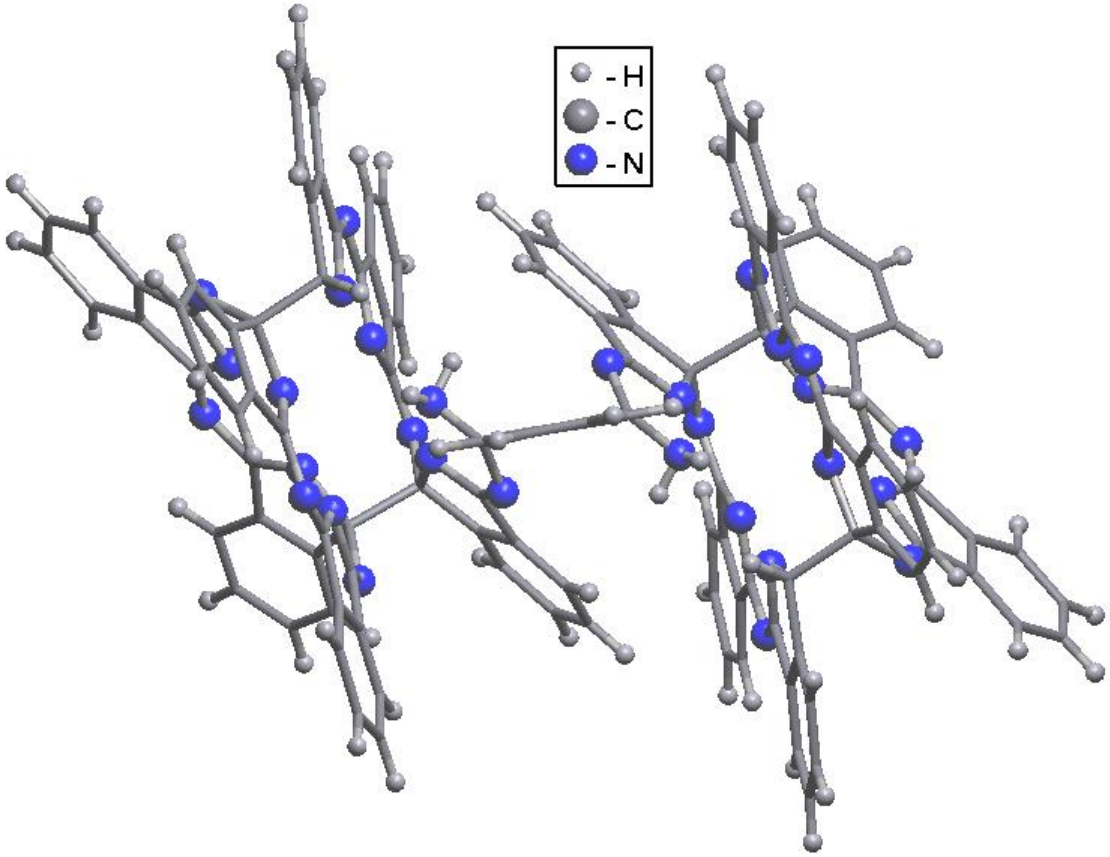
USAKOQ01



(C2H8N)2[Zn(SO4)2]

Z.Kristallogr.-New Cryst.Struct. 2021, **236**, 11.

JACSUF - C120H66N30_289071 {-6}

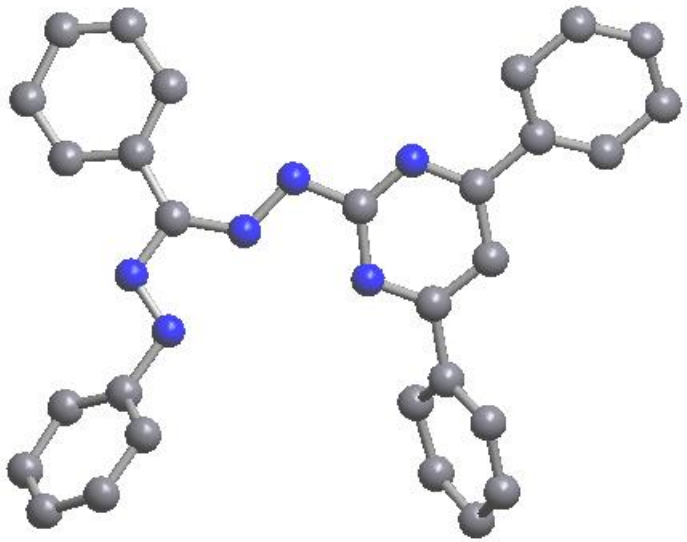


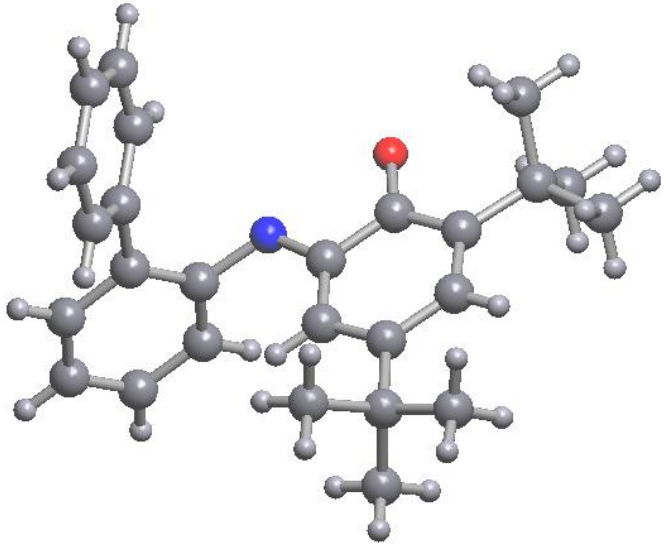
J.PORPHYRINS PHTHALOCYANINES 2017, **21**, 257

History of the problem

Difficulties in Determining the MF Charge in the Crystal Structure

- The complexity of the chemical composition and structure of MF
- Incompletely deciphered crystal structure (hydrogen atoms are not localized, some atoms are disordered over several positions, etc.)
- The presence of radicals that contain unpaired electrons





History of the problem

Methods for determining the MF charge for big data

- Calculation of the rigidity and energy of formation of the MF by the fast quantum method MOPAC (Stewart, 2016). For an MF with an optimal charge, the stiffness parameter should be maximum, and the formation energy should be minimum.
- Determination of the charge by comparison with the neutral form of MF, which is obtained by automatically adding the missing hydrogen atoms to it.
- Summation of positive and negative charges in MF smile.

Matthew G. Reeves, et al. Acta Cryst. (2019). B75

Molecular Fragment Descriptors

Electronic

- charge and multiplicity
- dipole moment
- polarizability

Geometric

- the total volume of the fragment
- total surface area
- conformational rigidity
- free space
- solid angle of contact

Chemical

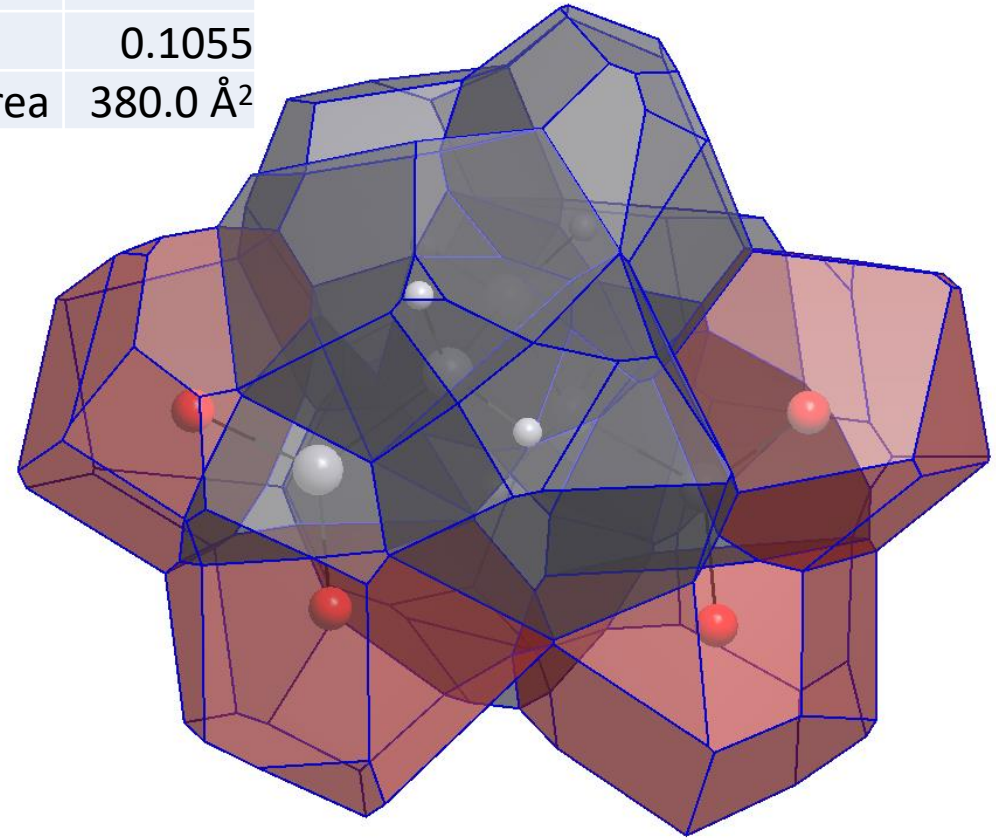
- elemental composition
- functional groups
- donors and acceptors of H-bonds
- metal coordinated to the ligand

Topological

- coordination number
- coordinating figure
- coordination sequences of atoms

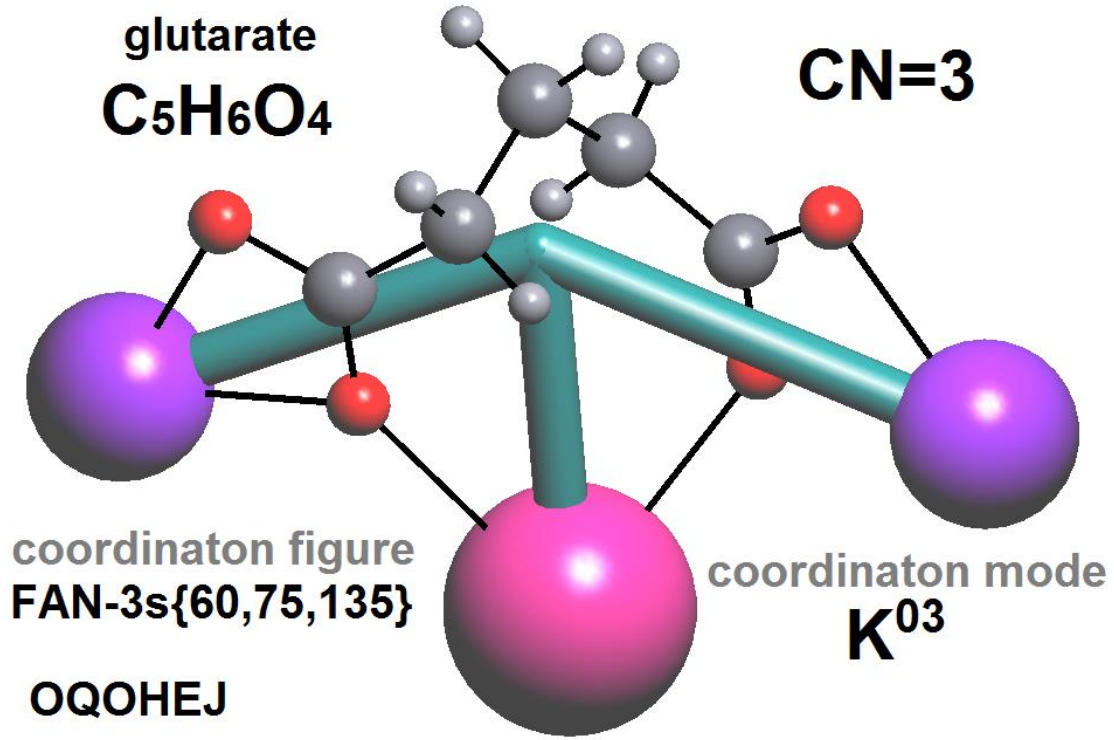
Molecular Fragment Descriptors

TotalVol	140.1 Å^3
G3	0.1055
TotCArea	380.0 Å^2



Molecular Voronoi polyhedral

$Rb[UO_2(glt)(Hglt)]\cdot H_2O$

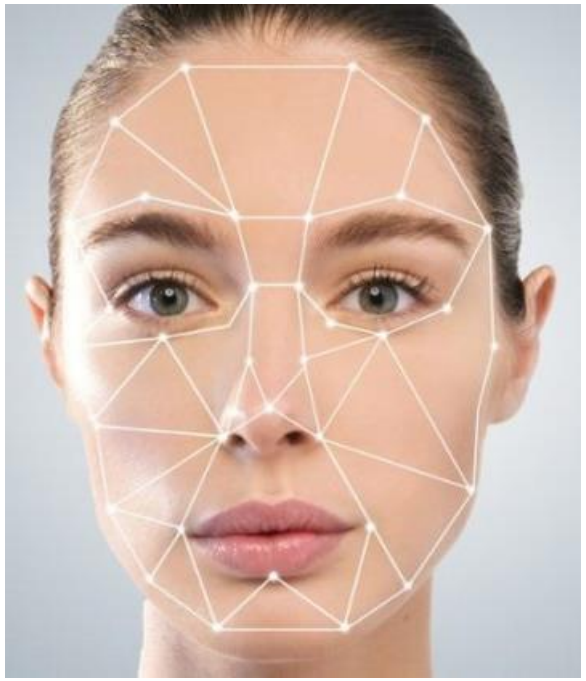


Smile: C(CC(=O)O)CC(=O)O

Polyhedron 117 (2016) 644–651

Molecular Fragment Descriptors

Algorithm for Determining the Shape of the Coordination Figure



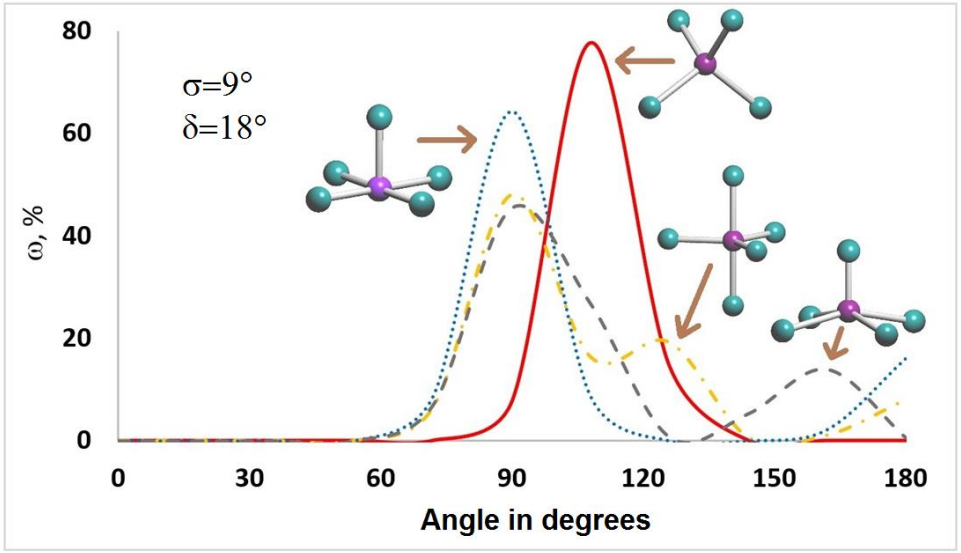
Algorithm

1. Smoothing function
2. Fingerprint calculation
3. Fingerprint comparison

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad (1)$$

$$F_i = 100 \frac{\sum_{k=1}^{180/\delta} f_k(\mu_i)}{\sum_{m=1}^{180/\delta} \sum_{k=1}^{180/\delta} f_k(\mu_m)} \qquad (2)$$

$$r = \sqrt{\sum_{i=1}^{180/\delta} \left(\frac{F_i - F'_i}{2} \right)^2} \qquad (3)$$

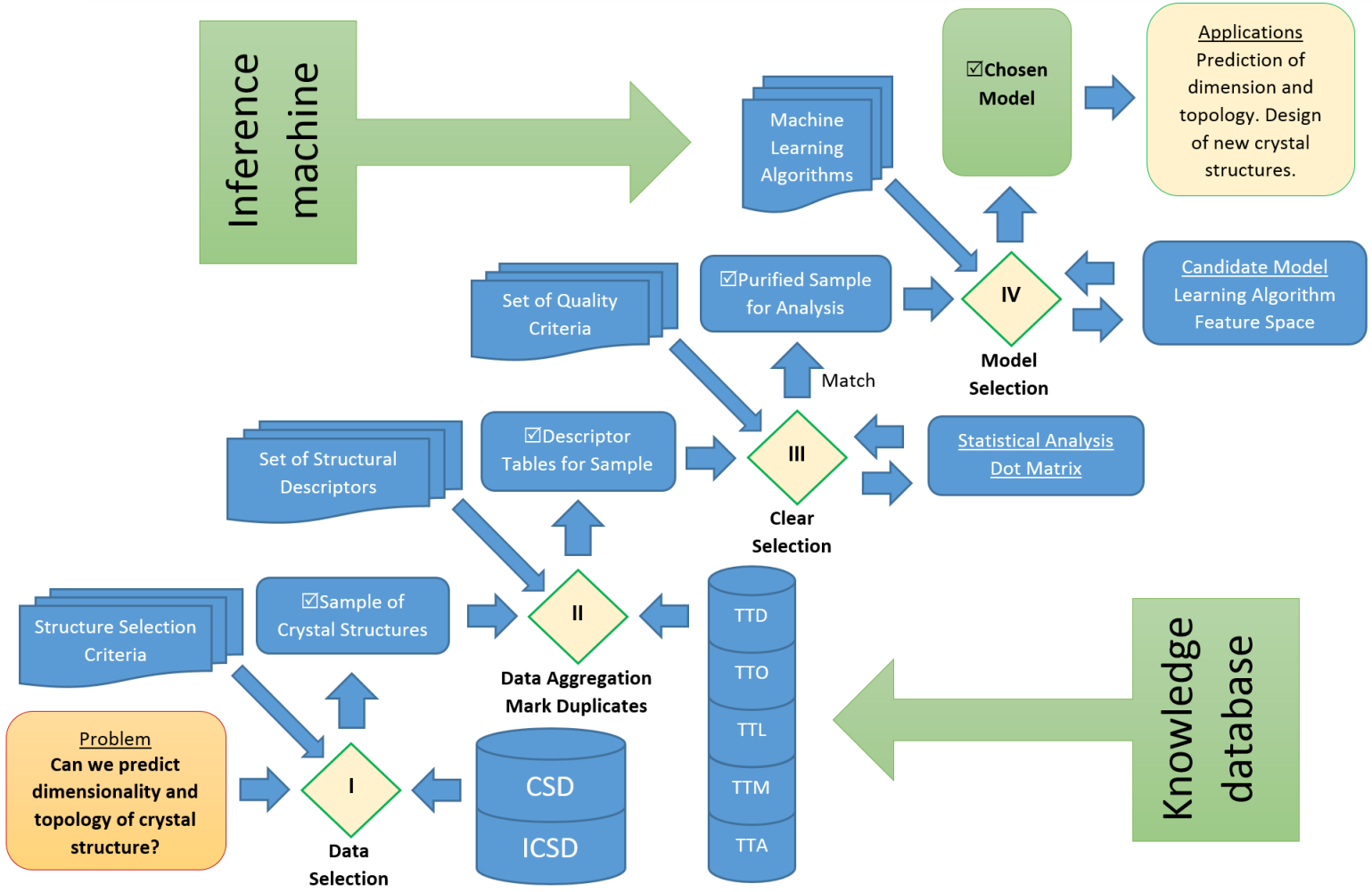


R. Dass, R. Rani, D. Kumar, Face recognition techniques: a review. Int. J.Eng. Res. Develop. 2012, 4, 70.

A. O. Lyakhov, A. R. Oganov, M. Valle, Comput. Phys. Commun. 2010, 181, 1623.

A. P. Shevchenko, I. A. Blatov, E. V. Kitaeva, V. A. Blatov Cryst. Growth Des., 2017, 17, 774.

Scheme for solving the problem



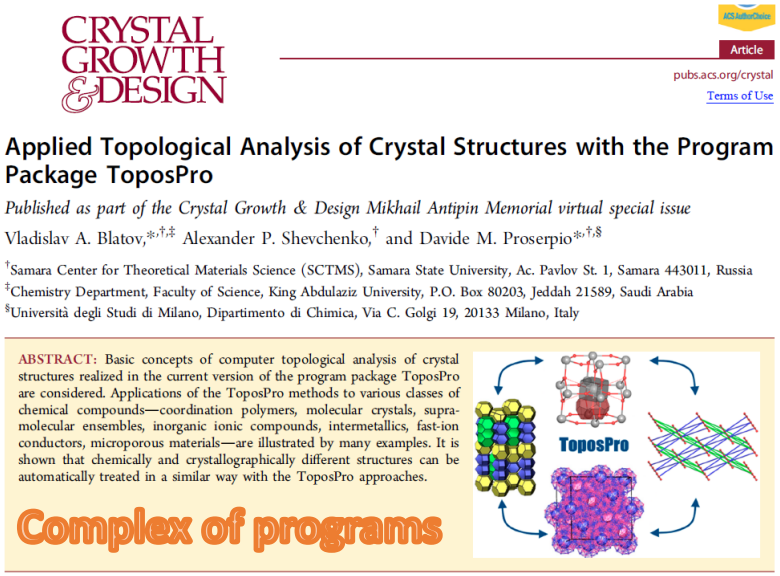
Scheme of machine data analysis.

??? Setting the problem, choosing the target variable!

- Selection of crystal-structural data;
- Create a sample descriptor table;
- statatistical analysis, cleaning and transformation of data;
- Development of a decision model.

$$$ Using the model in practice.

Data sampling and software tools


CRYSTAL GROWTH & DESIGN

Article

pubs.acs.org/crystal

Terms of Use

Applied Topological Analysis of Crystal Structures with the Program Package ToposPro

Published as part of the Crystal Growth & Design Mikhail Antipin Memorial virtual special issue

Vladislav A. Blatov,*[,†,‡] Alexander P. Shevchenko,[†] and Davide M. Proserpio*[,†,§]

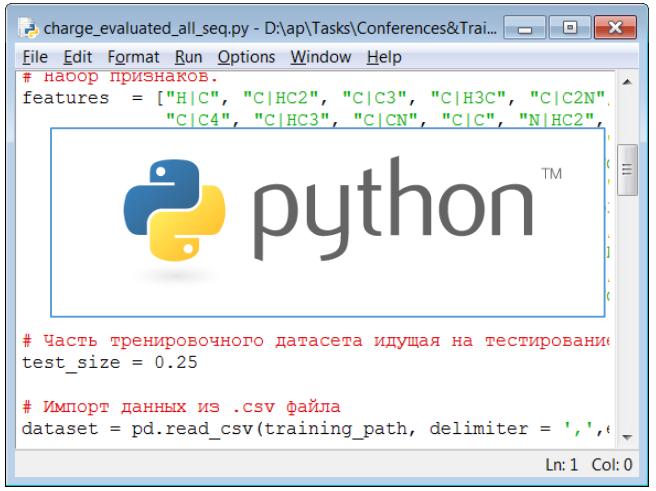
[†]Samara Center for Theoretical Materials Science (SCTMS), Samara State University, Ac. Pavlov St. 1, Samara 443011, Russia
[‡]Chemistry Department, Faculty of Science, King Abdulaziz University, P.O. Box 80203, Jeddah 21589, Saudi Arabia
[§]Università degli Studi di Milano, Dipartimento di Chimica, Via C. Golgi 19, 20133 Milano, Italy

ABSTRACT: Basic concepts of computer topological analysis of crystal structures realized in the current version of the program package ToposPro are considered. Applications of the ToposPro methods to various classes of chemical compounds—coordination polymers, molecular crystals, supramolecular ensembles, inorganic ionic compounds, intermetallics, fast-ion conductors, microporous materials—are illustrated by many examples. It is shown that chemically and crystallographically different structures can be automatically treated in a similar way with the ToposPro approaches.

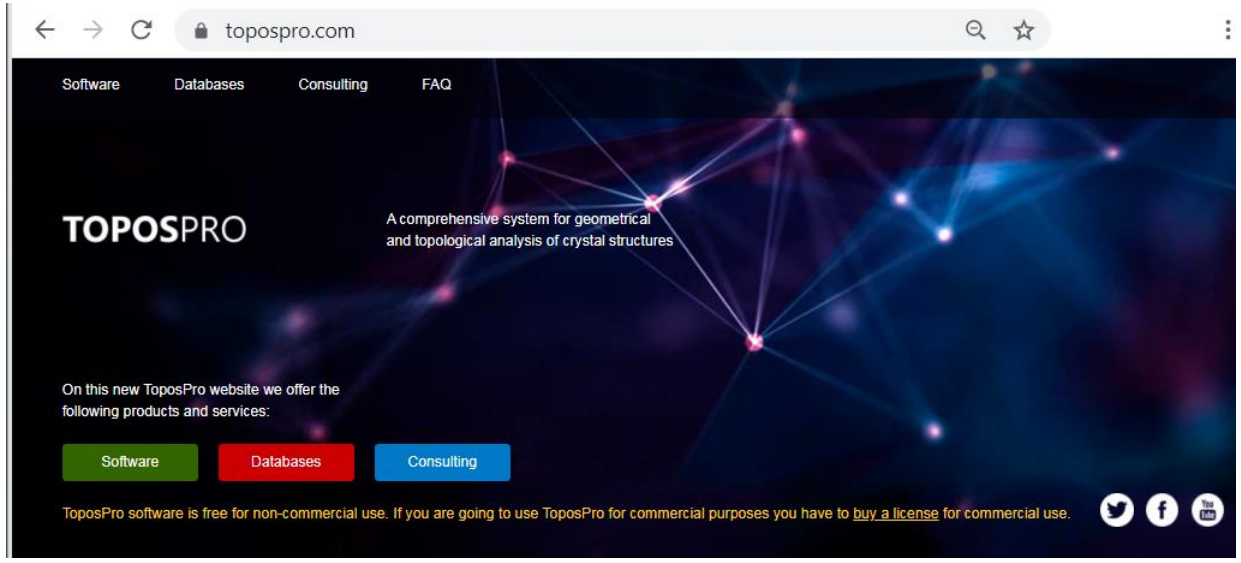
Complex of programs

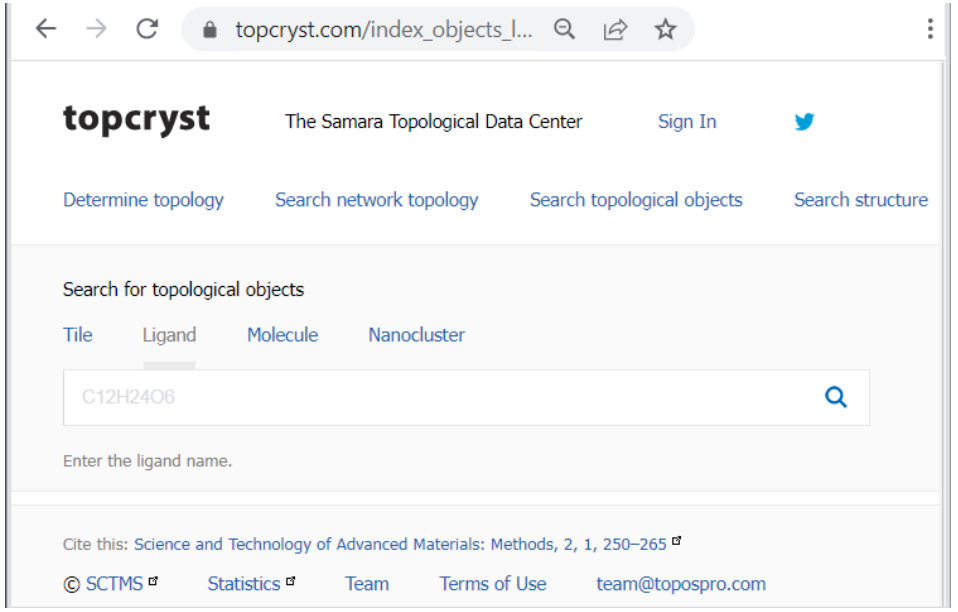
Libraries

TTD - topological types
TTO - topology occurrence
TTL - ligands in their complexes
TTM - organic molecules
TTA - metal atoms in their complexes
TTN - nanoclusters in intermetallics
TTT - tiles in zeolites



Services





Data sampling and software tools

Conditions for the selection of crystal structures:

- These are coordination compounds consisting of metal atoms, ligands, and outer-sphere particles, which are referred to below as the structural building units of a crystal.
- The structure contains only one type of ligand; there are no restrictions on the number and nature of metal atoms and outer-sphere particles.
- The positions of metal atoms are ordered, that is, they are completely occupied by a metal atom of the same chemical type.

Data sampling:

CCDC

38,595 structural studies,
36,821 crystal structures,
18,240 different ligands.

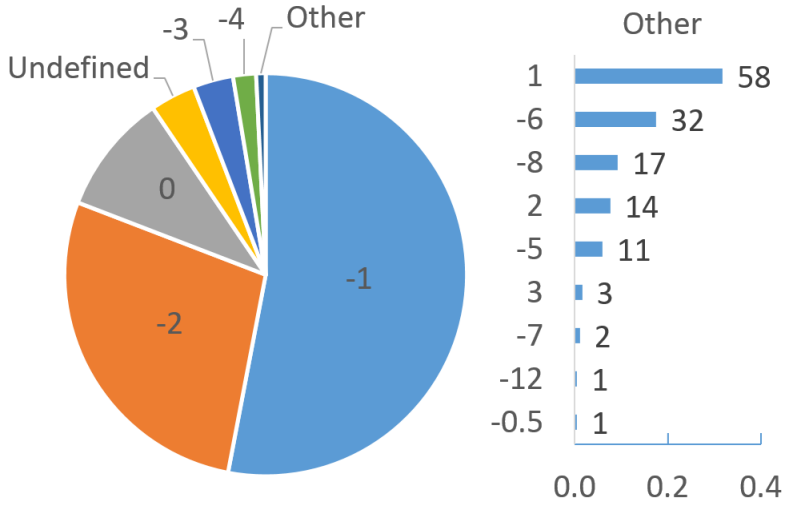
Data sampling and software tools

Table 1. The top twenty ligands by occurrence

	Ligand	N	w, %		Ligand	N	w, %
1	Cl_1305	1092	2.97	11	C5H5_1429	101	0.27
2	O_12	517	1.40	12	C3S5_4070	90	0.24
3	Br_1353	514	1.40	13	C5H7O2_9	88	0.24
4	I_1320	502	1.36	14	H2O_2	86	0.23
5	CO_1161	273	0.74	15	C10H15_1499	84	0.23
6	CN_1476	157	0.43	16	C7H3NO4_1838	78	0.21
7	CHO2_1	144	0.39	17	C32H16N8_2805	78	0.21
8	C2O4_74	129	0.35	18	S_1360	72	0.20
9	F_1809	126	0.34	19	O4P_1603	67	0.18
10	CNS_1405	122	0.33	20	C3H3N2_3597	63	0.17

Table 2. The top five coordination sequences CS_1 and CS_2 of ligand atoms

	CS_1	w_1, %	CS_2	w_2, %
1	H|C	41,8	H|C/C2	14,9
2	C|HC2	14,9	H|C/H2C	12,4
3	C|C3	5,8	C|HC2/H2C2	4,7
4	C|H3C	4,1	C|HC2/HC3	4,4
5	C|C2N	4,0	H|C/HC2	4,0



Distribution of ligands by their charge

Data sampling and software tools

Various feature spaces for ligand charge prediction

MF in monoligand coordination compounds from CSD.

- Nomenclature names of compounds
- MF smiles, which were calculated using the RDKit library
- **Weighted fractions of coordination sequences for ligand atoms**
- MF bit masks

Results and their use

Random Forest - Coordination Sequences

Only ligands of CHNOD composition are taken into account,

whose charge is equal to one of the values -8, -6, -5, -4, -3, -2, -1, 0, 1, 2

- One coordination sequence (185 features)

Accuracy = 0.749(7), Presicion = 0.384(37), Recall = 0.299(24), F1 = 0.331(22)

TEST: Accuracy = 0.754, Presicion = 0.334, Recall = 0.275, F1 = 0.296

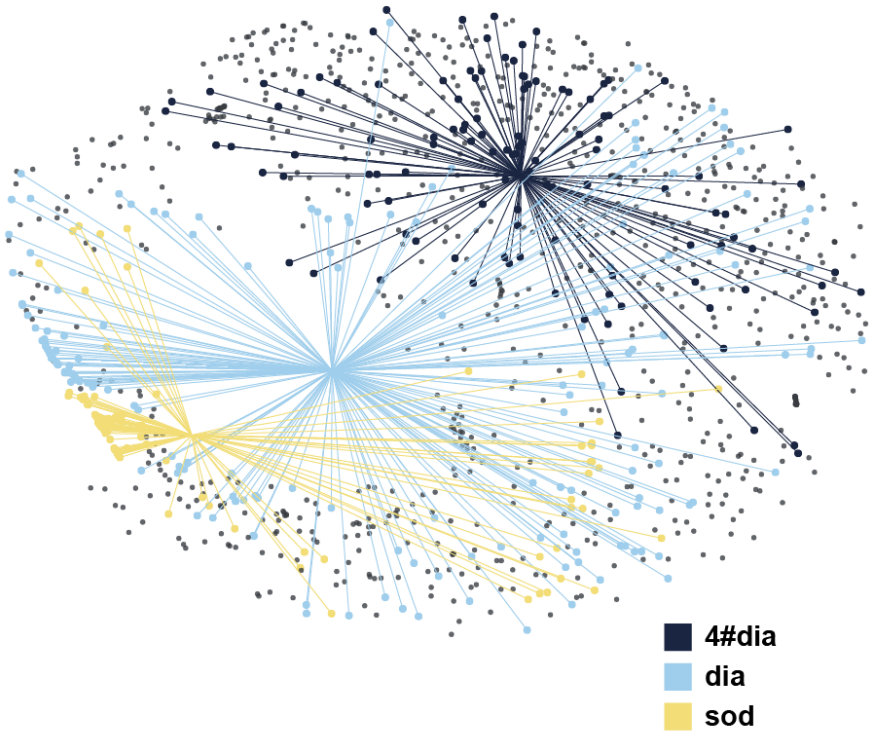
- Two coordination sequences (3,672 features)

Accuracy = 0.735(7), Presicion = 0.337(18), Recall = 0.246(16), F1 = 0.271(17)

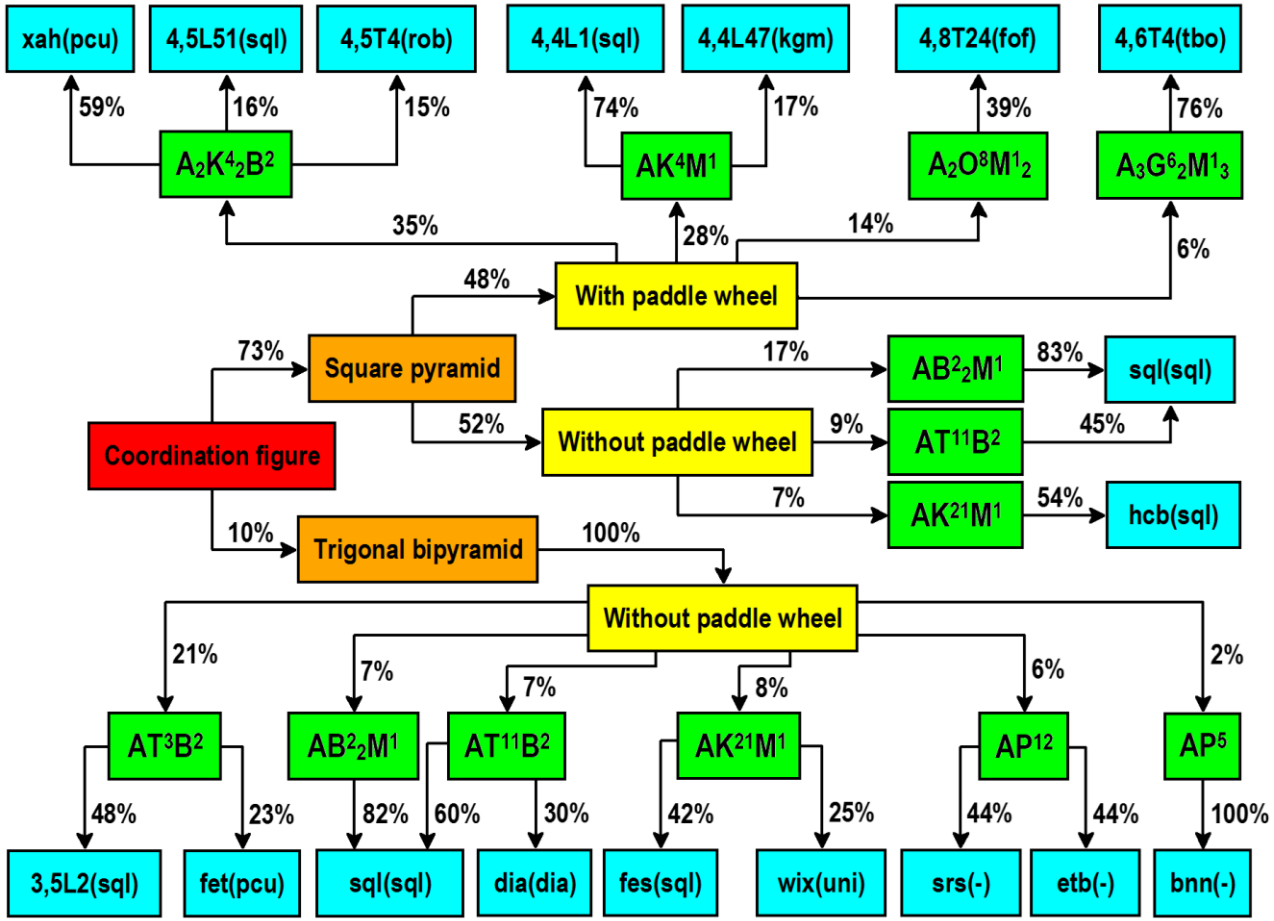
TEST: Accuracy = 0.751, Presicion = 0.322, Recall = 0.270, F1 = 0.287

Results and their use

Search for patterns



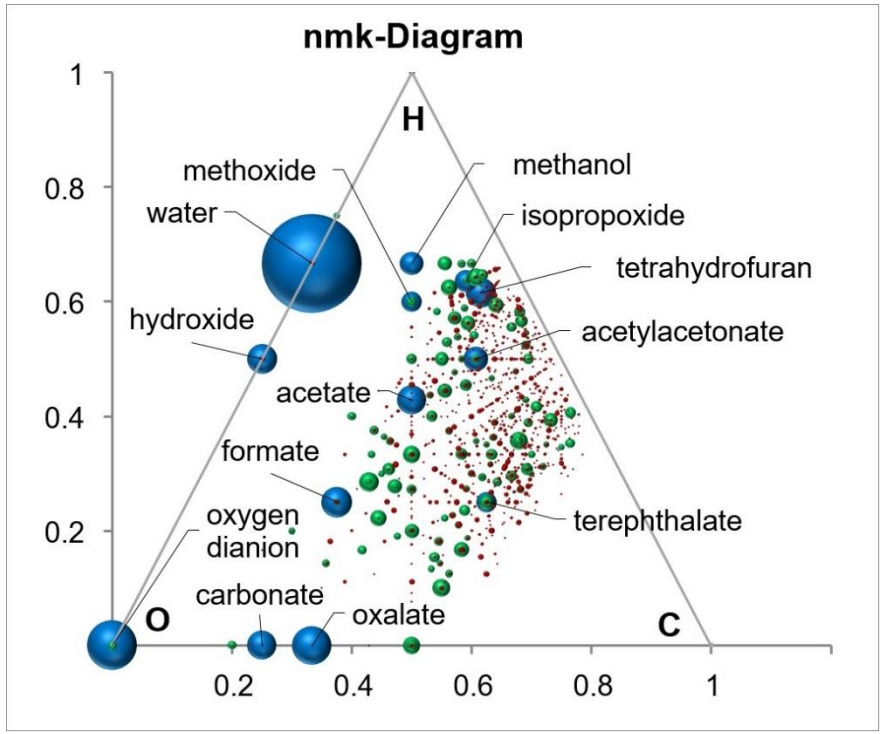
Geometric diversity within isoreticular groups of AB_2 MOFs.
Chem. Mater. 2021, **33**, 8289−8300



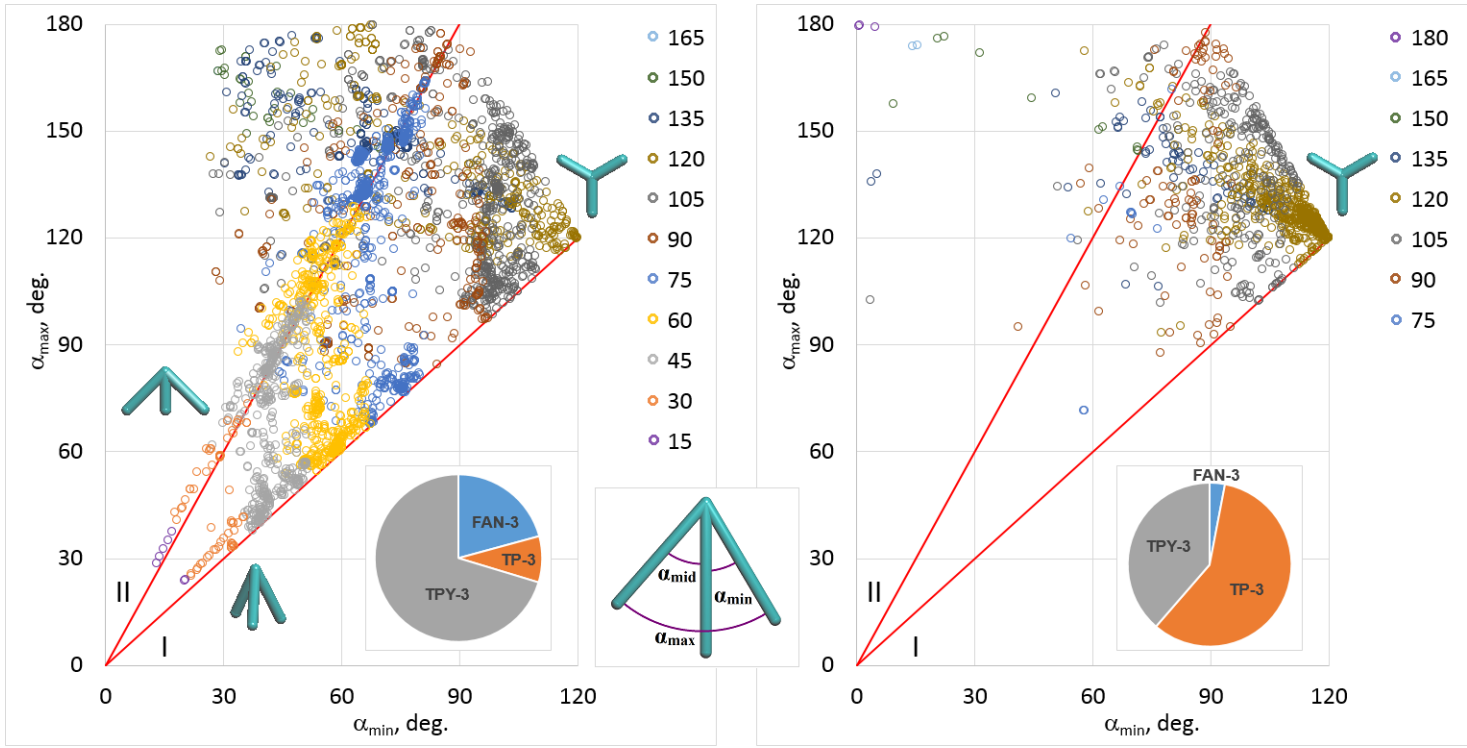
FP Tree
Cryst. Growth Des. 2017, **17**, 774−785

Results and their use

Classification of ligands and metal atoms



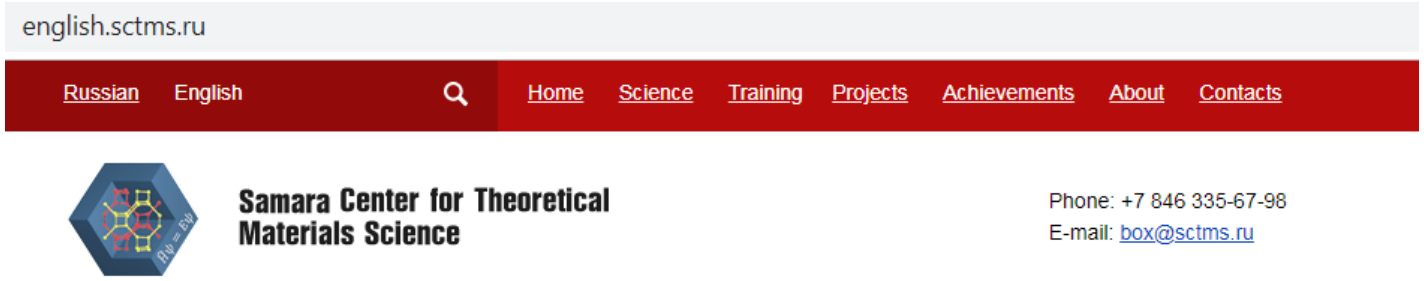
Search for patterns of CHO composition



Configuration spaces of angles in the coordination figures with CN = 3 for ligands (left) and metal atoms (right).

CrystEngComm, 2020, **22**, 7298

Thank you for your attention!



- http://topospro.com/ - complex of programs for the study of crystal structures and our topological databases that is calculated from structural data.
- https://topcryst.com/ - web-services for determining the topology of the crystal structure from the contents of the file, searching for the occurrence of topological types and selecting structural building units for the design of new coordination compounds.
- https://crystalpredictor.com/ - the oxidation state of a metal atom is determined based on the position of the atom in the Periodic Table and geometrical descriptors of the atomic Voronoi polyhedron.

Device for time-of-flight mass spectrometry

The work was supported by the Russian Science Foundation within the framework of the scientific project № 23-23-00387.