Available online at www.joac.info

ISSN: 2278-1862

Journal of Applicable Chemistry

2015, 4 (4): 1024-1144

(International Peer Reviewed Journal)

Critical Review (Cr)

ChemoInformatics
Part I: Molecular Descriptors in Omnimetrics Research

K RamaKrishna[1], Ch. V. Kameswara Rao[2] and R. Sambasiva Rao[3*]

1. Department of Chemistry, Gitam Institute of Science, Gitam University, Visakhapatnam, AP, **INDIA**
2. Department of Chemistry, Basic sciences and Humanities, GMRIT, Rajam 532 127, AP, **INDIA**
3. School of Chemistry, Andhra University, Visakhapatnam 530 003, **INDIA**

Email: karipeddirk@gmail.com, kamesh.chembolu@gmail.com, rsr.chem@gmail.com

Accepted on 11th July 2015

(Dedicated with profound respects to RSR's father R. Venkata Ramanaiah during his birth centenary year)

Conspectus

Background: In mid eighteenth century, the observation of increase toxicity of alcohols to mammals with decrease in solubility in water was the start of a new discipline - structure activity or specifically structure-toxicity relationships. The seminal contributions of Hammett in relating chemical reactivity (logarithm of rate constant/equilibrium constants) with change in substituent groups in aromatic acids and Hansch modelling of effective/lethal dose of compounds with partition coefficient during 1950 and 1970s initiated 'Linear free energy relationships (LFER)' and 'Quantitative structure activity relation ships (QSAR)'.

Molecular descriptors: The protocol of inspection of chemicals for their reactivity at the start up begins at looking into functional groups, cyclic rings, substituents and 2D-connectivity-structure. Constituent descriptors are simple number counts of atoms/bonds/rings in a molecule throwing light on compositional profile. The connectivity (symmetric) matrix of zeros or ones called adjacency matrix leads to distance matrix (of same size) of path was the foundation stone of topological chemistry.

Graph theory: Wiener index, column or row sum of adjacency matrix, proposed in 1947 for explaining boiling points of acyclic organic molecules was the start of mathematical (graph theoretical) chemistry and emergence of molecular descriptor research, now a coveted discipline. The continual advances of Wiener graph continues even in 2015 and occupying a pivotal position in application arena. Randic index and several distance derived descriptors viz. Hosoya, Rouvray, Detour and Zagreb topological indices are in vogue.

Electrostatic-, quantum-, Eigen- paradigms: Kier and Hall introduced electrostatic descriptors of local category based on non-quantum chemical ground. These charge distribution quantities use partial charge, topo electronic measure and surface area. The next major category involves 3D-optimized geometry available from X-ray/ neutron diffraction/ NMR experiments or computational (ab initio, DFT) quantum chemical approaches. Geometric molecular descriptors like shape/shadow indices speak out molecular size and shape. The core of WHIM and its clones is based on chemically valid 3D-geometry of a molecule. 3D-MoRSE, RDF_Wtd descriptors found extensive applications. The quantum chemical descriptors are well known CQC derived IP, EA, charges, multi-pole moments, electronic energies and thermodynamic quantities. The chemical hardness/softness, Fuki's parameters shed light on radical / electrophilic/nucleophilic reactions. Eigen matrix analysis has a key role in several molecular descriptor

categories primarily to get rid of correlation among descriptors. WHIM descriptors derived from Eigen values of 3D-weighted co-variance matrix encode shape, size and atom distribution. GWHIM takes into account of grid reflecting interaction field dimensions. The minimum and maximum Eigen values of atomic charge, polarizability or H-bonding ability instead of atomic number of diagonal matrix of Burden matrix and Eigen spectra of CQC-generated UV-Vis, IR, NMR profiles are coveted first-order probes per molecule. Many of topological and other types also were subjected to Eigen structure enabling use of classical hard least squares analysis. Information descriptors based on Shannon information theory mirrors mass distribution in the molecules. Bond Environment Descriptor (BED) and Augmented Environment Descriptor (AED) from first and second nearest neighbors and their bonds account for effect of fragments around an active atom.

Hybrid descriptors: The hybridization of descriptors of different origin opened a super-highway embracing the best of both information worlds to predict complex responses in real life materialistic compositions. The components of GETAWAY descriptors are geometric and distance matrix based one. The Topo-Geometric/ Electro-topological state atom (E-state)/ Topo-QC descriptors have high impact in statistical modelling and interpretation perspectives. The classical physico-chemical-biological parameters like logP, refractivity, Lipinski (rule of 5) alert index are also now in the band-wagon of molecular descriptors.

Applications: The applications of molecular descriptors over the last two decades engulfed almost all research disciplines under an umbrella term structure activity relationships (SAR). It comprises of computations of molecular descriptors (simple counting to advanced CQC, graph theory, matrix operations), pruning (selection) to a subset based on statistical, information and process-details basis, modelling strategies (MLR, PCR, NN, SVR etc.), validation procedures (residual analysis), ensemble/forest approaches. The target sciences of focus are chemistry, biology, pharmacy, material synthesis, environment are in the priority list. The typical subclasses of SXR include 'activeSite, AntiOxidant, Blood brain barrier, binding, complexation, drug, genetics, HIV/AntiHIV, inhibition, interaction, Logβ_{mlh}, pharmacokinetic, protein-Ligand-interaction, receptor and toxicity'. This qualifies these cutting edge investigations to be referred as "computational descriptors in Omni_metrics", in equilibrium with Chemometrics, Chemoinformatics, Mathematical chemistry, CQC and nature inspired algorithms.

Molecular descriptors from trustworthy packages for sizable number of compounds from academic research stand point have been used for HIV-1 integrase inhibition, activity against HIV-1 reverse transcriptase, Myorelaxant, Antimalarial activity, Serine protease active site, Ocular toxicity, Opiate/ antinidatory activity, PDE4 inhibition, CYP450 inhibition activity, BBB receptor crossing antagonists, Adenosine receptors, corticosteroids binding, Drug–protein interaction, PharmacoKinetics etc.

In biophysical chemistry, Protein-Ligand interactions, Putative proteins, β-blockers, drugs leading to Phospholipidosis, consequences of amino acid sequence, Chromosomal aberrations were studied with mathematical molecular descriptors. The variation of chromatographic/^{13}C NMR responses of large variety of compounds are interpreted and predicted with a pool of descriptors. The validity and accuracy of prediction of chemically significant mathematically/statistically derived constants (parameters) like logβ of metal-ligand complexes, ΔG, logP, Cell permeability and chemical properties -- solubility Boiling point, melting point, Viscosity refractive index and glass transition temperature-- is significantly increased over classical approaches with properly selected descriptors.

Software: The calculation of molecular descriptors started in manual mode mostly with paper and pencil. Later hand-held calculators played an important role for a molecule of large size. Leaving aside many historical in house programs, ADAPT by Jurs, Package of Kier and Hall are the stand alone tools. CODESSA by Katrizky in late 1990s is a full-fledged software package for generation of over 1000 descriptors and regression procedures for cause-effect model. DRAGON from University of Milan by Todeschini is a mega descriptor generator to over 4500 descriptors along with modelling tools. Some of the typical packages (ADRIANA, MOLGEN-QSPR, PreADMET..)in vogue are compared from the stand point of number and classes of descriptor outputted.

Keywords: Physico- bio-chemical properties, Hammett substituent constants, Hansch parameters, Weiner/Randic index, WHIM, Topological, geometric, electro-topological, quantum-chemical, Eigen, spectra, CODESSA, DRAGON

Contents

Molecular descriptors in OmniMetricS Research

1. Introduction

1.1 Historical perspective explanatory variables and cause-effect relationships
1.2 Molecular descriptors
1.3 Waves in descriptor research

2. Applications of descriptors in Structure X relationships (SXR)

- SHIVR — HIV/antiHIV
- SActR — Activity
- SActSitR — ActiveSite
- SAntiOxiR — AntiOxidant
- SToxR — Toxicity
- SInhR — Inhibition
- SBindR — binding
- SBBBR — Blood brain barrier
- SInteractR — Interaction
- SPharmacoKineticR — PharmacoKinetic
- SReceptR — Receptor
- SProtLigIntR — Protein-Ligand interaction
- SDrugR — Drug
- S Seq R — Sequence
- SGenetR — Genetics
- Contributions of Katrizky
- SRespR — Response
  - SChromatR — Chromatography
  - S NMR R — NMR
- S PhysChemPar R — Physico-Chemical_parameter
- SPropR — Property
  - Aqueous solubility
  - Boiling point
  - Glass transition temperatures
  - Viscosity of ionic liquids
  - Miscellaneous:

3. MethodBase_MolDes (MB. MolDes)

- Constitutional
- Topological
- Environmental
- Electrostatic
- B-CUT
- Geometric
- 3D-Morse, RDF
- Quantum chemical
- Thermodynamic
- Orthogonal
  - Eigen value
  - WHIM, GWHIM
  - GETAWAY
- Information
- Hybrid
  - Topo-electrostatic
  - Electro-topological state atom (E-state)
  - Topo-Geometric descriptor
  - Topo-QC descriptor
  - Information content + Gravitational index
- Physical_chemical_biological
- Recent descriptors

4. Software packages

5. State-of-knowledge of molecular descriptors

6. Exploratory super-highway for future MolDes research

INTRODUCTION

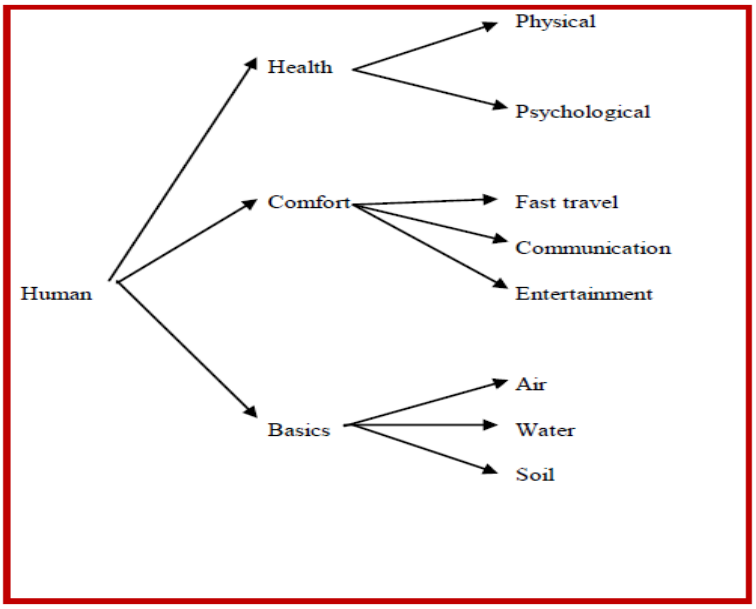
Chemical science has a primary concern of energetics in bond-making and bond-breaking between chemical moieties. H-bonding, electrostatic/ van der Waals attraction, dispersion, stacking interactions, although of small magnitude, have a dominant contribution in biological-macro-molecular interactions. A hierarchical representation of structure of a chemical moiety and its interactions with environment/radiation reflects the intricacies in perception of materials. A zero order representation is name or empirical formula of a compound. A linear (or first order) format is molecular representation showing the functional groups. A 2D-structure explicitly shows nature of bonding (including H-bond) and graph-theoretic-connectivity. Chemists were comfortable to show a third dimension on a 2D- sketch for many decades. A 3D-representation of a molecule (XYZ coordinates or Z-transform etc.) explicitly point out bond-length, bond-angle and dihedral angles of a conformer. The 4-D picture many a time in unfolded form is informative about conformers, but limited to vacuum/gas phase. In other words, only more rigorous format enables one to have details in solid state, liquid phase or in presence of a solvent/salt/micelle/cell etc. In a nut shell, the functional properties of molecules are understood at the moment with complimentary, redundant, unique structural details from 1D to 4D-representation.

Historical perspective explanatory variables and cause-effect relationships

Greeks were intelligent to choose orthogonal explanatory variables as X (or independent) and thus later multi-linear regression was adequate. With the available orthogonalization, project pursuit, rotations of variable space after transformation are a solace to arrive at acceptable results for correlated variables. But, this is at the cost of either loosing physical significance of transformed variable space or high explainability/predictability with unknown combination of variables.

In 1863, it was observed that toxicity of alcohols to mammals increased with decrease in their solubility in water. Later, in 1890 this effect was attributed to lipophiliciy. Hammet explained chemical reactivity (ionization, rate of reaction) of meta- and para-substituted aromatic acids in terms of substituent constant (σ). Taft extended this idea to ortho and aliphatic analogues. Hansch modeled biological response (ED_{50}, LD_{50}) with hydrophobicity measured as partition coefficient of the compound between n-octanol and water. Second order terms in multi-linear parametric equations accounted for the non-linear trend. Hansch and his group applied the quantitative structure activity relationship (QSAR) model to drugs, agrochemicals etc. Free and Wilson proposed binary regression model indicating the presence or absence (yes/no 1/0) of an atom (N,O,S), group (C=O, NH_2) or sub-structure (n-membered ring) in a compound. In Hansch like models, this factor was independently proposed as indicator variable. Of course, the consequence is increased statistical correlation between X and y and decrease in variance of residuals in y. The goals of scientific activity are to enhance the health and quality of life mankind and have fewer effects from even natural disasters (chart 1.1).

Chart 1.1: Focus of scientific pursuits



Typical centres of high technology

- CERN
- NASA
- USA-Env
- US-Tox
- FDA
- Space Lab

Molecular descriptors

The classification of descriptors into different categories is closely related with scientific evolution over last 75 calendar years. The representation of a molecule in zero to four dimensions throws light on simple atom/bond/ring counts, identification of groups /functional groups/moieties, details of bonding/ connectivity, spacial (three-dimensional) orientations of molecules/varieties of conformers. The classes of molecular descriptors are, mathematical (graph theoretical) calculations from 2D-connectivity in chemical structure and 3D-optimized geometric one of a conformer of a molecule. Others include molecular fingerprints consisting of numerous bit positions, spectral data derived tensorial descriptors. The derived

and indirectly calculated vector/matrix of descriptors enable one to probe into electro-static potential, electron density, multi(mono- to hexadeca-) poles, Eigen spectra etc. The newer and modified descriptors are in the direction of reducing degeneracy for a homogenous set of compounds and to include intrinsic but critical properties operating under extreme conditions (P, V, T, Concentration)/ environment (solvent, matrix, cell). A bird's eye view of evolution of molecular descriptors can be broadly divided into four waves, of course with merging border lines.

Waves in descriptor research

First wave

Wiener initiated the descriptor research in 1947 and the number of molecular descriptors now exceeds 4000. Many teething problems were circumvented over years and the tool occupies a niche in interdisciplinary academic as well as industrial pursuits.

Second wave

Kier et al. and Randic made seminal contributions for over a quarter century past 1970s in developing molecular connectivity indices χ_s based on graph theory. In descriptor space, χ_s are low dimension depictions of structure of a molecule (chart 1.2, KB. 1.1).

Chart 1.2: Molecular connectivity indices χ_s

- Correlation between χ_s
  - **Remedy :** Orthogonalisation of χ_s
- Modifications in χ_s
  - Valence
  - Augmented valence
  - Electro-topological state of AIM

KB. 1.1: Graph theoretical information

If	Topological index is same for two structures
Then	It is called graph invariant
If	Isomorphic graphs
Then	Graph invariant
If	Isomorphic compounds
Then	Isomorphic graphs

Bonchev in 2001, proposed overall connectivity index, which surmounted the degeneracy of connectivity index even when higher order path clusters are considered. Mandel found over all connectivity superior for physico-chemical properties of C3 to C8 alkanes. These developments of a vector of descriptors per compound are the highlight during this period.

Third wave

The Eigen value analysis of spacial distribution and spectra of compounds in 1990s was a new facet in mathematical analysis of physico-chemical responses of molecules. The progress continued and Eigen value based descriptors of a molecule were calculated from CQC derived IR, UV-Vis and NMR spectra. WHIM descriptors of Todeschini encode information about molecular size, shape, symmetry and atom distribution.

WHIM

- \+ Orthogonal
- \+ Surmounts limitations of correlation among explanatory variables

Chart 1.3 Select list of researchers in molecular descriptors

Balaban, A.T.	Kier, L.B.
Balasubramanian,K.	Kim, K.H
Basak, S.C.,	Klein, D.J.
Bonchev, D.	Klopman, G.
Burden, F.R.,	Krygowski, T.M.,
Crippen, G.M.	Kubinyi, H.

There is a treasure of around nine thousand references up to the year 2007 incorporated in the monographs (chart 1.3) of Todeschini group [1, 3-5] published in 2000 and 2008. The formulae and definitions are exhaustively spread in journal papers besides reviews and software manuals/fliers. DRAGON with more than four thousand and CODESSA with around fifteen hundred descriptors are both from academic research groups of Todeschini from Italy and Katrizsky of Florida (US). These packages are a treasure of knowledge, yet easy-to-implement tools for interdisciplinary researchers for generating breadth wise information store without really going deep into the intricacies.

Devillers, J.	Li, L.-F.
Diudea, M.V.	Livingstone, D.J
Estrada, E.	Meyer, A.Y.
Fujita, T.	Mezey, P.G.
Gasteiger, J.	Pogliani, L.
Gilvez, J.,	Randic, M.
Gutman, I.	Rekker, R.F.
Hall, L.H.,	Rouvray, D. H.
Hansch, C.,	Schultz, H.P.
Hopfinger, A.J.	Taft, R.W.,
Hosoya, H.	Todeschini, R.,
Jurs, P.C.	Trinajstic, N.
Kaliszan, R.	Trinajstic, N.,
Kamlet, M.J.	Wiener H
Katritzky, A.R.,	Xu, L.

The two perceptive means (eyes) hither to in vogue in science are experimental and computational (statistical/ theoretical/ simulated) probes. The coveted knowledge based approach and tensorial computation of this century are well in practice in our laboratory since three decades. We attempted to perceive the molecular descriptor research through this third eye. Typical scrutinized reports during this decade are incorporated in reference section. Our long term goals include object oriented display and implementation in MATLAB, the necessary conditions, formula, m-file, typical data sets showing success as limitation/ failure of a descriptor/set and available chemical interpretation. The derivation of a descriptor from a more general version, relation with other descriptors, and its area of influence are underway. The quality of results of course is dependent under the supervision of third eye of an expert and/ or international standard protocol.

The nascent molecular descriptors and their all-pervading applications mostly of this decade [1-200] are reviewed here. In continuation of our series of publications [195-200] in complex equilibria in aquo-organic mixtures, Chemometrics, kinetometrics, environmtrics, piscimetrics, speciometrics, omnimetrics and internet chemical education (HAI, ICE, Hot_Ice), We don't intend to duplicate literature citations or mathematical equations. Yet, nascent disciplines viz. chemoinformatics, toxicoinformatics, chemical_genoinformatics, chemical_biological_informatics pooled up large knowledge bits apart from (numerical, nominal) data bases worth exploring for sparkles/hot_ice or cold-diamonds to have leap in the present day expertise.

During loud thinking of ways and means of picking up phenomenon relevant descriptors, it was found up to late 1990s even published heuristics are limited. During the last one decade there are scanty reports about the bio-chemical/pharmaceutical relevant categories of descriptors. Ever since multivariate-SXR started, the elimination of variables is based on inter correlation of X or nearly same numerical values. After preliminary scrutiny, the further, the basis of elimination of descriptors is with low cc of Y with X. In the final set of models the descriptors and their type are documented as if trivial.

02. Applications of molecular descriptors in SXR

The theoretically computed descriptors are at most fundamental level and reflect quantitative picture of several factors of structure of compounds. Thus, their use in prediction and explanation of physical-chemical-biological (PCB) properties and activity in bio- and eco-systems have been preferred to the earlier (successful) macro- and microscopic explanatory factors. With an initial set of a single digit to few hundreds of descriptors, predictive models in physico chemical investigations, environmental management studies, drug discovery from leads and material science resulted in many success stories. The similarity measure in descriptor space is a probe to achieve set of compounds with desired characteristics. Diversity measures are appropriate to discriminate compounds with desired characteristics from the remaining set. The subtle sub goals viz. classifying drug like molecules from non-drugs and toxic from non-toxic groups is achieved with a large number of measures of similarity, diversity in multi-dimensional descriptors space. The statistical and mathematical tools employed for clustering, classification and discrimination are broadly categorized into supervised and unsupervised learning techniques. Software for LDA, QDA, PCA, HCA, SVM, rough sets, ridgelets and SOM-Kohonen are available as freeware and commercial products. A virtual library of size 10K to 200K is developed based on variation of substituent, side chains etc. by commercial and academic research groups.

chart 2.1: Abbreviations: Structure $$$ Relationship

Abbreviation	Structure	$$$	Relationship
SActR	Structure	Activity	Relationship
SActSitR	Structure	ActiveSite	Relationship
SAntiOxiR	Structure	AntiOxidant	Relationship
SBBBR	Structure	Blood brain barrier	Relationship
SBindR	Structure	binding	Relationship
SComplexR	Structure	Complexation	Relationship
SDrugR	Structure	Drug	Relationship
SGenetR	Structure	Genetics	Relationship
SHIVR	Structure	HIV/AntiHIV	Relationship
SInhActR	Structure	Inhibition Activity	Relationship
SInteractR	Structure	Interaction	Relationship
SlogβR	Structure	$Log\beta_{mlh}$	Relationship
SPharmaco KineticR	Structure	Pharmaco-kinetic	Relationship
SProtLigIntR	Structure	Protein-Ligand interaction	Relationship
SReceptR	Structure	Receptor	Relationship
SToxR	Structure	Toxicity	Relationship

Acronym	Abbreviation
Lethal dose	LD_{50}, LD_{75}
Effective dose	ED_{50}, ED_{75}
Inhibitory conc.	IC_{50}, IC_{75}

Molecular Descriptors in_action
ADME/toxicity prediction
Cause-effect models
Virtual screening
Diversity analysis
Library design

During last two decades, SPropR, SActR and many such correlations (chart 2.1) encompass every discipline and breadth wise growth is noteworthy. However, the depth wise inventions always have a lag in the translation, implementation and wide application in inter- and intra- and cutting edge-cross disciplinary research tasks as usual. Bulk and macroscopic physical chemical biological parameters which were instrumental in MLR models now play a secondary role. The objectives and quality of reports are diverse and there cannot be a utopian framework (save international agencies' stipulations) as it changes with scientific progress. Typical literature reports on applications of descriptors in diverse fields of research in physical chemical and bio-sciences follow.

🕮 SHIVR (Structure HIVRelationships)

HIV-1 protease inhibitors: Maw and Hall [61] applied E-state and molecular connectivity indices in predictive model of HIV-1 protease inhibitors (chart 2.2). Huuskonen [78] used TIBO derivatives against HIV-1 reverse transcriptase (chart 2.2b). Saíz-Urra et al. [154] modelled HIV-1 integrase inhibition compounds with GETAWAY, WHIM etc. (chart 2.2c). The model explainability increased from 62.4% to 72.5% after elimination of 5 outlier chemical compounds from the dataset. Wisniewski and Castleman [60] reported SActR model for compounds against HIV-1 reverse transcriptase with E-state descriptors (chart 2.2d). Hall [74] reported binding affinity to human serum albumin (chart 2.2e).

Chart 2.2a: SAntiHIVActivityR — S HIV R

Descriptors-Model	
E-state indices	2 Hydrogen
Molecular connectivity	shape index

Molecules: HIV-1 protease inhibitors
- 32 : Trn
- 15 : Pred

Response: In vitro enzyme binding

Model_Statistics		
Training		
r^2	s_{PRESS}	q^2
0.86	0.60	0.79
Prediction		
MAE	0.82	

Chart 2.2b: SInhibitionR — S HIV R

Descriptors: *E-state*

Properties: *Log P*, *MR*

Response: *HIV-1 RT inhibition*

Inf.Bits.
HIV-1 RT inhibition of TIBO derivatives
Influencing factors lipophilic electronic

Model	Residual statistics					
	Training			Prediction		
	#	r^2	s_{PRESS}	#	r^2	s_{PRESS}
MLR	41	0.85	0.56	24	0.80	0.64
	Compounds	TIBO derivatives against HIV-1 reverse transcriptase				

Chart 2.2c: Descriptors for HIV-1 integrase inhibition

WHIM	GETAWAY	3D-MORSE
RDF	Geometrical	Randić

Chart 2.2d: SAntiHIVActivityR — S HIV R

Descriptors-Model2	
E-state indices	four (2, 4, 8, carbon) atoms level
log *P*	MR

Molecules: TIBO derivatives
- 41 :Trn
- 24 :Test

Response: HIV-1 reverse transcriptase

Model			
Training		*Prediction LOO*	
r^2	s_{PRESS}	q^2	s_{PRESS}
0.85	0.52	0.80	0.56

Chart 2.e: SBindAffR — SActR — S HIV R

Descriptors
- E-state
- Topological
  - Chi

Properties	Log P, MR
Response_ biological	Binding affinity to human serum albumin
Response_ Instrument	HPLC _RI, Immobilized albumin
Compounds	94 drugs
Software	MDL QSAR

Inf.Bits.
- Randomization of response input
- No chance models
- Five-membered heteroatomic rings have a negative factor
- Six-membered heteroatomic rings have positive effect
- HIV-1 RT inhibition of TIBO derivatives

Model	Residual statistics					
	# drugs	r^2	S	s_{PRESS}	q^2	10% LGO CV
Training	94	0.77	0.29	0.69	0.70	0.69
Test (external set)	10				0.74	
	10% LGO CV	10% leave-group-out cross validation				

🕮 SActR (Structure Activity Relationships)

The activity is confined to biological response in human beings, animals and insects, bacteria/ viruses. These studies were primarily in vitro and then the results were extrapolated for planning investigations in animals. However, due to the insurmountable gap between ex-vivo (in vitro) and vivo activity, a new era of in vivo investigations started in 90s. The broad discipline biological activity has firm offspring. In many neurological disorders, JNK3 signaling pathway plays a crucial role (chart 2.3a). Ijjaali et al. [10] used BCUT and P_VSA descriptors in developing SActR models with different biological activity thresholds from IC50 data . Pinheiro et al. [163] investigated molecular descriptor model for antimalarial activity against P. falciparum by soft modeling approach (chart 2.3b).

Myorelaxant: Sharma et al. [21] applied DRAGON software to calculate molecular descriptors (chart 2. 3c) of cromakalim analogues in developing ATP-sensitive potassium channel activation.

Chart 2.3a: Molecular descriptor modeling of JNK3 signaling path way	
X(MolDesc)	
1) BCUT	**2)** P_VSA
Response: JNK3 activity IC50	Database: Aureus Pharma' AurSCOPE Kinase knowledge database

Chart 2.3b: Model for antimalarial activity against P. falciparum with molDesc				
X(MolDesc)				
Molecular softness	Total surface area	Randic's index, path-1 molecular connectivity-average	Bond information index	Shape index based on paths of length 2 of Kier
🔔 van der Waals' volume	🔔 WHIM-3D indices (L1v and Tv)		**Model** ❑ PCA ❑ KNN ❑ SIMCA	

Chart 2. 3c: Molecular descriptors of low explainability for A2B agonist effect				
X(MolDesc)				
1) BCUT	**2)** Functional	**3)** Topological	**4)** Constitutional	**5)** Atom centered fragments

6) Empirical	**7)** BCUT_modified	**8)** 2D autocorrelations	**9)** Galvez topological charge indices
Software : DRAGON		Descriptors : 470	
Compounds: 89: Adenosines			**Model** Combinatorial Protocol in MLR
Response: Myorelaxant (ATP-sensitive potassium channel) activation of cromakalim			

Inf.Bits.

- The activity can further optimized by changing substituent groups in chromans.
- The increase in polarizability with increasing path length or higher atomic mass with decreasing path length have a positive effect on activity
- High bond order of carbon atoms attached to heteroatom as well as no or less branching augments activity

Human intestinal absorption: Wegner et al. [92] filtered 2900 molecular descriptor for 196 molecules in predicting by GA coupled with Shannon Entropy Cliques algorithm (chart 2.4).

Chart 2.4: Binary classification of compounds of interest in human intestinal absorption

Descriptors

- RDF
- Conjugated electrotopological state
- Topological polar surface area

Molecules	196
Response (Measured)	Human intestinal absorption
Descriptors	2934

Software	JOELib and MOE
Model	Binary classification
	▶ Adaptive boosting algorithm
Feature selection	GA + Shannon Entropy Cliques

Chalcones: The derivatives chalcones possess anticancer activity. Rybka et al. [18] developed a best linear predictive SXR model for anticancer activity of chalcones against HT-29 human colon adenocarcinoma cell lines (chart 2.5).

Mason and Beno [9] employed BCUT chemistry space and 3D- multiple-point pharmacophore descriptor values to arrive at molecular diversity/similarity in combinatorial virtual libraries

Chart 2. 5: Molecular descriptors in SXR modeling of Chalcones

X(MolDesc)

1) BCUT	**2)** Topological	**3)** Electronic
4) Geometrical	**5)** QC	
		Best model
Response: anticancer activity against HT-29 human colon adenocarcinoma cell lines	Compounds: 136 Chalcones Descriptors >1000	10-member rings BCUT weighted by van der Waals volume

🕮 SActSitR (Structure ActiveSite Relationships)

Mason and Beno [9] developed a combinatorial library for Ugi condensation reaction and a serine protease active site using BCUT chemistry space and multiple four-point pharmacophore fingerprints. SAA was employed for simultaneous optimization of combinatorial reagent selection and diversity of product shape.

STOxR (Structure Toxicity Relationships)

Chemical toxicity: Singh et al. [88] applied molecular descriptors in predicting toxicity of diverse chemical list of Organization for Economic Co-operation and Development (OECD) with classification and regression STOxR models (chart 2.6).

Toxicity of Phenols: Ertürk and Saçan [147] used WHIM descriptors in toxicity studies of phenols towards *Chlorella vulgaris*. The WHIM (with T total size index/weighted by atomic masses descriptor) based STOxR model has better predictive power for even anilines. The validation set consisted of toxicity of 58 compounds to freshwater alga *Pseudokirchneriella subcapitata*.

Opiate/ antinidatory activity: Dervarics et al. [97] introduced a chirality-sensitive flexibility (CSF) descriptor in SXR models for compounds of opiate/ antinidatory activity with better performance compared to pharmaco pair distance measure (chart 2.7).

Cytotoxicity: Saíz-Urra et al. [13] found 2D-autocorrelation descriptors successfully modelled cyto toxicity of 37 naphthoquinone ester derivatives against oral human epidermoid carcinoma (chart 2.8).

Chart 2.6: Ensemble based **decision treeboost/treeforest** models for toxicity of OECD listed chemicals

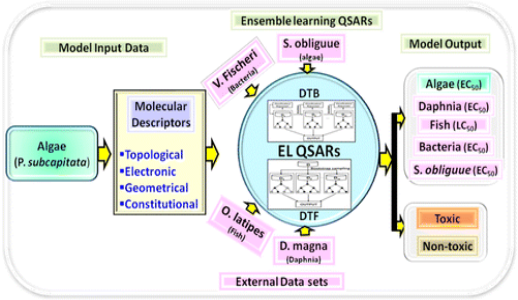
Response: Chemical toxicity	Species: Algae (*S. obliguue*), daphnia, fish, and bacteria
Model Ensemble learning Classification and regression ▪ Decision treeboost ▪ Decision tree forest Algs. Stochastic gradient boosting Bagging algorithms	Similarity index Tanimoto Statistic Kruskal–Wallis

	Classification accuracy %	
	DTB	DTF
All algae data	97.82	99.01
Four test species	[92.50 to 94.26]	[92.14 to 94.12]

Descriptors

Topological Constitutional
Geometrical Electronic

	Regression R^2 (measured vs model)			
Species →	1	2	3	4
Model	0.918	0.15	0.905	0.21
Test	[0.575 to 0.672]	[0.18 to 0.51]	[0.605 to 0.689]	[0.20 to 0.45]



Courtesy of Ref 88

Chart 2.7: SActR with a new descriptor for chirality-sensitive flexibility

Descriptor: Pharmacophore point pair distance

\+ Prediction of features of the active conformation

− Not able to describe chiral arrangement of pharmacophores

Remedy: Chirality-sensitive flexibility (CSF) descriptor

Basis: Definition of a plane by three pharmacophore points;

▶ Calculation of distance between a pharmacophore point and the plane

Model : 3+3D-SActR		
#	Compounds	Response:$$$ Activity
37	Endomorphin analogues	Opiate
38	PGF2α analogues	Antinidatory

Chart 2. 8: Model compounds for oral human epidermoid carcinoma

X(MolDesc)				
1) BCUT	**2)** Galvez topological charges indexes	**3)** Randic's molecular profiles	**4)** geometrical	**5)** RDF
		Resp = MLR (2D-autoCorr)	Comparison	
Response: cytotoxicity of these compounds against oral human epidermoid carcinoma	Compounds: 37 naphthoquinone esters	> 84.2% of variance GA: descriptor selection		

Cytotoxicity of benzophenazines: Saíz-Urra et al. [15] compared MLR-GA models for IC_{50} of topoisomerases I and II versus drug-resistant human small cell lung carcinoma line cell H69/LX4 (chart 2.9). 3D-MoRSE descriptors for 64 benzophenazines explained 82.2% of the variance of observed experimental activity after elimination of outliers (chart 2.9).

Chart 2. 9: Cytotoxicity of benzophenazines with differents sets of descriptors

3D-MORSE	WHIM	RDF
Topological,	Randić	Geometric
2D autocorrelations	Galvez topological charge indexes	BCUT
Response : inhibition of the topoisomerases I and II expressed as cytotoxicity		

Insect toxicity: Grodnitzky and Coats [69] predicted insect toxicity of mono-terpenoids with no experimental insecticide data or those compounds which are not yet synthesized from S Insecticide Activityt R using molecular descriptors (chart 2.10).

Chart 2.10: S InsecticideActivityt R SActR

Response (Measured)	LD_{50} values for house fly toxicity
Compounds	Aromatic mono-terpenoids
Task	Prediction
Model	MLR

Descriptors
- Mulliken Population
- GETAWAY

Inf.Bits.

Predictive models

- In plants, monoterpenoids defend against herbivores, plant-directed pathogens or competing plant species.
- Results of SActR pave way to good lead compounds for synthesis of effective insecticides

Ocular toxicity: Solimeo et al. [51] reported ocular toxicity predictive consensus model with DRAGON calculated molecular descriptors along with chemical environmental ones for a pruned data set obtained from a national data center (chart 2.11).

Chart 2. 11: Animal ocular toxicity relationship with descriptors

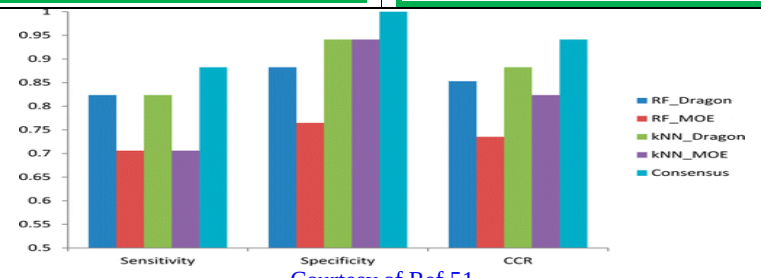
Descriptors
- Dragon
- Molecular Operating Environment

- *Response :* Animal ocular toxicity data
- *Method :* Rabbit Draize test

- *Data Source :* National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods
- Pruning
  - Eliminate duplicates, mixtures, inorganics
  - ⟶75Compounds

Models
- K-Nearest neighbors
- Random Forest

Model testing	
Model	**CCR**
All consensus	72 to 87%. 0.93



Courtesy of Ref 51

Mosquito larvicidal activities: Pasquale predicted mosquito larvicidal activities of a series of prepared chalcones similar in structure to the most active molecule by SActR models using molecular descriptors (chart 2.12).

Chart 2. 12: Molecular descriptors in probing into larvicidal activity

Dragon descriptors	
Constitutional	Topological
Electronic	Geometrical

Model characterstics
- Response: Larvicidal activity
- Task : Prediction
- #Descriptors : 1000
- #Compounds : 28 chalcone derivatives

Inf.Bits.

Chalcones having one or both electron-rich rings showed high toxicity
activity of chalcones was reduced by electron-withdrawing groups

SAntiOxiR (Structure AntiOxidant relationships)

Antioxidant activities: Li et al. [149] applied SAntiOxR model for 214 tripeptides using molecular descriptors as explanatory variables (chart 2.13, table 2.1)

Chart 2. 13: SAntiOxidantR with WHIM descriptors

X

Molecular Surface-WHIM	Hydrophobic Electronic Steric Hydrogen (HESH)
Vectors of Hydrophobic Steric and Electronic properties (VHSE)	Divided Physico-chemical Property Scores (DPPS)
Isotropic surface area–electronic charge index (ISA–ECI)	*Z*-scale

- *Y* : Measured antioxidant activities with ferric thiocyanate method

Table 2.1: SAntiOxidantR model stastistics

Elimination of outliers by Hotelling's T^2 method and residual analysis six significant models

Descriptors	Model/Training			Prediction
	R^2	Q^2	RSD_c	R^2
DPPS, HESH, ISA–ECI, VHSE	0.6	0.5	0.39	> 0.7)
MS-WHIM, *Z*-scale	0.6	0.5	0.44	----

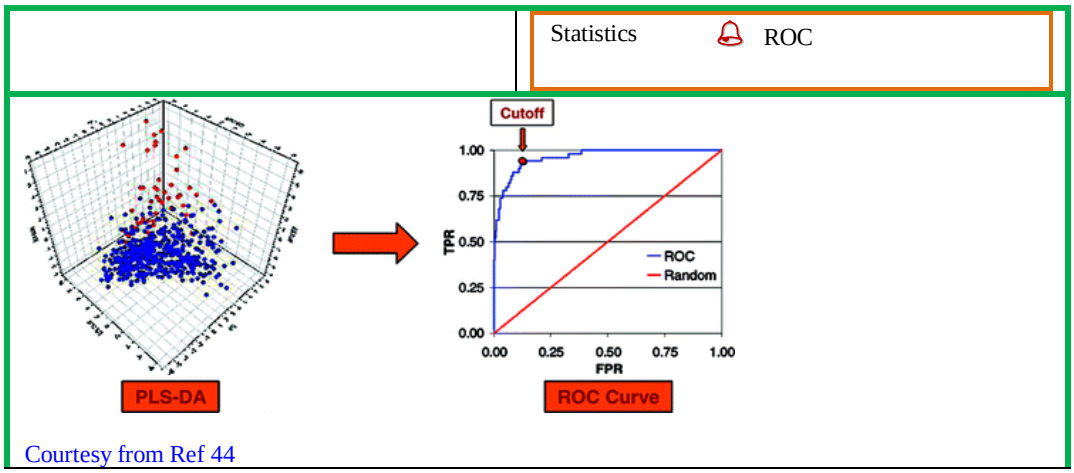
SInhR (Structure Inhibition Relationships)

PDE4 inhibition: Rizzi and Fioni [44] employed discriminate analysis of PLS components for a variety of descriptors to pick up PDE4 inhibitors from a pool of drug like compounds (chart 2.14). The DRAGON descriptor set from 2D-structure performed better as is evident from ROC curve analysis.

Chart 2.14: Sbioconcentration_factors R — SActR

Descriptors
- DRAGON
- EVA descriptors
- QikProp

Response	PDE4 inhibition
Task	Virtual screening
	Classification
Model	PLS_DA



Zheng et al. [155] made an exhaustive predictive SXR modeling with mono- and bis-quaternary ammonium salts (42 in numbers) which are antagonists to neuronal nicotinic acetylcholine receptors (nAChR) (chart 2. 15). These compounds mediate nicotine-evoked dopamine release. This project resulted in discovery of six molecules with IC_{50} measured values less than 0.1 µM at nAChR subtypes.

Chart 2. 15: Predictive SXR with WHIM descriptors

Descriptors	Model/Training	
	r^2	r^2_{cv}
log ($1/IC_{50}$)	0.76	0.64

% Explainability	Descriptor
~44.0	Length of the *N*-alkyl chain attached to the quaternary ammonium head group
~20.0	Log P Moriguchi
~13.0	Molecular surface area,
~12.6	WHIM index/unweighted first component shape directional
~7.8	Molar refractivity Ghose–Crippen
2.6	LUMO energy

Test set

Correct classification ▶ Fourteen out of 18	18 molecules with IC_{50} values >1 µM.
Prediction of activity ▶ 12 among 13	Newly synthesized biologically important compounds 11 Active antagonists 02 Inactive

Inf.Bits.
SXR_NN model is an apt aid in drug discovery

Inhibition activity: Mattioni and Jurs [100] developed SInhibitionR computational_NN models using molecular descriptors as explanatory variables (chart 2.16). CYP1A2 is an important member of CYP450 superfamily and accounts for 15% of total CYP450 presence in human liver. A large number of SActR

models with 7500 compounds using molecular descriptors are studied using NNs, SVR and domain approach (chart 2.17) for inhibition activity [52]. Franke et al. [83] performed a multiphase screening of compounds of increasing activity of inhibitors of human 5-lipoxygenase (5-LO) (chart 2.18) showing the feasibility of lead structure identification through natural-product-derived screening libraries. CYP450 1A2 inhibitory modeling by Novotarskyi et al. [52] (chart 2.19) and adenosine kinase inhibitory activity by González et al. [24] (chart 2.20) employed descriptors as causative factors.

Chart 2.16: SInhibitionR models with molecular descriptors SActR

Descriptors	
Electronic	Topological
Geometric	

Models	Classification
MLR	kNN
Computational NN	

Enzyme	I#-H#-O#	Training		Prediction	
		Rms	R2	Rms	R2
CA-I	8-5-1	0.105	0.994	0.208	0.98
CA-II	9-5-1	0.140	0.992	0.231	0.971
CA-IV	8-5-1	0.147	0.992	0.221	0.991

Chart 2.17: SInhibitionR SActR

Descriptors	
E-state	IDIDA-SMF
Dragon	

Data source	PubChem BioAssay database	
Data set	3745: inhibitors	3741:-non inhibitors

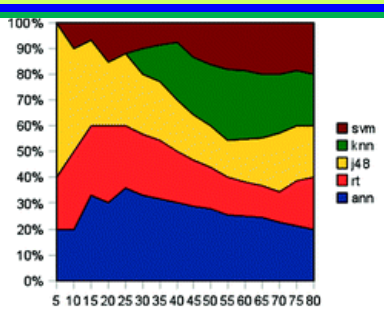
Machine Learning algorithms (MLAs)	
Associative NN	k-Nearest neighbors
Random Tree	C4.5
Domain approach	SVR
Fragment-based	

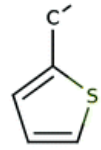
	% Correct classification	
Test set		Domain approach
Internal	83	90
External	68	78

Response (Measured)	CYP450 inhibition activity
Task	Prediction

80 in-silico SActR models
- Same meta procedure
- Descriptors
- Machine Lrn methods

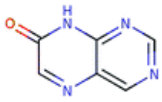
Inf.Bits.
Domain approach >> Best models





2-methylthiophene

CYP1A2 Inhibitors: 151
CYP1A2 Noninhibitors: 14
Model Sensitivity: 100%
Model Specificity: 43%



pteridin-7-one

CYP1A2 Inhibitors: 905
CYP1A2 Noninhibitors: 74
Model Sensitivity: 100%
Model Specificity: 5%

Courtesy from Ref 52

Chart 2.18: Sequential ligand-based approach a novel class of 5-LO inhibitors

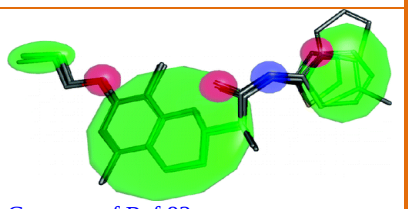
Step:	**One**	**Input:** 430 compounds	 Courtesy of Ref 83
Step:	**Two**	**Scaffold-hopping:** Search for similarity with topological (**CATS** 2D) pharmacophore descriptor **Output:** 18 compounds	
Step:	**Three**	seed ⟵ Find two most potent molecules **Ligand-based virtual screening:** Analyze with focused natural-product-derived combinatorial library **Output:** Best molecules	
Step:	**Four**	+ Molecules with potently suppressed 5-LO activity in intact cells	

Chart 2.19: Molecular descriptors in CYP450 1A2 inhibition modeling

Descriptors_ packages	Descriptors
Dragon	ISIDA
ISIDA	▶ fragments-base descriptor
	Dragon
	▶ 2D topological descriptors (E-state)
	▶ 0D – 3D

Chart 2.20: Adenosine kinase inhibitory activity

X(MolDes)			Adenosine kinase inhibitory activity (32,1) = fn(Desc)
Geometric	Constitutional	GETAWAY	Expl: <70%
2D-auto correlation	BCUT	3D-Morse	
WHIM	GETAWAY	Topological	
❖ TOPological Sub-Structural Molecular Design (TOPS-MODE)			Expl >77% Compounds: pyrrolo[2,3-d]pyrimidine nucleoside analogues

BBB receptor crossing: Cammarata et al. [102] proposed NNs models for SActR of BBB receptor crossing of glycine/NMDA receptor antagonists with molecular descriptors (chart 2.21).

Chart 2.21: SInhibitionR NN models with molecular descriptors

Descriptors		Models	Descriptor pruning
Electronic	Topological	MLR	GA
Geometric	Polar surface	Computational NN	
		RBF-NN	

Compounds	109

Response	log(IC50)	Range : -0.57 to 4.5

Soybean 15-Lipoxygenase inhibitory activity: Camargo et al. [111] reported MLR and PLSR models for SXR between Inhibitory activity of Soybean 15-Lipoxygenase and molecular descriptors (Chart 2.22).

Chart 2.22: SInhibitionR NN models with molecular descriptors

Descriptors		Models	Descriptor pruning
Quantum chemical	Topological	MLR	GA
Geometric		PLS	

Solvent-accessible surface area; Average distance/distance degree descriptor; LUMO

Compounds	Mono-, di-, and trisulfides derived from the essential oil of garlic
Response	soybean 15-lipoxygenase inhibitory activity

SBindR (Structure binding Relationships)

Adenosine receptors: González et al. [19] proposed TOPological Sub-Structural Molecular Design (TOPS-MODE) procedure for prediction of affinity for A1 adenosine receptors with explainability of 79% of variance of binding (chart 2.23).

Protein-flavonoid interaction: Rolo-Naranjo et al. [107] employed protein geometrical parameters and a quantum mechanical descriptor, Binding Local Site (BLS).

PocketAlign algorithm: Yeturu and Chandra [110] introduced PocketAlign, a new algorithm to probe into structural superpositions in binding sites. This approach encodes shape descriptors from geometry and classification of chemical groups (Chart 2.24).

Binding to Serotonin Receptors: Hong et al. [53] selected ligands for binding to serotonin Receptors 5HT1E and 5HT1F (chart 2.25) by CODESSA descriptors.

PTK binding by flavonoids: Oblak et al. [42] put forward SActModel for biological activities of binding flavonoids to PTK using descriptors generated from CODESSA (chart 2.26).

Binding affinity of corticosteroids: Maw and Hall et al. [57] reported predictive modelling of binding affinity of corticosteroids with E-state and shape descriptors (chart 2.27**).**

Chart 2.23: TOPological Sub-Structural Molecular Design of A1 adenosine receptors

X(MolDesc)		
TOPological Sub-Structural Molecular Design (TOPS-MODE)	**Response:** Affinity of A1 adenosine receptor	Compounds : 32
Model : MLR Explainabilit: 79% of the variance		

1) BCUT	**2)** Constitutional	**3)** Topological
4) Geometrical	**5)** Molecular walk	**6)** RDF

Inf.Bits.

- Contribution of biological activity to different fragments
- Direct structural interpretability

Chart 2.24: PocketAlign

+ Alignment exploration through different schemes

Captures similarities in abilities for ligand recognition

+ Probing into similarity of sites
+ Detecting residues and atoms contributing to similarity

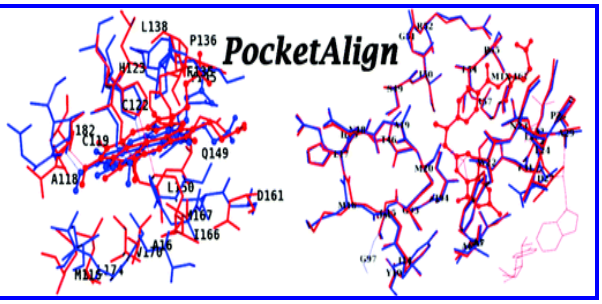


Chart 2.25: S Serotonin Binding R — SBindR

Descriptors Packages
- DRAGON
- MolConnZ
- MOE

Response:
Binding to Serotonin Receptors 5HT1E and 5HT1F

Task : Selectivity of Ligands

Descriptors Pa
- Frequent Subgraphs
- Molecular Hologram Fingerprints

Models
- **k Nearest Neighbor**
- **Automated Lazy Learning (ALL)**
- **PLS**

Chart 2.26: S logCMC R — SPropR

Descriptors
- CODESSA
- Electrostatic
- Quantum chemical

Compounds: 104 flavonoid derivatives

Response : $p56^{lck}$ protein tyrosine kinase (PTK) inhibition

Model
Descriptors : Orthogonalization
Cause-effect relation: MLR

Inf.Bits

Influencing factor in binding flavonoids to PTK is maximal total interaction for a C−O bond

Chart 2.27: Corticosteroid binding globulin

Descriptors
- E-state ▶ Kappa shape index
- Hydrogen E-state descriptors

Molecules — 31 steroids
Response (Measured) — Binding affinity
Task — Prediction

SRespR_ models				
	r^2	s	r^2_{press}	s_{press}
	0.81	0.51	0.72	0.62
Prediction	0.70 : r^2_{LOO}			

Leave-group-out (LGO) **for model-validation**

- Data set divided into random groups of 20% each
- Remove each observation three times from data set
- Cal residuals

Inf.Bits.

- E-state model predicts pK of binding values for unexplored compounds.

🕮 SBBBR (Structure Blood brain barrier Relationships)

BBB concentration ratio in vivo: Rose, Hall and Kier [68] estimated logarithm of blood brain barrier concentration ratio in vivo of around twenty thousand compounds without resorting to 3D-structure from SlogBBBR model (chart 2.28).

Chart 2.28 SActR

Descriptors

- E-State index for hydrogen bond donors
- Hydrogen E-State index for aromatic CHs
- Valence molecular connectivity index (second order difference)

Inf.Bits.	
Molecules that penetrate BBB have	
Descriptor values	Groups
Large HST(arom) values	Aromatic groups
Small values of HST(HBd)	Fewer or weaker H−Bond donors
Smaller d2χv values	Less branched molecules with fewer electronegative atoms

Descriptor	**Encodes**	
	MAE	RMS
Test	0.38	0.47
Prediction	27 correct out of 28	
Estimation of log BB	Drugs & drug like compounds: 20 039	

🕮 SInteractR (Structure Interaction Relationships)

Drug–protein interaction: Cao et al. [80] reported PyDPI (drug–protein interaction using Python software) to compute structural and physico-chemical descriptors for proteins and peptides from amino acid sequences. These descriptors are compared with those standard packages (chart 2.29)

Chart 2.29: Molecular packages for comparison with PyDPI	
Molecular Operating Environment (MOE)	Chemical Computing Group
MODEL	Molecular Descriptor Lab
Kappa shape indices	kappa module

SPharmacoKineticR: Yap et al. [157] proposed predictive SR models (chart 2.30) for total clearance of a compound using descriptors from DS-WHIM, DS-GATAWAY etc.

Chart 2. 30: Predictive SPharmacoKineticR

Compounds : 503	
Descriptors	
DS-Mixed	DS-WHIM
DS-3DMoRSE	DS-ATS
DS-RDF	
Y : total clearance (CL_{tot}) of a compound	

Model	*Average-fold errors*
GRNN	1.63 to 1.96
SVR	1.66 to 1.95
k-nearest neighbour	1.90 to 2.23

🕮 SReceptR (Structure Receptor Relationships)

Benzodiazepine receptor: Carotti et al. [162] classified 2-phenyl-2,5-dihydropyrazolo[4,3-*c*]quinolin-3-(3*H*)-ones (PQs) into subsets (full and partial agonists, antagonists, and inverse agonists) of varying efficacies using overlapped hyperspace of molecular descriptors (chart 2. 31).

Adenosine receptors: Sharma et al. [16] applied modified BCUT descriptors in modeling 2-chloro-N6-substituted-4′-thioadenosine-5′-uronamide compounds which have agonistic activity against human A3 adenosine receptors (chart 2.32).

González et al. [159] studied A_1 adenosine receptors agonist effect and A2B agonist effect making use of descriptors as explanatory factors (chart 2.33). Marzio, Todeschini et al. [164] proposed MLR of acute response in freshwater fish species for nine congeneric aromatic hydrocarbons using WHIM descriptors. Also, it was found that linear regression with Hansch's parameters has adequate predictive ability for non-polar compounds.

🕮 SReceptors_AffinityR:

Vitamin D receptor: González et al. [12] performed SActR modeling of vitamin D receptor affinity of 1α,25-dihydroxyvitamin D3 analogues (chart 2.34).

Adenosine receptors: González et al. [8] used BCUT descriptors for explaining more than 80% variance in experimental agonist effect of adenosine analogues towards A3 adenosine receptors (chart 2.35).

Adenosine analogues: González et al. [20] reported a predictive A2B agonist model of explainability 70% with RDF descriptors. The data set consisted of 89 adenosine analogues and any other type of descriptors could not account more than 47% of variance in response (chart 2.36).

Adenosine receptor: González et al. [22] proposed a MLR for affinity of A1 adenosine receptor explaining 84% of variance. The other models using different kinds of descriptors could reach only 75% explainability (chart 2.37).

SReceptors_agonistR: González et al. [14] made use of molecular descriptors for N^6–arylcarbamoyl, 2-arylalkynyl-N^6–arylcarbamoyl, and N^6–carboxamido derivatives (29 compounds) in investigating A_{2A} adenosine receptors agonist effect. The model explained >78% of variance in the experimental activity (chart 2.38).

SReceptors_agonistR: González et al. [158] compared the efficacy of different types of molecular descriptor in SActR predictive modeling of A2A adenosine receptors agonist compounds (chart 2.39).

Dopamine antagonists: Kim et al. [11] classified dopamine antagonists based BCUT and MOLCONN-Z descriptors using SIMCA procedure (chart 2.40).

Chart 2. 31: Classification of High affinity central benzodiazepine receptor ligands in molecular descriptor space

3D-WHIM CLOG P	Classification PCA K-nearest neighbors LDA
Inf.Bits. LDA >k-nn	

Chart 2.32: Agonistic activity of human A3 adenosine receptor ligands

X(MolDesc)	
Software: DRAGON	488 Topological descriptors
1) BCUT _modified	**2)** Functional group classes
Response: Human A3 adenosine receptor agonistic activity of 2-chloro-N6-substituted-4′-thioadenosine-5′-uronamides	**Model:** Combinatorial Protocol in Multiple Linear Regression (CP-MLR)

Significant descriptors			
Average valence connectivity index of order zero	Sum of topological distances between O and Cl, T (O···Cl) from the TOPO class	Lowest eigenvalue n.2 of Burden matrix/weighted by atomic masses, , from the BCUT class	Number of secondary aliphatic amides, nCONHR, from FUNC class

Chart 2.33: Descriptors for adenosine receptors agonist compounds

Model GTA:	Geometrical	Topological	Atom-Weights AssemblY (GETAWAY)
Response: A_1 adenosine receptors agonist effect		Explainability: 77%	N^6-arylcarbamoyl, 2-arylalkynyl-N^6-arylcarbamoyl, and N^6-carboxamido derivatives: 32
Model TG	Topological	Galvez Topological Charges indexes	
Model RGW	Randić	Geometrical	WHIM
Inf.Bits. R^2 of Models TG and RGW 0.70 ie. failure of model			

Chart 2. 33b: Descriptors for adenosine receptors agonist compounds

Response: A2B agonist effect			adenosine analogues: 89	
Constitutional	Topological	Galvez Topological Charges	GETAWAY	**Models** **RDF Desc** Explainability: 77%

3D Morse	Randić	Geometrical	WHIM	– Alt_models with other types of Descriptors fail (Expl : 47% variance
BCUT	2D autocorrelation	Molecular walk		

Chart 2. 34: Molecular descriptor modeling of vitamin D receptor affinity

WHIM	Topological Galvez charge indices	2D autocorrelations	BCUT
	Response vitamin D receptor affinity	Compounds : 86	
Inf.Bits. Model_explainability: WHIM (71%) >> others (38%)			

Chart 2.35: Model for antimalarial activity against P. falciparum with molDesc

X(MolDesc)

1) BCUT	**2)** Galvez topological charges indexes	**3)** Randic's molecular profiles	**4)** Geometrical	**5)** Topological
		Resp = fn(BCUT) ; Failed Models		

Response Predicting affinity toward A3 adenosine receptors	**Compounds** 32 adenosine analogues	>80% of variance	With molDes (2 to 5)

Chart 2.36: Molecular descriptors of low explainability for A2B agonist effect

X(MolDesc)

10) BCUT	**11)** WHIM	**12)** Topological	**13)** Constitutional	**14)** Molecular walk count
15) Geometric	**16)** GETAWAY	**17)** QC	**18)** 3D Morse	**19)** 2D autocorrelations,
20) Galvez topological charge indices				

Compounds 89 adenosines	
Response: A2B agonist	

	% Variance explainability
Best model : Resp = fn(**RDF**)	70
Models [1 TO 11]	<47

Chart 2.37: Molecular descriptor modeling for N6-(substituted-phenylcarbamoyl) adenosine-5′-uronamides analogues

X(MolDesc)	
1) BCUT	**2)** Randic
3) Geometric	

Compounds:	89 adenosines
Response:	A2B agonist

Best model	% Variance explainability
TOPological Sub-Structural Molecular Design	84
Models [1 TO 11]	75

Chart 2.38: Descriptors for adenosine receptors agonist compounds			
GETAWAY	3D-MORSE	WHIM	2D
Geometrical	BCUT	Galvez Topological Charges indexes	Radial Distribution Function
Topological	Randić	Autocorrelation	
Inf.Bits.			
Bulkiness and stereoselectivity are key factors in affinity for A_{2A} adenosine receptors agonist compounds			

Chart 2.39: Descriptors for adenosine receptors agonist compounds			
Model GTA	Geometrical	Topological	Atom-Weights AssemblY
Response: A_{2A} adenosine receptors agonist effect			N^6-arylcarbamoyl 2-arylalkynyl-N^6-arylcarbamoyl N^6-carboxamido } derivatives : 29
Inf.Bits.			
Model GTAexplainabiliity 77% >> $\left[\left\{\begin{array}{l} Topological \\ Galvez\ Topo\_Charges \\ Geometrical \\ WHIM \end{array}\right\} descriptors = 70\%\right]$			

Chart 2.40: Model for antimalarial activity against P. falciparum with molDesc		
X(MolDesc)		
1) BCUT	**2)** Molconn-Z	
		Models
Response: predicting affinity toward A3 adenosine receptors	Compounds: 1328 Activity: Dopamine antagonists database: MDDR	SIMCA 80% Success
Inform.Bits		
Functional feature hypotheses + topological descriptors better in classification		

SProtLigIntR (Structure Protein-Ligand interaction relationships)

These investigations comprise of interactions of simple proton, water molecule or small ligands to drugs, proteins and other bio-molecules. Many of them are essential for life processes, while some others result in life threatening scenarios.

Ligand–receptor interactions: Ijjaali et al. [10] used BCUT descriptors to choose a firm for purchase of compounds relevant to ligand–receptor interactions.

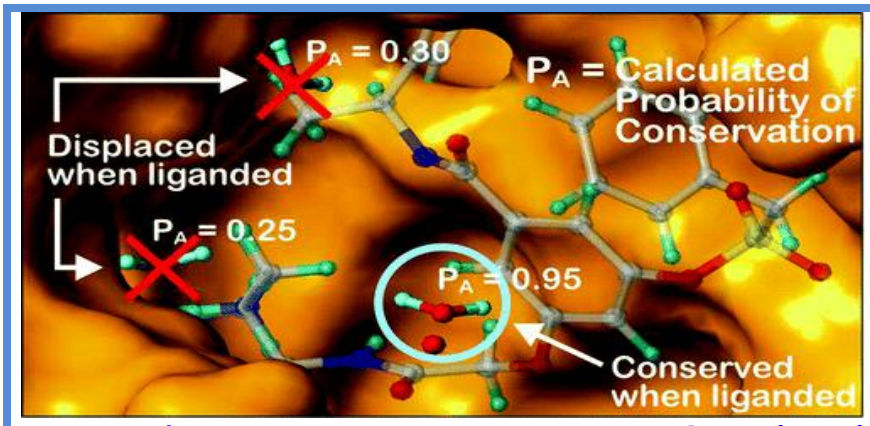


Fig. 2.1: Bound water in proteins Courtesy from Ref 96

Putative proteins: Amadasi et al. [96] estimated both conserved (making at least two hydrogen bonds) and non-conserved water molecules in putative proteins using geometric descriptor (rank) and HINT score. The bound water molecules gain 0.6-2.0 Kcal mole^{-1} of binding energy (fig. 2.1).

🕮 SDrugR (Structure Drug relationships)

β-blockers: Quigley and Naughton [144] correlated physico chemical parameters of β-blockers employing molecular descriptors with success (chart 2.41).

Chart 2.41: Variation of physico chemical parameters of β-blockers with molecular descriptors

Descriptors		Molecules	β-blockers
Connectivity index	Order 1 and 2	Response	log P from ClogP program
	Simple	Method	Catalase protection assay
	Valance		
	Differential		
Wiener	Hyper-Wiener		

Descriptor: Equilibrium electronegativity

Basis: SEMO ; Topological concept

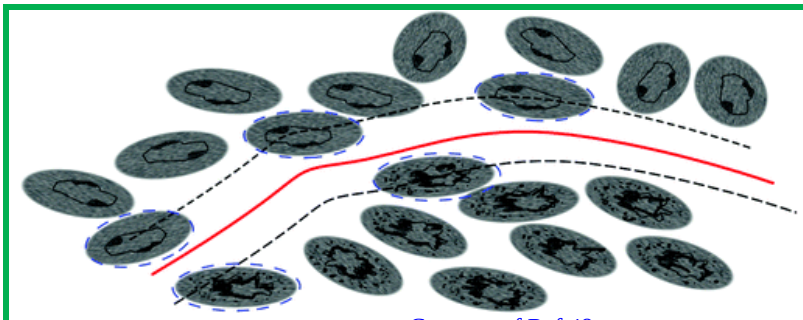
1) Cal Group electronegativities (A) taking into consideration of chemical environment
2) Cal relative bond lengths(B), equilibrium electronegativity (C)
3) SEMO_TOPOL_Desc = fn(A,B,C)

Phospholipidosis: Many cationic amphiphilic drugs intervene with cells and the processes resulting in excess accumulation of phospholipids and this adverse effect is called Phospholipidosis. It is detected by electron microscopy of cells by the presence of lamellar inclusion bodies. This is one hurdle in development of drug development.

Lowe et al. [48] (chart 2.42) predicted phospholipidosis employing a larger data set using DRAGON software for molecular descriptors.

Chart 2.42: SBioProcessR SProcessR

Descriptors		Models
🕮 DRAGON	**Bio-process: phospholipidosis**	**RandomForest**
🕮 E-descriptors	**Task : Prediction**	**SVM**



Courtesy of Ref 48

Inf.Bits

Phospholipidosis Models

- Efficient with circular fingerprints
- Low quality with E-Dragon descriptors or a hybrid ones

🕮 S Seq R (Structure sequence relationships)

The variation in properties of tri- and higher peptides with change in amino acid sequence is a hard core and is in borderline between biophysical and biochemical research. A few literature reports indicate the success stories of employing molecular descriptors.

🕮 SGenetR (Structure Genetics relationships)

Chromosomal aberrations in Chinese hamster lung cells: McElroy et al. [108] employed molecular descriptors in the classification of organic compounds in the context of chromosomal aberrations in Chinese hamster lung cells (chart 2.43).

Mutagenicity halo acetic acids: Pérez-Garrido et al. [151] predicted direct mutagenicity haloacetic acids in drinking water as chlorination by-products making use of spectral moment descriptors and others (chart 2.44). The experimental data for TCAA—Trichloroacetic acid, BDCAA—Bromodichloroacetic acid, and TBAA—Tribromoacetic acid confirm the validity of predictive SmutaGenR models.

Chart 2.43: SChromosomeR NN models with molecular descriptors SGenetR

Descriptors

- Electronic
- Topological
- Geometric
- Polar surface

Classification Models

- GA
- kNN

297	Compounds causing chromosomal aberrations in Chinese hamster lung cells	
		Aberrant cells
Response	Clastogenic	>10%
	Nonclastogenic	<5%

	% correct classification	
Compounds	Training	Prediction
297	86.5	80.0
279	85.7	85.7
182	91.5	94.4

Chart 2.44: Descriptors for Mutagenicity haloacetic acids

- Geometrical
- RDF
- WHIM
- Eigenvalue_based
- 2D_autocorelation
- Information

Bio concentration factors: Bordás et al. [43] used DRAGON descriptors in developing predictive models for bio-concentration factors (BCFs) of POPs (persistent organic pollutants) in different plants with ensemble NN models (chart 2.45).

Table 2.2 incorporates a few test cases of activity data sets analysed with molecular descriptors and supervised regression methods.

Chart 2.45: Sbioconcentration_factors R SActR

#Descrip	Package/type
1660	DRAGON
150	VolSurf
11	QC

Predictive models
- MLP_NN-BP
- Ensemble
  - 10-fold (MLR)
  - 10-fold leave-10% -out

Response (Measured): BCFs of POPs in different plants

Task: Model

Compounds: 20 Polychlorinated dibenzo dioxins/dibenzofurans; 14 Polyhalogenated biphenyls

	Training			Validation	
Regression	#	R^2	q^2	DE	F
Model-1	27	0.940	0.922	0.155	392.1
Model-2	27	0.921	0.898	0.161	140.4

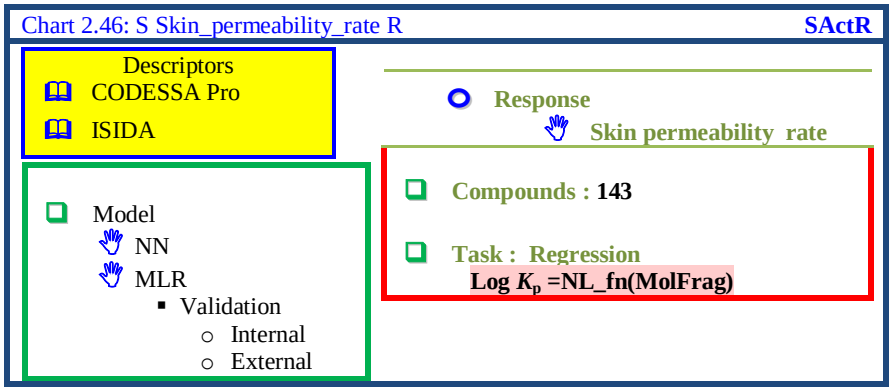
Table 2.2: Typical studies in S Act R

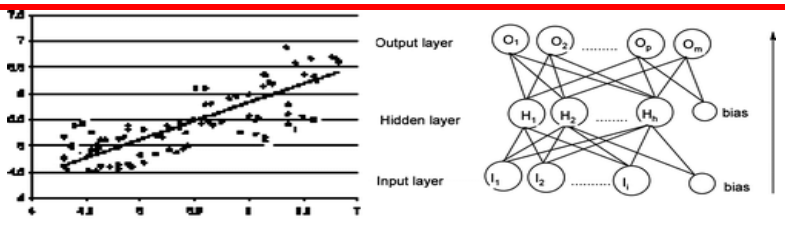
Activity	#	Molecules	Descriptors	Model
Inhibitory effect		Benzothiazole	• Classical variables	○ Stepwise ○ regression
Muscarnic activity	2000	Compounds	• TOPO	○ Trend vector ○ analysis
Enzyme P56 Protein tyrosine kinase	105	Flavonoid inhibitors	• QC • Classical • physico-chemical	○ CNN
Herbicide PI 50	30	Triazines	• TOPO • Orthogonalized	○ MLR ○ PCA ○ Dominant ○ Component ○ analysis
Human intestine absorption	86	Drugs	• 728 Descriptors • CORINA • HYPERCHEM • ADOPT MOPAC	○ MLR ○ CNN
IC50	37	Bis amidines Isohelical pentamidines	• Similarity index • Charge index • Shape index	○ MLR

Log(1/C)	39 60	Carboquinones Benzodiazepines	• LogP • MR	NN
BR	256	Diaminodihydrotriazines	• Hydrophobicity • Size • Hammett σ	NN
BR	128 129 130	4-substituted Guanidino Thiazole	• Wiener • χ	Correlation
Carcinogenic activity binary(exo-or-endo-)	11	PAH	• Ordinal data • NMR ^{13}C	
Mutagenic activity		Nitro sub PAH	• MW • # N • #Ar rings • #methyl groups • Estate index	
Base line toxic Pol narcosis LC50	172		• QC	PLS MLR

Contributions of Katrizky

Katritzky et al [26-32, 35-38, 40-41], developer of CODESSA from University of Florida, reported predictive skin permeability rate (chart 2.46), solvent effect on decarboxylation of 6 nitrobenzisoxazole-3-carboxylic acid (chart 2.47), molecular adduct formation of β-cyclodextrin with organic guest molecules (chart 2.48), thermodynamic free energy relationships (chart 2.49) for solvation of 80 solutes in varying number of solvents (upto 80), SToxR multi-parametric models for aquatic toxicity of Poecilia reticulata (guppy) (chart 2.50), six-parametric SpropR model (chart 2.51) for melting points of ionic liquids, six parametric SXR model for gas phase lithium cation basicity of organic compounds with quantum chemical descriptors (chart 2.52), multi- (Lipophilicity Hydantoin by HPLC-RT, fragment descriptors in SPropR regression models of cmc of anionic surfactants (chart 2.53), critical micelle concentration (chart 2.54),4-variate linear models for boiling point of azeotropic mixtures (chart 2.55), and melting points of potential ionic liquids with CODESSA (chart 2.56).





Multilinear Regression model Neural Network model

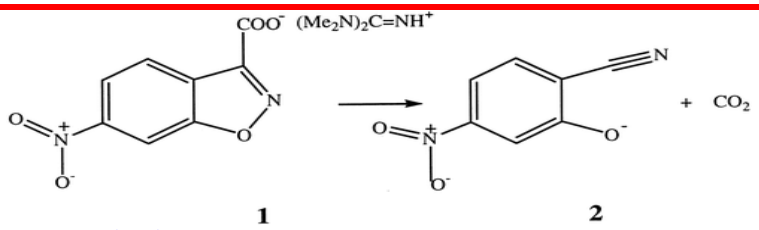
ISIDA LogKp = f (molecular fragments)

Courtesy of Ref 27

Chart 2.47: S Reactivity **R** **SReactR**

Descriptors
CODESSA

- Response
  - decarboxylation rate
- Compound : 6 nitrobenzisoxazole -3-carboxylic acid
- Solvents: 24
- Task : Solvent effect on rates



1 2

Courtesy of Ref 47

Chart 2.48: S Complex **R** **SReactR**

Descriptors
CODESSA Pro

Model statistics			
Model	R^2	R_{cv}^2	#
7 Parameter	0.796	0.779	218
Fragment trail	0.943	0.848	195

- **Response**
  - Complexation

Complex

- **218 organic guest molecules + β-cyclodextrin**
- **Task : Molecular complex reactivity**

Inf.Bits

- Fragment trail + multi-parametric model
  + Better pragmatic utility

Chart 2.49: S Solvation_FreeEnergies **R** *SΔGR*, **SSolR, SThermodR**

Descriptors
CODESSA Pro

Model statistics	
Compounds	R^2
n-propylamine	0.996
Toluene	0.604 (Lowest)

- **Response**
  - **Complexation**
- **Compounds**
  - **80 solutes**
- **Solvents**
  - **15 to 82**
- **Task : Regression**

Chart 2.50: S Aquatic_Toxicity R **SToxR**

Descriptors
CODESSA

Model statistics		
Model parameters	Toxins Class	R^2
2	1	0.96
5	2	0.92
5	3	0.85
--	4	unsatisfactory

- **Response**
  - **Aquatic Toxicity of Poecilia reticulata (guppy)**
- **Compounds**
  - **Environmental Pollutants**
- **Solvents**
  - **15 to 82**
- **Task : Prection Regression**

Chart 2.51: S MeltingPoint R SPropR

Descriptors
CODESSA

Model statistics	
Model parameters	R^2
6	0.788
2	0.0.713

- **Response**
  - **Melting point of Ionic liquids**
- **Compounds**
  - **126 Pyridinium bromides**
- **MP range**
  - **30−200 °C**
- **Task : Prection, Regression**

Chart 2.52: S Basicity R SParR

Descriptors
CODESSA

- Response
  - Gas-phase lithium cation basicity

Model	R^2	R^2_{CV}	F	s^2
6-parameter	0.801	0.785	133.11	8.78

Inf.Bits

+ Descriptors have encoded knowledge of
  - Reaction equilibrium
  - Electrostatic interaction between a lithium cation and a base

Compounds
- 205

Descriptors
- Minimum net atomic charge
- HOOMO energy
- Total point-charge component of μ
- ....

Chart 2.53: S Lipophilicities R — SPropR

Descriptors
CODESSA Pro

derivatives
- Geometry Opt
  - AM1

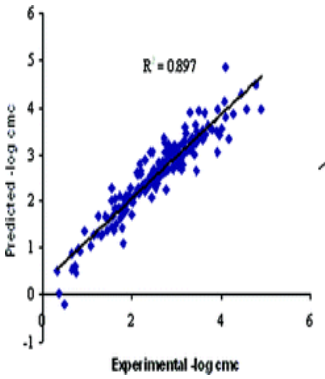
Response
RP_HPLCRT ⟶ Lipophilicities

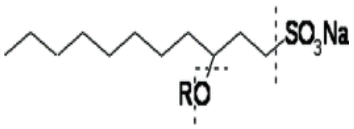
Inf.Bits
- Lipophilicities have higher correlation with molecular descriptors
- Entropy effect predominates

Chart 2.54: S logCMC R — SPropR

Descriptors
CODESSA Pro

Compounds: 181 anionic surfactants
Response : critical micelle concentration





Inf.Bits
- Fragment approach is superior
- Intermolecular interactions between anionic surfactants and water (consequently cmc) explained by
  - Hydrophobic fragments expressed by topological and geometrical descriptors
  - Hydrophilic fragments expressed by constitutional, geometrical, charge descriptors

Chart 2.55: S BoilingPoint R — SPropR

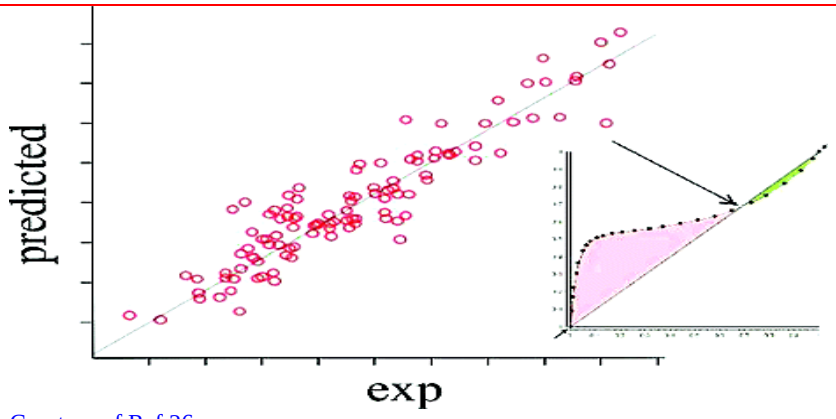
Descriptors
CODESSA Pro — 56 topological indices

- Model
  - four-descriptor MLR
- Validation
  - Internal
  - External

- Property:
  - Boiling point
- Compounds : 426 azeotropic mixtures
- Task : Regression

Inf.Bits

- Descriptors are related to three components of the enthalpy (heat) of vaporization



Courtesy of Ref 26

Chart 2.56: S logCMC R SPropR

Descriptors
CODESSA

Compounds: 104
Response : Melting point (MP)

Compounds #	Set	R^2	Model
57	A	0.7442	5-Parmater equation
29	B	0.7517	
45	D	0.6899	
18	C	0.9432	3-Parmater equation

Set D: Benzimidazolium bromides

Sets A,B,C:N-substitution in imidazolium bromides

Inf.Bits

- Size and electrostatic interactions in the cations are major contributors for MP

SRespR (Structure Response Relationship)

The response of an instrument for a chemical/biological system revolutionized the knowledge processing to understand, predict and arrive at materials of desired characteristics within reasonable limits. The tensorial order of data (zero to three) in standalone and hyphenated instruments per compound is a signpost of the intricate details of the structure mirrored in response. The studies pertaining to one compound are primarily aimed at chemical structure elucidation. The similarity and diversity investigations involve a small to a large set of compounds. Quantitation on the other hand, deals with one or many compounds at different concentrations. In table 2.3, the investigations reported on ^{13}C NMR and chromatography using molecular descriptors is presented.

Table 2.3: A few typical studies in S Res R

Property	#	Molecules	Descriptors	Model	Software
HPLC RI	126	PAHs	• Total energy • Polarisability • IP	MLR AMI	MOPAC

			• μ • Sub polarity	○ MLRI ○ MNDO ○ PM3	
NMR ^{13}C	24	Trn	• 24 Descriptors	○ MLR ○ NN-BP	ADAPT
NMR	13	Compounds	• 150 Topo • Geo • Elec	○ MLR(10) ○ NN	
Structural variation	154	PCB	• 52 QC	○ PCA	HYPERCHEM

For, ^{13}C NMR data, the quantum chemical (charge) and environment (nearest neighbour) descriptor play an important role. In chromatography, the characteristics of the stationary, mobile phases and the compounds dictate the type of descriptors to be used as explanatory variables. Size and ionization potentials are of prime importance in modelling MS response. However, many other descriptors of different types enhance the power of the model and accuracy of prediction. Thus, the cycle of descriptor research jargon is repeated and the best model is proposed based on case-by-case inspection. The experimental response with sophistication in instrumentation and automation is reliable only in the hands of pure experimentalists of the time tested laboratories. That is why the response data already published or compiled by authentic agencies are the source of data in molecular descriptor activity.

⊥ SChromatR (Structure Chromatography Relationship)

Aldehydes: Ren et al. [131] applied MLR in structure GC_ response_ relationships of aldehydes and ketones on different polar stationary phases (chart 2.57).

Saturated esters: Liu et al. [132] put forward structure GC retention relationship model for a series of saturated esters on several stationary phases with MLR with acceptable statistics (chart 2.58)

Flavor/fragrance Molecules: Rojas et al. [133] [used DRAGON software to model GC retention index of fragrance and flavors compounds with nearly five thousand molecular descriptors of various categories (chart 2.59).

Oxyethylenes: Voelkel et al. [166] proposed topological descriptors in SGC_ Reten_ Relationship model of oxyethylene derivatives (chart 2.60).

Heterocyclics: Farkas et al. [161] predicted gas chromatographic retention indices for heterocyclic (O-, N-, and S-) compounds on standard nonpolar polydimethyl siloxane stationary phase employing molecular descriptors (chart 2.61).

Oxygenated organic molecules: Liu et al. [132] found polarizability based descriptors could explain variation in Kovat retention index for a set of diverse oxygenated organic structures (Chart 2.62**)**.

Reversed-phase fused-core HPLC: D'Hondt et al. [148] developed peptide_ SRetentionR for 16 chromatographic systems using WHIM, 3D-MoRSE, RDF and log P descriptors of amino acids.

Reversed-phase HPLC_ pesticides: D'Archivio et al. [150] found inferiority of WHIM and GETAWAY descriptors compared to quantum chemical or physico-chemical parameters (log kow and pKa) in SRetentR of pesticides.

HPLC_RT: Novotarskyi et al. [52] put forward ensemble MLR and MLP_NN models for retention indices of HPLC data with SIR, state descriptors etc. (chart 2. estate-18) as explanatory variables (chart 2.63).

The variation in variation GC retention index values for alkylbenzenes (chart 2.64), Chlorinated dibenzodioxins [140] (chart 2.65), fatty acid methyl esters [137] (chart 2.66) and Terpenoids [138] (chart 2.67) are reported. Selection of subset of compounds with similar or orthogonal character is calculated with WHIM (chart 2.68).

Similarity of chromatographic systems: Gyseghem et al. [156] selected a set of 10 compounds with orthogonal properties using WHIM descriptor data. Earlier, a set of 68 active pharmaceutical substances were tested for orthogonality/similarity of chromatographic systems.

NLO response with QC descriptors**:** Xu et al. [152] applied 3D-SXR models for nonlinear optical responses of organic chromophores with quantum chemical descriptors. The R2 values for prediction of β and γ *are* 0.943 and 0.963.

Chart 2.57: Atom based descriptors in SGC_ response_ Relationships

Descriptors		Molecules	Aldehydes and ketones
Atom-type-based AI topological		Response (Measured)	GC retention data (I_R)
Xu index		Polar stationary phases	HP-1, HP-50, DB-210, and HP-Innowax
Model	MLR		
Statistic	r>0.995		

Inf.Bits.

- Molecular size has a dominant contribution to I_R
- Atom types or groups have smaller effect
- Order of correlation with Kováts retention indices is [Topological > physicochemical > quantum-chemical]

Chart 2.58: Saturated esters SGC_ Reten_ Relationships

Descriptors		Molecules	Saturated esters
Polarizability effect index		Response (Measured)	GC Kováts retention index
Odd–even index		Stationary phases	SE-30, OV-7, DC-710, OV-25
Steric effect index			XE-60, OV-225, Silar-5CP
Model : MLR			
CC	0.9989 and 0.9977		
LOO-CV	0.9988 and 0.9975		
PredError	0.5–0.7%		

Inf.Bits.

- Contributing factors to RI of structures with polar groups
  - Dominat: Polarizability, molecular size, branching
  - Influencing: Steric effect

Chart 2.59: Predictive SGC_ Reten_ Relationships of

Descriptors		Flavor/fragrance Molecules	1208
Software :	Dragon	Response (Measured)	GC retention index
#MolDesc:	4885	Glass capillary gas chromatography column	OV-101

Selection of descriptor categories

Phase I ▶ All descriptors
Phase II ▶ 3D-descriptors
Phase III ▶ All but excluding conformational descriptors
Methods ▶ y randamization
▶ Domain analysis
Statistics ▶ LOO
▶ LeaveMoreOut

Model : MLR
Variable selection ▶ Replacement method

Inf.Bits.
Contributing factors to RI
High relevance: solvation connectivity index of first order
Non-influencing: information in 3D-descriptors

Chart 2.60: oligooxyethylenes SGC_ Reten_ Relationships

Descriptors
Topological

Molecules — Oxyethylene derivatives
▶ Alcohols
▶ Thioalcohols
▶ Alkylamines

Chart 2.61: GC_RI indices of heterocyclic compounds

X(MolDesc)

1) Connectivity	**2)** Constitutive **3)** Zero-dimensional	**4)** Galvez Topological Charges indexes
5) GETAWAY		

Response: Boiling point Molecular weight	Model : Resp = fn(Moldes)	
Variable selection	Subset selection	PLSC
Response = fn(moldes)	MLR	PLSR
Validation	CV	

Chart 2.62: SGC_ GCReten_ Relationships for oxygen-containing organic structures

Descriptors
Molecular polarizability effect index
Modified molecular polarizability index
Modified inner molecular polarizability index

Molecules	Diverse oxygen-containing organic structures
Response (Measured)	Kovat's retention index
Stationary phases	OV-1 and SE-54

Model ▶ MLR
Statistic ▶ r > 0.99

Inf.Bits.

- Topological > quantum chemical descriptors
- Descriptors encode molecular size, branching and polar functional groups

Dataset	Compounds	Mobile phase (% v/v AcN)
1	26 Neutral pesticide compounds	40–65
2	Phenoxy acidherbicides	30 and 70

Models

- MLR + GA
  - 7 Variables
- NN-BP
- Validation
  - Leave-m-Out

Chart 2.63: SHPLC_R_IR — SRespR

Descriptors

- E-state
  - I Group
- Connectivity
  - chi and kappa
- Structure-information representation (SIR)

Response (Measured): HPLC-RI

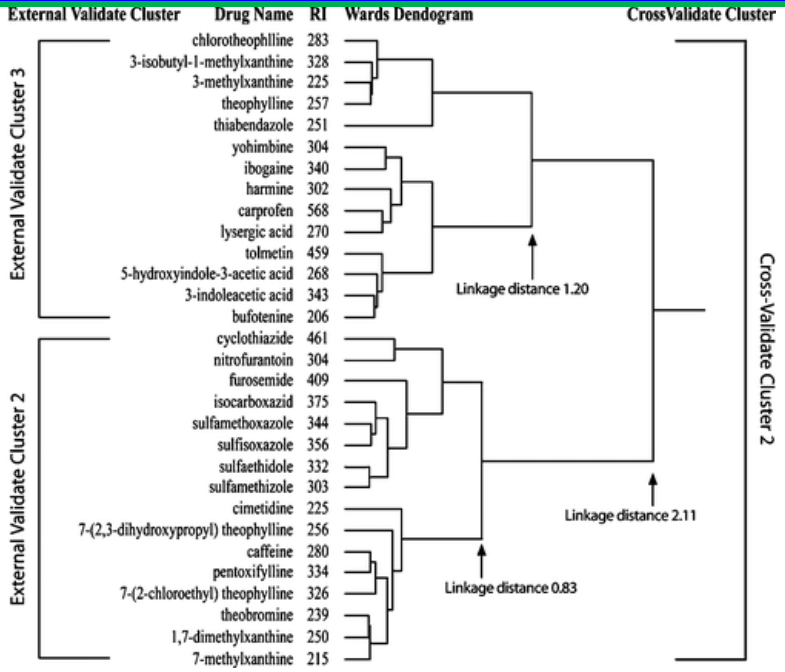
Task: Prediction

Predictive models

- MLP_NN-BP
- Ensemble
  - 10-fold (MLR)
  - 10-fold leave-10% -out

Model statistics

	Training			Validation		External		
	#	r^2	MAE	r^2	MAE	#	r^2	MAE
MLR	498	0.65	83	0.49	73	104	0.49	78
NN		0.93	30	0.84	41			



Courtesy from Ref 81

Chart 2.64: alkylbenzenes

Descriptors
- Topological descriptors
- Geometric
- electronic

- Molecules: 150 alkylbenzenes
- Response (Measured): GC _ Kovats_RI

SRespR_ models
- Descriptor selection: Automated Data Analysis and Pattern Recognition Toolkit (ADAPT); Max (information)
- SRespR Model: MLR; Computational NN

Inf.Bits.

Prediction error of RI Inter laboratory reproducibility

Chart 2.65 : Chlorinated dibenzodioxins

Descriptors
- Topological: First-order molecular connectivity ($^1\chi$)
- structural: Mono-, di- and trichloro substitution patterns

- Molecules: Chlorinated dibenzodioxins 37 tetra- to octachlorinated congeners
- Response (Measured): GC_RI
- Non-polar stationary phase: DB5
- Task: Prediction

SRespR_ models	
Descriptor selection	ADAPT; Max (information)
SRespR Model	Regression
Model Validity	Average (RI_Exp -RI-Cal) : 1.8 RI units
Test dataset	Di- and trichlorinated PCDD congeners

Inf.Bits.
- Molecular size (bulk property) influences RI of PCDD
- $^1\chi$ Index (reflecting molecular size) explains > 97% of the variation in RI
- Patterns of chlorine substitution increases/decreases numerical values of RI

ADAPT: Automated Data Analysis and Pattern Recognition Toolkit

Chart 2.66: Fatty acid methyl esters

Descriptors
- Molecular mass
- Number of double bonds
- # rotatable bonds
- Topological descriptors

- Molecules: 130 Fatty acid methyl esters (Branched, saturated and unsaturated)
- Prediction: 37 Fatty acid methyl esters
- Response (Measured): GC _RI

Models (fixed #descriptors)
- Modified best subset selection
- Forward selection
- Pair-wise correlation
- PLS
- Lasso method

Validation model
- Internal
- External

- Non-polar stationary phase: polydimethylsiloxane

Inf.Bits.
- Prediction error of RI Inter laboratory reproducibility error

Chart 2.67: S GC _ Kovats_RI R

Descriptors
- Topological descriptors

Descriptors' selection
- Combined data splitting-feature selection

- Molecules: 573 Terpenoids
- Response (Measured): GC _ Kovats_RI
- Model: NN

Inf.Bits.
Prediction error of RI Inter laboratory reproducibility

Chart 2.68: Selection of subset of compounds with similar or orthogonal character
- Test molecules
- Cal WHIM descriptors
- Cal auto scaled PCs
- Apply Kennard and Stone (K&S) algorithm
- Cal correlation coefficients based weighted average-linkage (WAL) dendrograms
- Draw color maps
- Evaluate orthogonality of system
- Inference: Subset of 10 compounds have orthogonality

Validation model
- Training → Explainability > 90% Variance in y
- Prediction → Rel. error > 3.2%

⊥ S NMR R (Structure NMR Relationship)

COSY (^{13}C-^{13}C): Beger et al. [75] reported structure NMR information relationships using molecular descriptors (chart 2.68b). This study combined CQC_NMR chemical shift information and atom-to-atom distances of the molecule in SRespR modeling.

Chart 2.68b: 2D-NMR COrrelation Spectroscopy (COSY) ^{13}C-^{13}C | SActR

Descriptors
E-state
Topological ○ Chi

Molecules	30 Steroids
	Binding to corticosteroid binding globulin
Response	Y= (^{13}C-^{13}C COSY)
Descriptors	X = (^{13}C-^{13}C distance spectra)

Dimension reduction and cause-effect models	
Orthogonalization/	PCY = PC(Y)
	PCX =PC(X)
Regression Model	PCY = fn(PCX;par)

Model	Statistics			
	r^2	Leave -m-out		
		q_1^2	q_3^2	q_{10}^2
PCR	0.96	0.92	-----	---
NN_parallel distributed	0.96	-------	0.78	0.73

S PhysChemBiolPar R (Structure Physical_Chemical_Biological-parameter Relationships)

The data from an instrument or scientific experiment if operated in Physical_ Chemical_ Biological_ Marine framework generates free parameters called constants. These are popularly fixed in the mind as chemical constants, biological constants etc. In general, a single (scalar) or n^{th} order tensor (hitherto first order are common) of constants are in reduced dimension compared to the data generated (0 to 4^{th} order or way) for this purpose. Further, these constants are in a specialist's (chemist/ physicist/ Biologist/ Geologist/ marine scientist) space and contain a lot of implicit information. But, it is obscure to non-specialists who may be even experts in other fields. For example, rate constant in chemistry varies with temperature, dielectric constant and/ or chemical structure; but the important information outpoured is how fast the said reaction takes place. Further, they are not like fundamental constants Avogadro number in chemistry or Plank constant in Physics and so on. The variations of such physico-chemical (Henry, log-mobility, biochemical binding) constants are explainable in terms of molecular descriptors cited in table 2.4.

Table 2.4 : Select cases of S PCC R

PCB parameter	#compounds	Descriptors	Model	Software
Henry's constant	63	• 144 descriptors • 53 GA • SAA • 9 TOPO Geo	○ MLR	ADAPT MOPAC
Ion multiple constant	168	• 83 TOPO • 23 Geo • 48 Electronic	○ NN ○ 6-2-1 ○ MLR	HYPERCHEM ADAPT EHNDO (in house) selection of descriptors SA,GA
Ion multiple constant	70 Tr 70 Te	• 79 TOPO • 17 Geo • 5 Electrostatic • 25 CPSA • 126 Total	○ MLR ○ NN	MOPAC
log($1/K_m$)	89 Peptides	• 29 Parameters	○ PLS ○ Q-PLS	
Logk	57 Unsaturated	• QC	○ PLS	
LogP	219	• Topological • Geometric • Information	○ MLR	

The applications of molecular descriptors brought a renaissance in cause-effect relationship modeling. In the classical modeling, experimental (logP, dielectric constant) /derived parameters ρ, σ *etc.* were from 2 to 8. Leaving aside the improvement in modeling technology, the number of X variables increased to nearly 4000. This posed the same old hurdles like correlation in multi-channel responses. The correlation may be in chemical (X) space and/ or Y space. These correlations are again are statistical, chemical or chance based.

Logβ of metal complexes of ionophores: Tetko et al. [58] applied SlogβR model for metal complexes of ionophores with E-state molecular descriptors (chart 2.69). The hurdles in comparing different methods, descriptors in trust worthy comparisons are difficult with small number of datasets.

log(1/K_i): Zheng et al. [153] modeled measured log(1/K_i) values with of 104 VMAT2 ligands using different categories of molecular descriptors (chart 2.70).

S ΔG R: Katrizky et al. [37] proposed structure solvation free energy models for a solute in a series of solvents employing advanced descriptor selection procedure with their CODESSA PRO software (chart 2.71). Boudene et al. [101] reported SNMRR and STherR of a series of methylaluminoxane (MAO) cages of varying repetition units with a follow up good predictivity of ΔG_r (chart 2.72).

Log P: Tetko et al. [76] performed an extensive ensemble NN modeling for nearly 13 thousand logP values of organic molecules to develop ALOGPS (v2.0) software for prediction purposes (chart 2.73). Panek et al. [90] applied Kohonen_NN in unsupervised classification of 88 substituted aromatic compounds in molecular chemical descriptor space followed by cause-effect model of measured properties using Counter-Propagation_NN (chart 2.74).

log*D* uranyl species: Varnek, Katrizky et al. [35] performed an exhaustive study of modeling of uranyl extractants with molecular descriptors, generation of virtual library, prediction of logD, synthesis of best hypothetical compounds and finally testing them for extraction behavior. Two compounds met the criteria endorsing the success of SXR as a preamble before attempting synthesis of lot compounds (chart 2.75) as in irrational brute force approach.

Caco-2 Cell permeability: Randić and Pompe [25] made use of around 900 descriptors from different software packages (chart 2.76) to arrive new molecular entities (NMEs) by screening PBC and BCSs based on Caco-2 Cell permeability and dose Number.

SPermeabilityR: González et al. [23] reported a TOPological sub-structural molecular design to account 90% variance of permeability coefficient for 63 compounds and assessed the efficacy of many types of molecular descriptors (chart 2.77).

- Hansch parameters vs molecular descriptors:

Chiu and So [65] showed that E-state descriptors as well as Hansch (π and MR) parameters yield similar results in SPropR for a variety of compounds. And molecular descriptor sets based upon the approach of Kvasnicka, Sklenak, and Pospichal are capable of substituting F and R parameters. The hydrophobic, steric/polarizability, electronic (field and resonance) characteristics of the substituents are reflected in π, MR, F and R.

Chart 2.69: SComplexationR

Descriptors
E-state ▶ Sub-structural molecular fragments
Counts of atoms determined for E-state atom types

Molecules	Ionophores
Response (Measured)	logK_1 for ML logβ_2 for ML_2 M: Ag^+ ; Eu^3

	Models
MLR	RBF_NN
SVR	Maximal Margin Linear Programming
K_NN	Associative NN

Model_Statistics
- 5-fold external-CV
- Robust statistical tests
  - Bootstrap
  - Kolmogorov-Smirnov
- Wilcoxon signed-rank test

Inf.Bits.

Predictive models
- [SMF descriptors + E-state counts] superior to [E-state indices]
- Average (MLR) with SMF has good predictability
- Results with [SMF] ; E-state counts
- Less significant models with E-state indices

Chart 2.70: SXR model for VMAT2 with molecular descriptors

Descriptors		
WHIM	GETAWAY	Topological
RDF	Geometrical	Aromaticity

Descriptors	Model/Training		Prediction	
	R^2	*RMSD*	Q^2	LOO RMSD
NN model (experimental vs predicted)	0.91	0.225	0.82	0.316
Non-linear NN	0.93	0.382	-----	------

Inf.Bits.
- SXR model reflected critical role of size, steric structure and atomic distribution
- Sheds light on RDD of new chemical moities as ligands of VMAT2
- Directing synthesis of new futuristic molecules

Chart 2.71: S ΔG R — SPropR

Descriptors
CODESSA Pro

Compounds
500

Solvents
69

Model	R^2	SD
2-pyrrolidone	0.837	0.1
dipropyl ether	0.998	0.02

- Descriptor selection
  - Advanced variable procedure
- CV
  - ▶ LOO

Chart 2.72: Predictive modeling of ΔG_r with molecular descriptors

Step: Zero Input: 56 methylaluminoxane cage structures $[AlMeO]_n$, n = 6:12

Courtesy of Ref 101

Step: One Geometry opt with DFT

Step: Two Calculation of molecular descriptors
Geometric, electronic

Step: Three **Response:**
- Internal energy
- Enthalpy
- Gibbs free enthalpy (ΔG_r),

Step: Four Supervised Learning
Identification of opt descriptors

Step: Five Predictive modeling
Input : opt descriptor
Prediction of $\Delta G_{r\ of}$ $[AlMeO]_n$ n= 6 to 33

Chart 2.73: SActR

Descriptors
E-state
○ Atom and bond type
○ #Atoms
• Hydrogen
• Non-hydrogen

Molecules	12908 organic molecules
Daatabase	PHYSPROP Syracuse Research Corporation
Response	log *P* *logarithm(partition coefficient of a solute in n-octanol/water)*
Software	ALOGPS v 2.0

ALOGPS v 2.0

Step: Zero **Molecular structure : A**tom and bond-type Estate indices number of hydrogen and non-hydrogen atoms

Step: One Preliminary selection of indices by MLR 75 input

Step: Two Combination of descriptors
Several atom-type or bond-type indices with similar physicochemical properties

Step: Three Ensemble (50) NNs
- Training
- Each NN contains 10 hidden neurons
- partition algorithm

Step: Four Prediction

Subset	r^2	Leave -one-out	
		RMS	MAE
12, 777	0.95	0.39	0.29

Five Inter series prediction
ALOGPS >> better prediction ability compared to tested methods

If Molecules in the test and in the training subsets same
Then Similar SPropR results

If Molecules in the training and in test sets belong to different series of compounds
Then Less efficient performance of all methods
Reason: Different (diverse set) of molecules for trainingand test

Chart 2.74: Kohonen and counter propagation NNs in

MolDisc modeling of aromatic substituted compounds

Descriptors

WHIM	RDF	**3D-MoRSE**
Atoms in Molecules	Electron Localization Function	**G ETAWAY**

Step: **One** **Input:** 88 nucleophilic and electrophilic substituted aromatic compounds

Step: **Two** **Descriptor calculation**

Step: **Three** **Classification_unsupervised: Kohonen NN**

Step: **Four** Response:
- *n*-octanol/water partition coefficient
- Dipole moment
- Molecular refractivity

Step: **Five** Supervised Learning : Counter-Propagation_NN

Step: **Six** Testing of MolDesc space

3D-MoRSE	Three-Dimensional Molecule Representation of Structures based on Electron Diffraction
GETAWAY	Geometry, Topology and Atom-Weights Assembly
WHIM	Weighted Holistic Invariant Molecular

Chart 2.75: S logD R | SParR

Descriptors
- CODESSA Pro
- TRAIL
- Fragments

D : Distribution coefficient of uranyl extracted between phosphoryl-containing podands from water to 1,2-dichloroethane

Phase I Modeling

Model	Descritpors
1)	Classical structural
	Physicochemical
2)	Fragment

Best Models

a) Sequences of atoms and bonds atoms with their close environment
b) Descriptors (CODESSA PRO) + fragment descriptors (TRAIL)

Phase II Testing

i. Generation of virtual combinatorial library of 2024 podands (CombiLib)

ii. Prediction of log*D*

Inf.Bits

\+ SXR predicted logD values successfully for two compounds

Phase III Synthesis, experimental determination
- Hypothetical compounds were synthesized
- Experimental logD

Chart 2.76: S Cell permeability R | SParR

Descriptors
- CODESSA
- 56 topological indices
- Variable connectivity index $^{1}\chi^{f}$ (not availabe in CODESESS)

Model equation (CODESSA)
BP = 38.12 $^{1}\chi^{f}$ − 37.56

r	RMS °C	Fisher ratio
0.9915	4.21	5691

- **Property:** Boiling point
- **Compounds : 100 alcohols**
- **Task : Prediction**
- **Models: MLR**

Chart 2.77: Permeability of polyethylenes as a function of molecular descriptors

X(MolDes)			permeability coefficient (63,1) = fn(Desc)
Geometric	Constitutional	GETAWAY	Expl: <90% (Compounds :63)
2D-auto correlation	BCUT	3D-Morse	
WHIM	3D-Morse	Topological	
TOPological Sub-Structural Molecular Design (TOPS-MODE)			Expl >77% Compounds: Low-density polyethylene Temp : 21.1 °C

S Process R (Structure Process Relationships)

Jelcic [160] put forward solvent molecular descriptors for elucidating the effect of solvent on the poly(D, L-lactide-*co*-glycolide) (PLGA) nanoparticle size in the emulsification– diffusion method. The experimental data correlated well with information from molecular descriptors (chart 2.78) and this is a reliable probe to have an insight into thermodynamics and the molecular nature of the polymer nanoparticle formation in the solvent environments.

Chart 2.78: Molecular descriptors for solvent effects on sizes of nanoparticle

Model	Response	Fn(MolDesc)
Coarse	Nanoparticle sizes	hydrophilicity index Randic shape descriptor
Rigoruous		GETAWAY Directional WHIM 2D Morse

Sorption: Schüürmann et al. [71] reported E-state descriptors predictor model for sorption of organic molecules into soil (chart 2.79). Andersson et al. [47] (chart 2.80) used a combination of topological, geometric and quantum chemical descriptors in prediction of soil sorption studies.

Chart 2.79: SlogKR SPropR

Descriptors	
E-state	24 Fragments
Molecular weight	Bond connectivity

Response (Measured)	log K_{oc}
Task	Prediction

Models	
MLR	
PLS	Linear
	Non-linear
Continuum Regression	

Data set	571 *Compounds*	r^2	Standard error
Training	80%	0.852	0.469 log units
Test set	20%		

Inf.Bits.

- Model >> linear solvation energy relationships
- Analysis for subsets of H-bond donors/acceptors & non-polar/weakly polar compounds

Chart 2.80: SActR

Descriptors

- DRAGON
- CODESSA
- HyperChem
- logP
- Combination
  - Connectivity indices
  - Geometrical
  - Quantum chemical

- Response : Soil Sorption
- Task : Prediction

Models

- Descriptor selection
- Multivariate(X-y)

S Reaction R

Homolysis of C-X bonds: Hiob and Karelson [33] predicted logk values for gas phase homolysis of C-X bonds with molecular descriptors (chart 2.81). The variation of bond lengths especially of C-C, C-H in different environments drew the attention of molecular groups. Molecular descriptors successfully modelled these precise and accurate numerical values.

Chart 2.81: S lok R SParR

Descriptors
- CODESSA

Response
- **Kinetic parameters (log k)**

C-X bonds
- **287**

- **logk :891**
- **Task : Prection**

SPropR (Structure Property Relationships)

Boiling point, melting point, refractive index and glass transition temperature are some of the widely studied properties of physico-chemical relevance. The explanation and prediction of these properties of a set of congeneric or diverse organic compounds found a niche as SPropR. It encompasses both linear and non-linear trends in descriptors and their transformed space. The advances in cause-effect modelling technology, selection of variables, similarity/diversity measures, residual analysis, best set of models, ensembles, forests and optimization methods

continuously uplift the quality of these chemoinformatic models. The soft from of the table is available in WINWORD, Excel and database formats. It can be readily sorted chronologically compound/ response/ modelling techniques wise. The link field (references) common to the literature enables picking up the contributions of an author(s) and references form a single or a group of journals specific period of time etc. The static knowledge base contains the up to date state of art of each of the sub goals in descriptor research like geometry of optimization, hard modelling, soft models, residual statistics and packages to calculate different types of descriptors, typical data sources.

Aqueous solubility: Catana et al. [70] tested continuum regression and linear andnon-linear PLS models for a large set of aqueous solubility of a large number of compounds (chart 2.82). Tetko et al. [59] proposed SLP_(33-4-1)_NN for aqueous solubility of around 1200 molecules with electro-topological E-state indices (chart 2.83). Abramov [86] obtained thermodynamic intrinsic aqueous solubility of compounds by indirect measurements and molecular descriptors. The accuracy and predictive power depend upon solid state contribution viz. solid state, ΔG_{fus} and nonsolid state, $\Delta G_{mix,}$. The results with SpropR and descriptor based models show the need for molecular descriptors accounting for long range periodic effect in the crystalline state. Liu et al. [34] predicted C60 solubility in different solvents with CODESSA descriptors as explanatory variables in SVR (chart 2.84). Pham-The et al. [55] made use of around 900 descriptors from different software packages (chart 2.85) to arrive new molecular entities (NMEs) by screening PBC and BCSs based on Caco-2 Cell permeability and dose Number.

Melting point: Varnek et al. [66] developed predictive structure melting point relationships with descriptors from DRAGON software with good success.

Boiling point: Lima Ribeiro et al. [146] reported physico-chemical properties modeling of PAHs in molecular descriptor space by soft regression techniques (chart 2.86).

Glass transition temperatures : Kim et al. [145] employed MLR and NN models to model and predict glass transition temperatures of electroluminescent molecules with different kinds of calculated descriptors based on graph theory, thermodynamics and quantum-chemical-electrostatics (chart 2.87). Mattioni and Jurs [98] reported 10-descriptor computational _NN SPropR models for glass transition temperature of monomer and polymers with good predictive statistical measures (chart 2.88)

Viscosity of ionic liquids: Mirkhani and Gharagheizi [89] arrived at SPropR model for ionic liquids using molecular descriptors as causative factors for variation of viscosity.

Chart 2.82: SSolubilityR SPropR

Descriptors	MOE package (22 Desc)
E-state	65 ISIS keys
1D and 2D descriptors	

Response (Measured)	Aqueous Solubility
Task	Prediction

Models	
NNs	
PLS	Linear
	Non-linear
Continuum Regression	

Data set	*Compounds*	r^2	RMSE
Training	930	0.935	0.468
Test set	177	0.911	0.475

Inf.Bits.

Predictive Aqueous solubility models

Linear methods [CR = PLSR) ≅ Non-linear models

Chart 2.83: SSolubilityR SPropR

Descriptors_Model1
- electrotopological E-state

Molecules — 1291 organic compounds

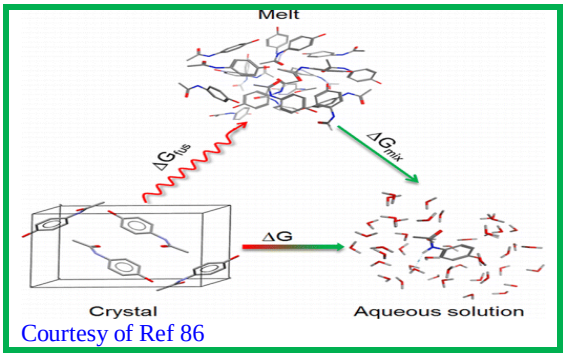
Response (Measured) — Aqueous Solubility

Descriptors-Model2
- Electro-topological E-state ▶ Molecular connectivity
- Shape ▶ Flexibility and
- Indicator indices

Model_Statistics
- 5-fold external-CV
- Robust statistical tests
  - Bootstrap
  - Kolmogorov-Smirnov
- Wilcoxon signed-rank test

Inf.Bits.
Predictive models
- Model2 has 12 hidden neurons
- Model1 >> Model2

	Model1
NN	33-4-1
r^2	0.91
RMS	0.62



Courtesy of Ref 86

Chart 2.84: S Solu R SPropR

Descriptors
- CODESSA

Compounds
- C_{60}

Solvents
- Diverse categories

Model	R^2	RMS
Lin	0.892	0.126
NonLin	0.903	0.116

Logk
- 891
- Task : Prediction
- Model : SVR

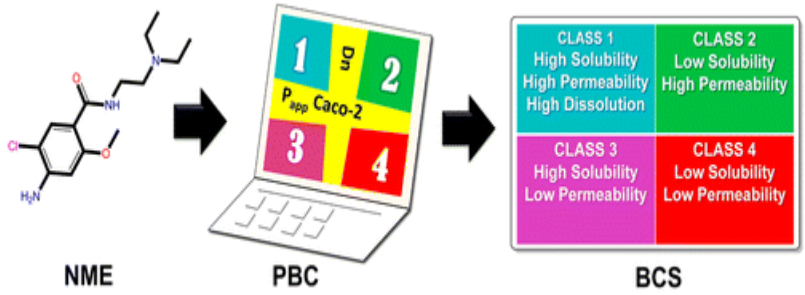
Inf.Bits

+ SVR >> Heuristi Regression

Chart 2.85: S SolubilityR, S Permeability R SPropR

Descriptors Packages
- Dragon (V6.0)
- 803 0-2D
- Volsurf+ (version 1.0.4_)
- 126 Desc

- **Response:**
  - **PBC solubility**
  - **Permeability**
- **Task : Classification**
- **Models: 18 Linear, Non-Lin**



Courtesy of Ref (dragon-12)

Chart 2.86: Predictive modeling in molecular descriptor space for properties of PAHs SPropR

X(MolDesc)

Connectivity Volume	Thermodynamic	Topological	HOMO,LUMO energies	Surface area
Molecular hardness	Polarizability	Atomic charges	Electrnoic	Steric

Response

- **Boiling point**
- **Octanol-water partition coefficient**
- **Retention time index**

Property	PLR	PLSR
BP	0.995	0.995
log Kow	0.975	0.976
RI	0.898	0.898

Chart 2.87: Molecular descriptor modeling of glass transition temperature

Descriptors
- 1 Topological
- 1 spatial
- 3 electrostatic
- 1 thermodynamic
- 1 structural

Molecules	103 organic electrolumi-nescent devices molecules
Response	glass transition temperature
#Descriptors	83

Model statistics

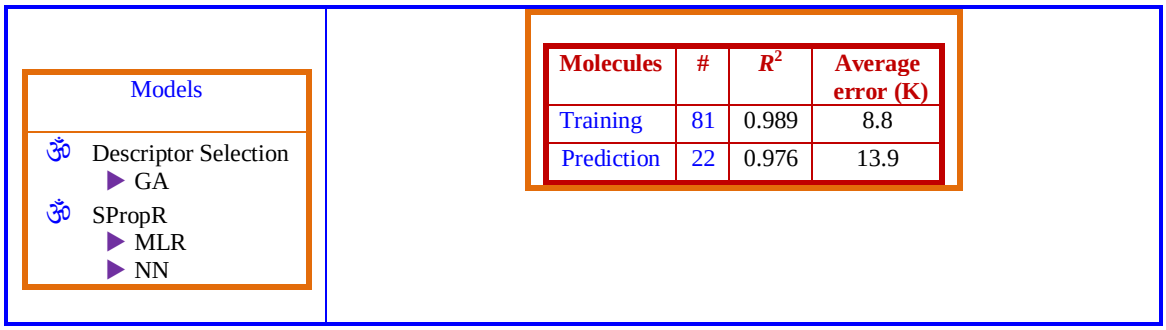


Chart 2.88: Glass transition temperature prediction with molecular descriptors using NNs

Descriptors	
Electronic	Topological
Geometric	

Models	Desc selection
MLR	SAA
Computational NN	GA

Monomer				
	RMS	R2	Compounds	165
Training	10.1	0.98	Response	Glass transition Temperate
Testing	21.7	0.92	Opt_descrip	10
			Committee	10 CNNs

Polymers				
	RMS	R2	Compounds	251
Training	21.1	0.96	Response	Glass transition Temperate
Testing	21.9	0.96	Opt_descrip	11
			Committee	10 CNNs

Table 2.5 describes a few typical investigations of molecular properties employing descriptors from a variety of packages.

Table 2.5: Typical studies in S Property R

Property	#	Molecules	Descriptors	Model	Software
	150	Alkanes (C_1-C_{10})		MLR	MDE Molecular distance edge vector
BP	299 572	Tetrahydrofuran Thiophene Furan Pyran Pyridine	TOPO QC CPSA Geo	MLR NN (20:10:1)	ADAPT
BP	185	Ethers, peroxides acetals, sulphur analogues	11 TOPO	MLR RBF	
BP	1020		GEO TOPO Information	MLR	Poll SYBYL .2 CONCORD 3.2.1
BP	9	Isomers of heptane	4 Orthogonal		
BP	532	Halo-poly halo alkanes	TOPO	MLR	

BP	134 209	⊥ Thiophenes ⊥ Furans/tetrahydrofurans	• 77 • TOPO • Electrostatic	O MLR	❖ ADOPT
BP	296	⊥ Hydrocarbons	• 40 total descriptors	O MLR O NN	❖ GA ❖ (selection of)
BP	185 22 44	⊥ Acyclic compounds ⊥ Ethers ⊥ Sulphides	•	O MLR	
BP		⊥ Halo polyhalo alkanes ⊥ C_1-C_{11}	• 12 descriptors	O MLR	
BP	52	⊥ Chloroethanes	• Descriptors	O MLR	
BP	155	⊥ Acyclic ethers	• TOPO • E-state • Balaban	O MLR	
BP		⊥ Haloalkanes (C1-C4) ⊥ C1 : 15 ⊥ C1-C2 : 62 ⊥ C1-C4 : 276	• TOPO • Information	O NN	
BP CT	298 Tr 154 Te	⊥ ⊥ Industrially important	• QC • TOPO	O NN MLR	❖ ADAPT ❖ MOPAC
BP Critical temperature	298	⊥ Saturated/unsaturated	• 8 descriptors	O MLR O NN	
BP Critical temperature	227 100 327	⊥ Non-nitrogen ⊥ nitrogen	• Descr selection • SA GA	O MLR O NN O 8-3-1	
BP	298 19	⊥ O,N,Cl,Br ⊥ Pred Compound	• 2 Geometric	O MLR(8)	❖ CODESSA
BP MP CP MV	150	⊥ Alkanes (C_1-C_{10})	• 8 TOPO (substructure)	O NN 8-7-1	
BP MP Crit temp MV	150	⊥ Alkanes (C_1-C_{10})	•	O NN O MLR	
Crit temp Crit PRES	165	⊥ Industrially important ⊥ organic compounds	•	O MLR	❖ MOPAC ❖ ADOPT
Critical temperature	76 165	⊥ Hydrocarbon	• TOPO • QC	O PCA O MLR	❖ CODESSA ❖ MOPAC 6.0
MP	443	⊥ Substituted benzenes	• Information • QC • Electrostatic • Geometric	O PCA O MLR	❖ CODESSA ❖ MOPAC
MP		⊥ Pyridinium ⊥ Bromide	• Information • QC • Thermodynamic	O MLR	❖ CODESSA ❖ MOPAC
Octane		⊥ Heptanes	• TOPO	O PLSR	

number			• Sub graph • Count		
Refractive index	125	⊥	• QC • Geometric	○ MLR	❖ CODESSA
Refractive index		⊥ Amorphous ⊥ Homopolymers	• QC	○ MLR (5)	❖ 655 CODESSA
Solubility	140	⊥ Aromatic ⊥ Aliphatic	• 144 Descriptors • 53 GA, SAA • 9 TOPO • GEO • Electrostatic	○ NN	❖ ADAPT ❖ MOPAC(PM)
Solubility	109 132 241	⊥ Hydrocarbons ⊥ Halogens	• Topological • Electrostatic • QC • Information	○ MLR	❖ CODESSA ❖ MOPAC 6.0
Solubility	150 54	⊥ Alkanes C_1-C_{10} ⊥ Alcohols & Hydrocarbons	• Cavity surface area	○ MLR	
Solubility	332	⊥ Compounds	• TOPO • GEO • Electrostatic • QC	○ MLR CNN	❖ ADAPT ❖ HyperCHEM ❖ MOPAC
Solubility	300	⊥	• TOPO GEO • Electrostatic	○ MLR	
Solubility in super critical CO_2	52 Tr 6 Te	⊥	• QC • GEO	○ NN ○ 7-2-1	❖ ADAPT ❖ HyperCHEM ❖ MOPAC
Solubility of gases	400	⊥ Compounds	• QC • GEO	○ MLR PCR	❖ CODESSA ❖ MOPAC
Solvent polarity	25 67	⊥	• Information • Geo	○ MLR PCR	❖ CODESSA ❖ MOPAC
Tg	88	⊥ Polymers	• QC • GEO	○ MLR(5)	❖ CODESSA
VP Solubility logP	411	⊥	• QC	○ PC	❖ CODESSA ❖ MOPAC 6.0
VP	479	⊥ PAH ⊥ Acids ⊥ Esters	• 25 • Constitutional • Geo • QC • TOPO	○ NN>MLR	❖ CACHE ❖ MOPAC (MP3)
VP	342 134	⊥ Tr ⊥ Te	• 92 TOPO	○ MLR	❖ POLLY 2.3 ❖ SYBYL 6.2 ❖ CONCORD 3.2.1

🕮 Miscellaneous:

Cyclic compounds: Hu et al. [67] [extended SPropR models to cyclic compounds with a hybrid method of projection pursuit and number theory based algorithms (chart 2.89). It generates uniformly distributed directions on unit sphere helping interpretation. The information hidden over the spaces are spanned in connectivity, Kappa and Atom-type E-State Indices.

Diversity of compound collection: Ma et al. [85] increased of an existing compound collection through BCUT chemistry space and prioritization algorithm. Further, it was probed into correlation of distance of chemistry-space and Tanimoto similarity index in weighted linear regression (Fig. 2.2).

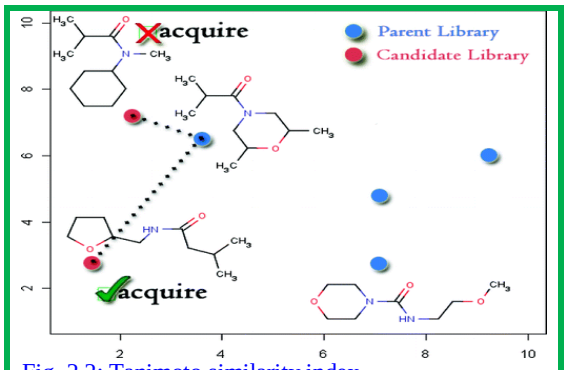


Fig. 2.2: Tanimoto similarity index
Courtesy of Ref 85

Isoxazoline and oxime derivatives: Wang et al. [91] used GA in selecting molecular descriptors and MLR for an optimum model for 33 isoxazoline and oxime derivatives of podophyllotoxin (chart 2.90)

Pesticides: Che et al. [93] reported SpesticidalActR of N-Arylsulfonyl-3-acylindole Arylcarbonyl Hydrazone derivatives against nematicidal activity with a variety of molecular descriptors (chart 2.91).

Uptake capacities of methane: Fernandez et al. [94] studied ~58,000 hypothetical Metal–Organic Frameworks structures for uptake capacities of methane, N_2, and CO_2.

Test Data : ~25,000 MOFs	
R^2	0.70 to 0.82

Scavenging activity: Argüello-García et al. [143] reported hard and soft regression models for SActR of scavenging behavior of thioallyl compounds with energy and surface area descriptors (chart 2.92).

Pereira et al. [46] modeled Parr electrophilicity index with supervised models based on different philosophies using molecular descriptors as causative factors (chart 2.93)

Chart 2.89: SPropR

Descriptors
- Molecular connectivity index
- Atom-type E-State index
- Kappa shape index
- chi cluster and path/cluster indices

Model: projection pursuit + number theory method

Inf.Bits.
- ☏ Important factor
  - 🔔 Number of quaternary atoms
- – Collinearity in topological descriptors is a serious problem

Descriptor	Encodes
Chi cluster Path/cluster	🕮 Cyclicity
Kappa shape	🕮 Significant information on size 🕮 Degree of cyclicity 🕮 Degree of centralization/ separation in branching
Atom-type E-State	🕮 Size 🕮 Branch number

	Cyclicity

Chart 2.90: GA-MLR model for isoxazolines in chemical descriptor space

Category	Descriptor
2D autocorrelation	GATS4e
Edge adjacency indices	EEig06x
RDF	RDF080v
3D MoRSE	Mor09v
Atom-centered fragment	H-052

Chart 2.91: Nematicidal activity with WHIM and GETAWAY descriptors

Category	Descriptor
WHIM	E1m
GETAWAY	R1m+
	R3m+
Burden eigenvalues	BEHm8
Edge-adjacency	EEig05x
	EEig13d

Molecules	54: l N-arylsulfonyl-3-acylindole arylcarbonyl hydrazone derivatives
Response (Measured)	Nematicidal activity (LC_{50})

Training		
Correlation coefficient	0.791	R^2
Leave-one-out cross-validation CC	0.701	Q2LOO
7-Fold cross-validation CC	0.715	Q27-fold
External validation		
Cross-validation CC	0.774	Q2ext
Root-mean-square error	3.412	RMSEtest

Inf.Bits

Nematicidal activity correlated with

- Molecular mass (r < 0)
- Molecular polarity (r > 0)

Chart 2.92: Aliphatic, heterocyclic amines

Descriptors

- Topological descriptor
- Polar surface area
- Electronic
- Total energy

Molecules	Thioallyl Compounds from Garlic
Response	Scavenging activity
Method	Catalase protection assay

properties — HOMO energy

Descriptor: Equilibrium electronegativity
- Basis: SEMO ; Topological concept
- Cal Group electronegativities (A) taking into consideration of chemical environment
- Cal relative bond lengths(B), equilibrium electronegativity (C)
- SEMO_TOPOL_Desc = fn(A,B,C)

Models
- MLR
- PLSR

Inf.Bits.
- Deeper knolwdge of antioxidant mechanism of garlic derivatives involving HOCl production

Chart 2.93: Structure Mayr electrophilicity relationship with Dragon descriptors

Descriptor Packages
- DRAGON
- Adriana.Code
- CDK

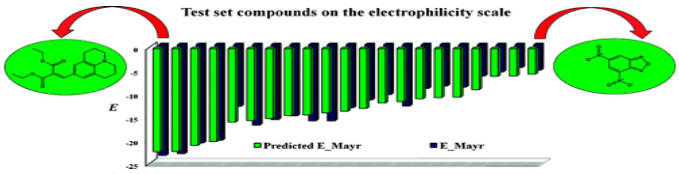
Models
- *k* nearest neighbors
- Model trees
- Random forests
- SVMs
- Associative NN
- Counterpropagation NN
- MLR

Descriptors in best model (SVM)
- Parr electrophilicity index
- ε_{LUMO}
- Hardness
- CDK descriptors
  - FNSA-3
  - ATSc5
  - Kier2 and
  - nAtomLAC

Model characterstics
- $R^2 = 0.92$
- *#Compounds : 21*

Correlations between Mayr electrophilicity and descriptors (r_D)

Descriptor	r_D
ε_{LUMO}	−0.82
Folding degree index	−0.80
Kier flexibility index	−0.78
Kier S2K index	−0.78



Courtesy of Ref 46

In binary classification of compounds using molecular descriptors of properties, naive Bayes classifier is suboptimal with a possible large bias (Fig. 2.3). ROC plot is instrumental in consolidating the descriptors with true positive and false positive characteristics for promiscuous compounds (Fig. 2.4).

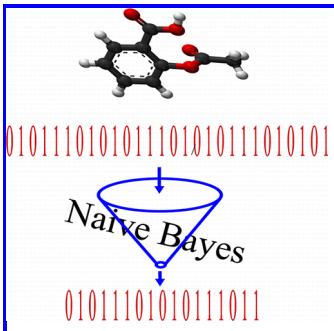


Fig. 2.3: Sub-optimal binary classification based descriptors

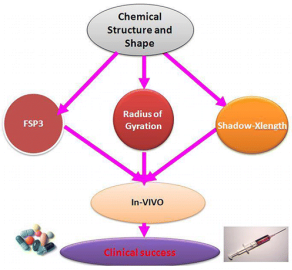


Fig. 2.4: Role of ROC in classification of compounds with true and false positive discrimination

03. MethodBase_MolDes (MB. MolDes)

Constitutional descriptors

Constitutional descriptors refer to atom/bond/ring/groups/substituent counts in a molecule along with bulk quantities viz. molecular weight, average molecular weight (Chart 3.1). In a nut shell, they reflect the compositional details of a molecule. The total number of moieties, the ratio or fraction is more informative compared to basic data of presence/absence of a group.With advanced spectroscopic/ quantum chemical/ reaction dynamics techniques, the numerical information from constitutional category is simple/ trivial and assimilated as preliminary from purists' chemistry point of view. However, they have a niche in molecular descriptor research of macromolecules/drugs/ materials so on.

Chart 3.1a: Constitutional descriptors

Number_of_atoms [1:108]	Rings total
Molecular size	RingsSize [3:12]
Molecular weight	Rings_6-mebered_WithCarbonAtoms
+ Independent of conformer	Circuit structure index
Normalized constitutional descriptors	Average of molecular weight $AveMolWt = \frac{MolWt}{N_{atom}}$
Ratio of molecular ID to number of atoms in molecule Count of paths to number of atoms Connectivity index to number of atoms in hydrogen reduced skeleton	Mean value of atomic composition index

Descriptor	Constitutional
Input	1D-strucutre
Application	Similarity searching and clustering
Significance	Molecular composition

Descriptor	Fragment
Input	1D-strucutre
Significance	Gross composition of molecule
Retains	Chemical information
Does not retain	Neighboring fragment influence
Remedy	Environmental descriptor
Limitation	Does not indicate connection

Count descriptors	
#$$$	
Bonds	Total non-Hydrogen Single Multiple : [Double triple] Rotatable 3.aromatic Rotatable 3.fraction Order (H-suppressed/depleted)

between individual atoms

Hybridization (1:4)
[primary, secondary, Tertiary, quaternary]
[SP3, SP2, SP]

- Including implicit hydrogens
- Between heavy atoms
- H-bond donors
- H-bond acceptors

Chart 3.1b: Functional Groups

		ammonium	phenols	phosphites
phosphates	phosphothionates	phosphodithionates	Phosphothioates	
aldehydes	amidine	anhydrides	hydroxyls	Azetidines
Aziridines	Beta-Lactams	carbonate	carboxylic acids	CH2RX
CHR2X	CHRX2	CR2X2	CR3X	CRX3
cyanates	disulfides	dithioacids	dithioesters	esters
ethers	Furanes	guanidine	hydrazones	hydroxyl
hydroxylamines	hypohalogenydes	Imidazoles	imides	imines
isocyanates	Isothiazoles	isothiocyanates	Isoxazoles	ketones
N azo-	hydrazines	nitriles	nitro groups	nitroso
N-nitroso	Oxazoles	oximes	Oxiranes	Oxolanes
Oxetanes	oximes	phosphanes	phosphates/ thiophosphates	phosphites/ thiophosphites
phosphonates/ thiophosphonates	phosphoranes/ thiophosphoranes	positively charged N	primary alcohols	primary amides
primary amides	primary amines	primary amines	Pyrazines	Pyrazoles
Pyridazines	Pyridines	Pyrimidines	Pyrroles	Pyrrolidines
quaternary N	R#CX	R=CHX	R=CRX	R=CX2
secondary alcohols	secondary amides	secondary amines	sulfates	Sulfenic (thio)
thio-/dithio-	sulfones	sulfonamides	sulfonates	sulfoxides
sulfuric	tertiary alcohols	tertiary amides	tertiary amides	tertiary amines
Thiazoles	thioacids	thiocyanates	thioesters	Thioethanes
thioketones	thiols	Thiophenes	Thiranes	Thiophenes
Triazoles	urea	water	X on aromatic ring	X on exo-conjugated C
(thio-) carbamates	124-Triazines	135-Triazines	acyl halogenides	
CH2X	CR2HX	CR3X	R=CHX	R=CRX
R#CX	CHRX2	CR2X2	R=CX2	RCX3
X-C on aromatic ring	X-C- on ring	X-C= on ring	X-C on conjugated C	donor atoms for H-bonds (with N and O)
acceptor atoms for H-bonds (N O F)	CH3R / CH4	CH2R2	CHR3	CR4
CH3X	CH2RX	CH2X2	CHR2X	CHRX2
CHX3	CR3X	CR2X2	CRX3	CX4
=CH2	=CHR	=CR2	=CHX	=CRX
=CX2	#CH	#CR / R=C=R	#CX	R--CH--R
R--CR--R	R--CX--R	R--CH--X	R--CR--X	R--CX--X
X--CH--X	X--CR--X	X--CX--X	X--CH..X	X--CR..X
X--CX..X	Al-CH=X	Ar-CH=X	Al-C(=X)-Al	Ar-C(=X)-R
R-C(=X)-X / R-C#X / X-=C=X	X-C(=X)-X	X—CH..X	X—CR..X	X—CX..X

ॐ Topological descriptors

Structure of a molecule is not a set of physical properties. But one of the ways of representation is by a molecular graph taking into account of atom-atom connectivity based on molecular orbital (MO) and graph theoretical principles. A notable consequence is it mirrors some of arisingproperties.A graph of a molecule is ID picture of the adjacency/connectivity of atoms. The search of sub graph in a graph is analogous to substructure identification in a structure of a given organic compound. Typical and popular types of sub graphs are path (P), cluster (C) and path-cluster (PC) and chain of different orders (Chart 3.2).

Chart 3.2: Principals of graph [128,129]

- Let $G(k,n)$ be the set of connected simple n-vertex graphs $V(G) = v_1, v_2, \ldots v_n$ with edge set $E(G) = e_1, e_2, \ldots e_m$
- Among all n-vertex graphs with edge or vertex connectivity k, graph $G = K_k \vee K_1 + K_{n-k-1}$, the join of K_k, complete graph on k vertices, with disjoint union of K_1 *and* K_{n-k-1}, is unique graph with maximum sum of squares of vertex degrees

Topological descriptors (Wiener to Kier and Hall connectivity indices) are global ones, as the entire molecule was searched for a fragment (path, cluster, path cluster of different orders). They could successfully model many physic-chemical properties depending upon the characteristics of the entire molecule. However, the limitation of these descriptors is that they do not explain phenomenon depending upon an atom or a group like nucleophilic attack, biological response (BR)of an atom at specified site. This led to the need of local characteristics like those depending on atomsof a molecule. Again Kier and Hall are the forerunners in this pursuit by proposing E-state index.

Mathematical perspective of chemical structure

Quantum mechanics had influence in progress of chemical bonding, apart from galaxy of chemical reactions researched. Even then, traditional chemists were looking into molecules in terms of bonding in two dimensional planar picture of chemical structure. The representation of bonded and non-bonded atoms in molecule in numerical second order tensor form started with adjacency/connectivity matrix.

Adjacency matrix

Adjacency matrix or connectivity is a numerical representation of how the atoms are connected in the molecule in 2D framework. It is a square symmetric matrix, where each entry is one if connected or zero if not connected to the other atoms. The row or column sum in an adjacency matrix represents connectivity number of that with the neighboring ones. Jurs and Katrizky proposed normalized descriptors.

Descriptor	2D
Input :	Atom and connection information (elements, bonds, formal charge)
Output:	Numerical values
Invariant to	Conformations of molecule
Don't require	3D-Atomic coordinates, conformers
Z	Atomic number Lone pair pseudoatoms LP (Z=0)

Descriptor	Adjacency matrix
Input	2-D structure
Basis	Undirected graph theoretical representation of hydrogen implied (formerly called suppressed) 2-D structure of a molecule
Algorithm	AdjMat (i,j) = 1 if i and j are neighbours = 0 otherwise
Application	Distance matrix

Heavy atoms	Z>1
Trivial atom:	LP or H with one heavy neighbour
Hydrogen count (h):	No. of hydrogen attached
Heavy degree d	No. of heavy atoms to which it is bonded

Descriptor	Extended adjacency matrix	Topo
Formulae	$a_{ij}^{E} = \begin{cases} a_{ij}^{\frac{\frac{\delta_i}{\delta_j}+\frac{\delta_j}{\delta_i}}{2}} & if\ i \neq j \\ 0 & if\ i = j \end{cases}$	a_{ij}^{E} Elements of adjacency matrix δ Vertex degree ▶ Number of lines meeting at each vertex $e\ln_i$ Atomic electronegativity π^* Conventional Bond order
	If hetero atoms $a_{ij}^{H} = \begin{cases} a_{ij}^{\frac{\frac{\delta_i^h}{\delta_j^h}+\frac{\delta_j^h}{\delta_i^h}}{2}} & if\ i \neq j \\ e\ln_i & if\ i = j \end{cases}$	$\delta_i^h = \delta_i\ ^*\ln_i$
	If bond multiplicity $a_{ij}^{HB} = \begin{cases} a_{ij}^{\frac{\frac{\delta_i^{hb}}{\delta_j^{hb}}+\frac{\delta_j^{hb}}{\delta_i^{hb}}}{2}} & if\ i \neq j \\ a_{ii}^{hb} = a_{ii}^{h} & if\ i = j \end{cases}$	$\delta_i^{hb} = \delta_i^{h} + \left(1 - \frac{1}{\pi^*}\right)$ $\begin{cases} 0.00 & if\ \pi^* = 1 \\ 0.33 & if\ \pi^* = 1.5 \\ 0.50 & if\ \pi^* = 2 \\ 0.67 & if\ \pi^* = 3 \end{cases}$

Distance matrix

Distance matrix (DistMat) or path of m-bonds is m^{th} power of adjacency matrix. It is also square symmetric second order tensor like adjacency matrix. Each element ($DistMat_{ij}$) corresponds the distance of i^{th} to j^{th} atom. The chronological developments in distance matrix amply demonstrate different frames of thought in conceiving the neighborhood.

Characteristic	Distance matrix
Input	H-suppressed 2D-connectivity structure
Basis	Adjacency matrix
Definition	▶ Number of edges on the shortest path from node i to node j. ▶ Each element is equal to the number of bonds on the shortest path form atom i to atom j
Application	▶ Computer search for functional groups

Characteristic	Adjacency matrix or connectivity table
Input	Hydrogen atom implied (formerly called suppressed) 2D-structure of a molecule
Basis	Undirected graph theory
Algorithm	AdjMat(i,j) =1 if i and

▶ Partitioning of atoms of a molecule into equivalent classes ▶ Sub structure searching	j atoms are connected/bonded = 0 else
	Application: Distance matrix

Algorithm: Powers of adjacency matrix

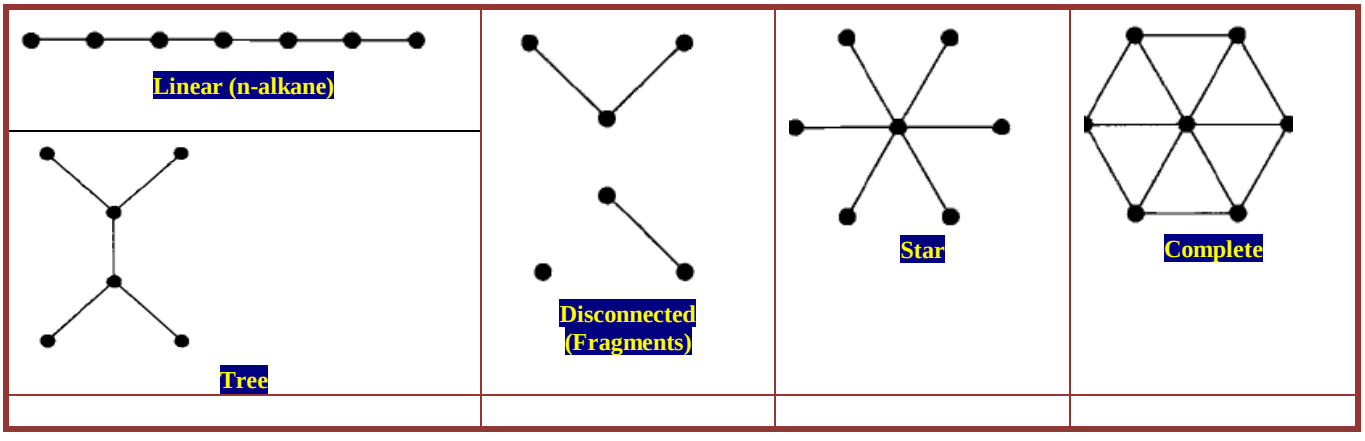
Formula		Order	Indicates
DistMat(1)	AdjMat	1	Path of one bond
DistMat(2)	$AdjMat^2$	2	Path of two bonds
DistMat(3)	$AdjMat^3$	3	Path of three bonds
DistMat(4)	$AdjMat^4$	4	Path of four bond

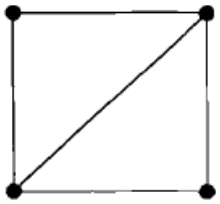
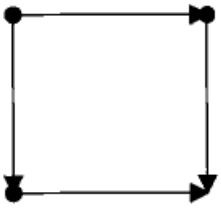
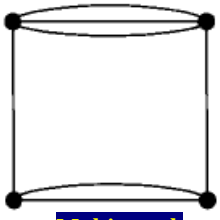
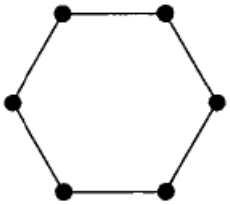
Sum of topological distances between $atom_i$...$atom_j$

$atom_i$...$atom_j$				
N..N	O..O	S..S	P..P	F..F
N..O	O..S	S..P	P..F	F..Cl
N..S	O..P	S..F	P..Cl	F..Br
N..P	O..F	S..Cl	P..Br	F..I
N..F	O..Cl	S..Br	P..I	Cl..Cl
N..Cl	O..Br	S..I		Cl..Br
N..Br	O..I			Cl..I
N..I				Br..Br
				Br..I
				I..I

Graph theoretical research in adjacency and distance matrices

Mathematical graph describes abstract vertices joined by edges. Each molecular graph is represented by a matrix, a polynomial, a sequence of numbers or a numerical index (called topological index).The order of graph equals to the number of vertices in the graph. In a multi graph, two vertices may be joined by more than one edge (Fig. 3.1).Estrada[116]proposed a generalized graph-theoretical matrix; the specific cases of it being adjacency and distance matrices.The vector–matrix–vector multiplication procedure and generalized invariant concept paved way to generalization of classical topological indices. In a connected graph, the chemical structure has at least two connectivity points. The number of chemicals exceed 10^{100}; chemists' representation molecules ranges from 0 to 6D, while a spectroscopist/ molecular descriptor scientist uses 100-300 dimensions in Eigen frame or spectral domain.



 **Graph**	 **Digraph**	 **Multi-graph**	 **Regular graph (Hexnane)**

#vertices : 2

$2K_1$A? K_2A_

#vertices : 3

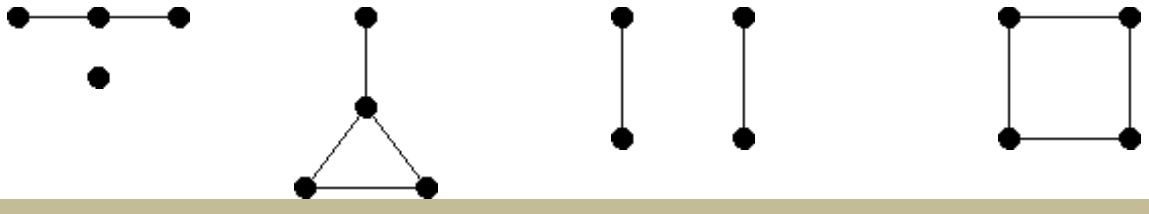
$3K_1$ = co-triangle**B** triangle = K_3 = C_3Bw P_3BO P_3Bg

Graphs are ordered by increasing number of edges in the left column

#vertices : 4

$4K_1$ = $K_{4C?}$ K_4 = $W_{3C\sim}$ co-diamond$_{CC}$ diamond = K_4 - e = 2-fan$_{Cz}$

co-paw$_{CE}$ paw = 3-pan$_{Cx}$ $2K_2$ = C_{4CK} C_4 = $K_{2,2C}$



claw = $K_{1,3C}$ co-claw$_{CJ}$ P_{4Ch}

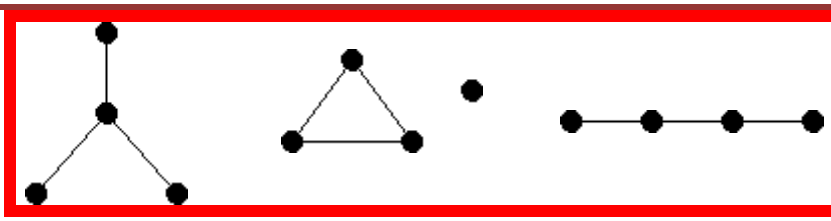


Fig. 3.1Typical types of graphs

Generalized graph-theoretical index and infinite number of derivable classical topological descriptors Topological

Descriptors

Generalized graph-theoretical matrix

$\xrightarrow[Specific\ cases]{}$ [*Adjacency matrix*; *Distance matrix*]

Two new molecular graph-theoretical vectors

Generalized graph-theoretical index

+ Infinite set of descriptors are derivable from it
+ Invariant
+ Classical_TIs are generalized

Generalized graph-theoretical index

Vector – matrix – vector multiplication procedure + Generalized invariant concept

↓

[*Generalized Classical TI*]

$$[Classical\_TIs] = \begin{bmatrix} Wiener\ index \\ Zagreb\ M_1\ and\ M_2\ indices \\ Balaban\ J\ index \\ Harary\ H\ number \\ Randic\ \chi\ index \\ valence\ connectivity\ index \end{bmatrix}\ as\ specific\ cases$$

Wiener index

Represented in 8-D space of the generalized topological indices in radial graphic form

Boiling point of octane isomers

Index	CC
Modified Wiener	0.944
Original Wiener	0.5392

Inf.Bits.

Modified_Wiener used for of C_l-C_g alkanes

Best model with one descriptor

Polarity number

Polarity number is half of the count of all paths of length three. It is directly obtained from distance matrix. Recently maximum/sum/absolute of sum of Eigen values of distance matrix weighted with a variety of properties (electro-negativity, mass, polarizability etc.) are proposed.

Descriptor	**Polarity number**
Input	H-suppressed 2D-connectivity structure
Definition	Wiener polarity index of a graph G is number of unordered pairs of vertices such that the distance between u and v is 3 i.e. Count of all parts of length three
Formula	$PolNum = 0.5^* \sum_{i=1}^{Natom} \sum_{j=1}^{Natom} distMat_{i,j}$
	+ good discriminatory power

```
%
%  PolNum.mR S Rao (8-6-15)
%
[PolNum] = om_polnum(AdjMat)

DistMat = om_distmat(AdjMat)
PolNum = 0.5 *sum(sum(DistMat))
```

- Bound for the Wiener polarity index of unicyclic chemical graphs calculated
- Maximum Wiener polarity index of unicyclic graphs determined

Wiener index

The popular Wiener index proposed in 1947 is the column or row sum of DistMat. In other words, it is equal to sum of distances between all pairs of vertices of H-depleted molecular graph of an organic compound.It is the start of a new era of explanatory variables and the saga continues even in 2015. Winer index and its clones continue to be principal descriptor set that correlating a chemical compound's molecular structure with experimentally measured data. It is based on connectivity (adjacency) matrix for alkanes to explain physico-chemical properties like boiling point (BP), change in heat (ΔH) etc. Further, it widely used in sub-structure search in silicon virtual library of compounds.

Descriptor	Wiener Index
Input	2-D Structure
Def	Sum of distances between all pairs of vertices of a connected graph. Or Sum of all topological distances in H-depleted molecular graph
Nomenclature	Path number in beginning
Basis	Trees with identical number of points have equal number of paths
Output	Increases with molecular size Decreases with branching or compactness
Formula	Wiener Ind = 0.5 * sum(sum(D))
Encodes	Cube of Van der waal's attraction between two parts of a molecule Measure of mean external contact area of molecule (3) Branchedness Of a molecule acyclic compounds
Aplicable to	Ayclic alkanes
Application	Chromatographic retention times-correlation extremely well monoalkyl and o-dialkylbenzenes

Wiener index

\+ Accounting for variation of physico-chemical properties (like boiling point) of organic compounds (alkanes)

\- Does not encode cis-/trans- isomeric or chiral phase information

\- Wiener number gives lower weights to peripheral bonds

\- Higher weights to internal C-C bonds

Remedy :

Reverse Wiener index

$$Wiener\_reverse = \frac{n^* \ n-1 \ ^*D}{2} - Wiener(G)$$

– Larger weights to peripheral bonds
– Smaller weights to internal C-C bonds

Expert opinion: Weiner number better than Randic

H. wiener J.Am. Chem. Soc., **69**,1947, 17- 20

Structural determination of paraffin boiling points

H. wiener,J. Am. Chem. Soc., **69**,1947, 2636-2638

Correlationof heats of isomerization and differences in heats of vaporization of isomers among paraffin hydrocarbons

Descriptor	All-path Wiener index	
	$ Type Index from $$ weighted	
Wiener	Z	

Descriptor	Balaban distance connectivity index	
$	**$Type Index from $$weighted**	
Balaban	Z	

	Mass	
	Van der Waals	Distance matrix
	Electronegativity	
	Polarizability	
Mean Wiener index		
Quasi-Wiener index (Kirchhoff number)		

	Mass	
	Van der Waals	Distance matrix
	Electronegativity	
	Polarizability	

Mean square distance index (Balaban)
Balaban centric index

Wiener polarity index [177]

Cycles of G have at most one common edge,

$$W_p\ G = M_2\ G - M_1\ G - 5*N_p - 3*N_h + |E_G|$$

Def: Wp(G) : Number of unordered pairs of

Vertices u, v of G such that distance $d_G\ u, v\ is\ 3$

Wp(G)	:	Wiener polarity index
M1(G)	:	First Zagreb index
M2(G)	:	Second Zagreb index
N_p	:	Number of pentagons
N_h	:	Number of hexagons
G	:	Molecular graph
N	:	Order of graph
u,v	:	Unordered pairs of vertices

Wiener Path

Total number of bonds between all pairs of atoms in the hydrogen suppressed graph

Descriptor | Partial Wiener Index | Topological

Formula

$$W = W_S + W_D + W_T + W_A = \sum_{b\in S} w_b^* + \sum_{b\in D} w_b^* + \sum_{b\in T} w_b^* + \sum_{b\in A} w_b^*$$

Bonds	
S	Single
D	Double
T	Triple
A	Aromatic

Descriptor Wiener Index for substituted compounds Topological

Formula

$$S_A = W_A + n_A * s_a + n_A * n_M + n_M * s_c$$
$$S_B = W_B + n_B * s_b + n_B * n_M + n_M * s_d$$
$$W = W_M + S_A + S_B + S_{AB}$$

Bonds	
M	Parent structure
A,B	Substituents
W	Weiner index
S_A, S_B	Topological substituent indices
N	Number of atoms

Lukovits, I.J. *Chern.Soc. Perkin* Trans., 2, **1988**, 1667-1671.

Decomposition of the Wiener Topological Index. Application to Drug- Receptor Interactions

Linear combination with Wiener index

A summation with due regards to algebraic sign of different topological indices played a critical role in descriptors research in explaining different types of phenomenon.

Descriptor	Lin_comb: Wiener + Szeged	Topological
Formula	$\eta(G) = Sz(G) - W(G)$ classification of graphs with $\eta(G) \le 3$	G : Molecular graph W(G) : Wiener index Sz(G) : Szeged index

	Classification of connected graphs which satisfy $\eta(G) \leq 4 or 5$	

Manso et al. [114] introduced a new topological descriptor (Fiindex) to predict boiling points of saturated/unsaturated and cyclic alkanes. It produced smaller deviations in prediction compared with other descriptors viz. Weiner, Hosoya and Randić.Chen and Chiu [171] proposed a novel set of Wiener indices and applied for a number of C_3–C_8 alkanes.

MolDesc	Fiindex	**Topological**
Application	Boiling point ▶ Alkanes ▶ Alkenes, alkynes cycloalkanes	
Robustness	Small deviations in property (BP)Compared to Weiner, Hosoya and Randić.	

Chemical interpretation of topological indices

Randić and Zupan [115] interpreted topological indices in terms of chemical bond contributions (chart 3.3). The proposed partition is straightforward for Wiener index and the molecular connectivity indices developed in terms of bond additive contributions.

Advances in Wiener index from Graph theoretical perspective

In organic compounds, branching is very typical to govern many physico-chemical properties. Graph theory provides natural mathematical frame for the quantitative codification of classical chemical bonding ideas (chart 3.4). However, there is no mathematical quantitativemeasure of it explicitly with complexity of branching. Here, star is the most branched graph, while path corresponds to least branched one.

Chart 3.3: Partitioning of molecular indices into chemical contributions

Inf.Bits. Partitioning of topological indices

- Balaban *J* index and the Harary index *H* partitioned.
  - Their development from the row sums of matrices is based on the partition of the path contributions to the bonds forming paths.
- Hosoya, modified-Hosoya indices partitioned based on the involvement of the individual bonds in the count of the *k*-disjoint edges of a graph
- Higher-order connectivity indices
- Topological indices are partition into bond contributions.
- Reviewed problems encountered with the interpretation of topological indices

Chart 3.4: Chemical tree

If	maximum degree of a tree is at most 4
Then	tree is a chemical tree

Li and Jalbout [173] reported hyper-Wiener index weighted by Mulliken bond orders are correlated with vibrational frequencies of formaldehyde derivatives. Hamzeh et al. [175] calculated the join, composition, Cartesian product, symmetric difference and disjunction of graphs under y-Wiener index. Wiener polynomials of a graph hyper-Wiener index of graph operations are generalized. Deng and Zhang [179] showed some pairs of trees and chemical trees indistinguishable revealing degeneracy of terminal_Wiener index. Liu and Liu [169]identifiedthe kth smallest and kth greatest indices of modified

Wiener indices for all k up to $\lfloor \frac{n}{2} \rfloor + 1$ among the class of trees of order n.The edge Szeged and edge Wiener indices of graphs are new topological indices hexagonal systems are also proposed.

In 2008, Wang and Zhang independently characterized trees with specified degree sequence that minimizes the Wiener index. Zhang et al. in 2010 pointed out that a corollary on maximizing the Wiener index was incorrect. Later, Çela et al. viewed this task as a quadratic assignment problem and arrived at an algorithm polynomial in time. Later information on the candidate trees for the maximum Wiener index is proposed.

Descriptor	Hyper Wiener Index	Topological
Formula	$^{m}Wiener = \sum_{m_{pij}} w_{ij} = \frac{1}{2} * \sum_{i=1}^{A}\sum_{j=1}^{A}\left[{}^{m}W \right]_{ij}$	$^{1}W, {}^{2}W, {}^{3}W, \ldots . {}^{D}W$ — Wiener numbers of different orders D — Max length of paths in graph i.e. topological diamater
	$Hyper\_Wiener = \widehat{W}(W) = \frac{1}{2} * \sum_{i=1}^{A}\sum_{j=1}^{A} w_{ij} = \sum_{m=1}^{D} {}^{m}W$	W' — Wiener operator W — Sparce Wiener matrix of m^{th} order

$W(G) = H'(G, 1)$	$H(G, x)$: Hosoya polynomial of a graph $H'(G, 1)$: First derivative calculated at $x = 1$	[187]
$WW(G) = HH'(G, 1)$	$WW(G)$, : Hyper-Wiener index $HH(G, x)$: Hyper-Hosoya polynomial $HH'(G, 1)$, : First derivative calculated at $x = 1$	[188]

Axioms for Wiener space

- Consider $(1,2)$-Sobolev space
- $W^{(1,2)}(U)$ on subsets U in an abstract Wiener space, which is regarded as a canonical Dirichlet space on U

Proved:

- $W^{1,2}(U)$ have smooth cylindrical functions as a dense subset if U is H-convex and H-open

- Hyper-Wiener index R proposed for acyclic structures
- Extended to cyclic molecules i.e. cycle-containing structures

[129]

Modified Winer index for trees	Topological

$$Wiener\ index\ of\ a\ tree\ T$$
$$Wiener\_Tree = W(T) = \sum_{e=1}^{T} n_1(e) * n_2(e)$$

$n1(e)$, $n2(e)$	number of vertices on the two sides of the edge e
T	Number of edges

Modified *Wiener index of a tree T*
(*Nikolić, Trinajstić and Randić*)

$$Mod\_Wiener\_Tree = {}^{m}W(T) = \sum_{e=1}^{T} \left[n_1(e) * n_2(e) \right]^{-1}$$

\+ The structural interpretation goes as greater weights of outer bonds compared to inner ones of a saturated hydrocarbon

Smallest and two greatest variable Wiener indices ${}^{m}W_{\lambda} \; T$	For all $\lambda \;\; \lambda \neq 0$ among trees on n vertices calculated
k^{th} smallest k^{th} greatest modified Wiener indices	for all k up to $\frac{n}{2}+1$ among class of trees of order n

Modified *Wiener index of a tree T*
(*Gutman, Vukičević and Žerovnik*)

$$Ext\_Wiener\_Tree = {}^{m}W_{\lambda}(T) = \sum_{e=1}^{T} \left[n_1(e) * n_2(e) \right]^{\lambda}$$

$$W(T) \xrightarrow{\text{modified to}} {}^{m}W(T)$$

$$W(T) \;\; {}^{m}W(T) \xrightarrow{\text{extended to}} {}^{m}W_{\lambda}(T)$$

Wiener index of a tree T

$$Wiener\_Tree = W(T) = \sum_{e=1}^{T} n_1(e) * n_2(e)$$

MolDesc	Pisanski-Zerovalk index	**Topological**
Alg.	**1)** Cal edge weight **2)** Sum all over pairs of vertices $W_{edge} = \sum_{i<j} \frac{P_{ij}^{b}}{P_{ij}}$ **3)** Add all edge weights $Pisa\_Zerrov = \sum_{b=1}^{B} W_b$	P_{ij} Number of paths of any length between vertices v_i and v_j P_{ij}^{b} Number of P_{ij} pathswhich include b bound W_b Edge weights W_{edge} Edge weights based on bond contribution
Characterstic	Global index	
For any graph $Pisa\_Zerrov = W$ is a valid end		If Acyclic Then Wiener index for acyclic graph $Pisa\_Zerrov = Wiener\ index$

q-Wiener index

Liu et al. [183] motivated by the theory of hypergeometric series, conceived q-analogs of Wiener index. The formulas are derived to compute*q*-Wiener indices of some compound trees. These generalize expressions are known earlier for Wiener index.

Derived descriptors from Distance matrix

The chronological developments in distance matrix amply demonstrate different frames of thought in conceiving the neighbourhood. Hosoya index, Rouvray index, Platt number etc are a few typical descriptors derived from distance matrix.

Descriptor	Hosoya index
Input	2-D Structure
Basis	Wiener index Extension to Wiener index
Def	Count of non adjacent edges (k) in a molecular graph. K edges chosen such that number two of them or adjacent
Alg	Length of shortest path connecting the ith and jth atoms in cyclic compounds
Application	Computer search for functional groups

Hoyasa index

\+ Accounting for variation of physico-chemical properties (like boiling point) of organic compounds (alkanes)

\- Larger weights to peripheral bonds

\- Smaller weights to internal C-C bonds

Descriptor	Rouvray index
Input	2-D Structure
Basis	Hosoya index/wiener
Alg	Sum of d_{ij} element
Related to	Rouvray Ind=2*wiener Ind
Compounds	Alkanes, alkenes, alkynes, arenas
Resp	MP, BP, Densitym, MR, Surface Tension, Viscosity
Model	Resp=a_1*(Rouvray Ind)^{a2}
Solution	L.S

Descriptor	Platt number
Input	2-D structure
Basis	Sum of first C-C neighbours of every C-C bond in the species
Feature	Improves correlation
Application	Multi parametric equation Resp=f(TI) TI: [wiener, polarity number, Platt]
Caution	cc (wiener, Platt) cc (Platt, polarity number) is worth studing

Average vertex distance degree
Square reciprocal distance sum index
Hyper-distance-path index
Reciprocal hyper-distance-path index
Mean distance degree deviation

Combination of distance matrix + adjacency matrix

Eigen analysis of weighted adjacency distance, distance matrices or their combination resulted in orthogonal matrices alleviating correlation problems.

Detourindex: The maximum distance matrix of Detour gave birth to new set of descriptors called Wiener* (W*) Harare* (H*) molecular topological index* (MTI*).

The difference between the descriptors using DistMat and DistMat* matrices contains new structural information. It is applied in SPropR for hydrocarbons.

Spanning tree number (log)		
Reciprocal hyper-Detour index	Distance/detour ring index of order	[3:9]
Distance/detour index		

Gravitational topological index: Here, topological distances instead of interatomic ones are used.

Descriptor	GravInd	FP	Scal	Topol
Def	Topological distance between ith and jth bonded atoms			

Basis	Intermolecular forces in bulk solvent media topological index based on topological distance	FP: Floating point Scal: one value for a compound i.e. tensor of zero order (Tens(0)
Limitation	– Varies little with isomers	

$$GravInd = \sum_{i=1}^{Natom-1} \sum_{j=i+1}^{Natom} \frac{m_i * m_j}{dist_{ij}^2}$$

m_i	:	Atomic mass of bonded atoms
Natom	:	Number of atoms in a molecules

Eigen value of (AdjMat + DistMat)

Descriptor	**Eigen value of (AdjMat + DistMat)**	Eigen value
Basis	Distance matrix	
Formulae	$eValues, eVec = eig\ DistMat + AdjMat$ $AscEig = AscOrd(EigValues)$ $\text{Large\_eig\_AD} = AscEig(1)$ $MolDes = \log\ \text{Large\_eig\_AD}$	x: $MolDes$ y: Physico chemical properties Model $y = fn\ MolDes; par$

Tensorial classification of descriptors:

Zero-order descriptors:The descriptors hitherto discussed yield only a single numerical value (scalar or zero order tensor) for each compound. Thus, Wiener, Hosoya, molecular volume, polarity number etc. is referred as zero-order descriptors.

Correlation among classical molecular descriptors

For a set of compounds, Winer, Hosoya indices and polarity number are correlated among themselves for a series of compounds or with some bulk properties. In the light of this, earlier results of two-three parametric regression models worth reinvestigation.

🕮 Randic Index

In 1975, Randic and Kier and Hall proposed connectivity index encoding degree of branching. It is calculated from adjacency matrix. This is the first non-empirical structural descriptors, which was renamed as molecular connectivity index (χ).

Vukičević[122]consideredchemical trees i.e. simple connected graphs with maximal degree at most 4 and thus path becomes extremal graph. The consequence is path alone remains in the analysisinstead both measures (star, path). Here, it is proposedgeneralized Randić index to be a suitable measure for branching.

$$Randic\_index = \sum d_u, d_v^{\ \alpha}$$

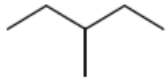
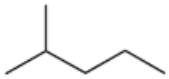
$$\alpha = \left(-\frac{1}{2}\right)$$

$If \quad \alpha = 1$
$Then \ \ it\ is\ weight\ of\ graph$
$Else\ \alpha \neq 0$

d(u)	:	Denotes the degree of a vertex u in G
d(v)	:	Denotes the degree of a vertex v in G
summation extends over all		

		edges i & v of G

Connectivity indices for hexane isomers

	Paths of length 2	Paths of length 3	Paths of length 4	Paths of length 5	$^0\chi$	$^1\chi$	$^2\chi$
	4	3	2	1	4.828	2.914	1.707
	5	4	1	0	4.992	2.808	1.922
	5	3	2	0	4.992	2.770	2.183

Andrew R. Leach, Valerie J. Gillet, **2007**, Springer, An Introduction to Chemoinformatics (Revised Edition).

- Randić index
- Determinants of the adjacency matrix
- Distance matrix to maximum wave length (λ_{max})

Correlated for conjugated polyenes

Descriptor	Randic Index	FP	VecCol	GraphTh

Def	Summation over all paths of length one in structural graph
Input	2-D Structure
Basis	Modified form of Zagrab group index
Model	Sum of bonds connecting pair of atoms other than hydrogen. Each atom is encoded by a cardinal number
Encodes	Degree of branching Weighted count of skeletal bonds Sigma bond (SP3 carbon) connectivity Dispersive nature of molecules
Characteristics	Non-empirical structural descriptor Not a measured property Not derived from or translated into a physical property Skeletal branching index or molecular
Algorithm	Powers of adjacency matrix
Property	Graph invariant
KB	If Branching increases Then Index decreases
Applicable	Acyclic alkanes Molecules can be classified as compact/spherical and extend

Randic Index	**Formula**	**Order**
Randic V0	$\sum_{i=1}^{N} \left(\frac{v_i}{N}\right)^{-0.5}$	0
Randic V1	$\sum \left(v_i * v_j\right)^{-0.5}$	1
Randic V2	$\sum \left(v_i * v_j * v_k\right)^{-0.5}$	2
Randic V3	$\sum \left(v_i * v_j * v_k * v_l\right)^{-0.5}$	3
Randic Vh	$\sum \left(v_i * v_j * \ldots\ldots\ldots * v_h\right)^{-0.5}$	

For hetero atoms

Randic V1-mod = $\sum \left(V\mathrm{mod}_i * V\mathrm{mod}_j\right)^{-0.5}$

Vmod: Sum of relative atomic weights of all adjacent atoms

– Limitation	No provision for compounds with hetero atoms, cyclic – and unsaturated structures It does not distinguish positional isomers Degeneracy • Example: does not distinguish cis-/trans-isomeric or chiral phase information Does not encode binding information
Remedy	Kier Hall Valence Molecular Connectivity

Randic index	[0:3]

Randic shape index	path/walk [2:5]
Kier alpha-modified shape index	path [1:3]
Kier and Hall	[1:3]
Kier and Hall shape index	[1:3]
2D Petitjean shape index	

Total adjacency	(+) graph invariant (-) highly degenerate
Randic	(+) increases with size (-) decreases with complexity

Kier symmetry index
Kier flexibility index
Kier benzene-likeliness index

Connectivity index	Randic	
	Average	[0:5]
	Valence	
	Average valence	
	Solvation	

Connectivity index	Modified Randic
	Reciprocal distance
	Reciprocal distance Squared

Narumi index (log)	Simple topological
Narumi topological index	Harmonic
Narumi topological index	Geometric

Pogliani index
Ramification index
log of product of row sums (PRS)

eccentric connectivity index
eccentricity
average eccentricity
eccentric

Schultz Molecular Topological Index (MTI)
Schultz MTI by valence vertex degrees

Gutman Molecular Topological Index • molecular topological index based on valence vertex degree
Gutman MTI by valence vertex degrees

Xu index

First Mohar index TI1
second Mohar index TI2

Detour index
Hyper-detour index

Descriptor	Comment	Ref
$Modified\_Randic\_index = \sum \max\ d_u, d_v^{-1}$	Represents a lower bound of Randić index	[124]
generalizations of R' and its counterpart, R'', real number α $Generalized\_R' = \sum \min\ d_u^{\alpha}, d_v^{\alpha}$	Lower bound for the generalized Randić index α : Real number	[124]
$Generalized\_counterPart\_R' = R_{\alpha}^{"}\ G = \sum \max\ d_u^{\alpha}, d_v^{\alpha}$	Upper bound	
Randić index of a triangle-free graph G with given minimum degree $\delta(G)$	Best-possible lower bound	[118]
Randić energy of a bipartite graph & Graphs whose adjacency matrix is partitioned into blocks with constant row sum	Sharp upper bound	[119]

Randić index of a graph of order n without isolated vertices is at least n−1
For a connected nonbipartite graph G with $n \geq 5$ vertices and girth $g \geq 5$, bound is best possible when the graph is composed of a cycle C_5 *on* 5 vertices & a tree T on $n-4$ vertices sharing a single vertex $S_z\ G\ -W\ G\ \geq 2*n-5$
For a connected nonbipartite graph G with $n \geq 4$ vertices, bound is best possible when the graph is composed of a cycle C_3 on 3 vertices & a tree T on $n-4$ vertices sharing a single vertex.
Proved that for every large enough integer w there exists a tree with Wiener index equal to w there exists $2^{\Omega^* \sqrt[4]{w}}$ non-isomorphic trees with Wiener index w [Wiener-18]
Concerning the minimum Randić index among all n-vertex connected graphs with the minimum degree at least k True

Conjectures		Ref
Aouchiche and Hansen ♎ Proved for: Graphs in $G(k,n)$ for which Randić index attains its minimum value for $k \le \frac{\pi}{2}$	α : Real number	[128]
♎ Generalized results of [extremal graphs are complete split graphs $K^*_{k,n-k}$, with two degrees $(k, n-1)$] to graphs with prescribed maximum degree q	▶ Number of vertices of degree k is $n-k$ ▶ Number of vertices of degree $n-1$ is k.	[128]

Randic Index	Formula $$$$
General Randić matrix	Matrix $R_\alpha = \left[R_\alpha \right]_{ij nn}$ of G is defined by $R_{\alpha\ ij} = d_i * d_j^{\ \alpha}$ if vertices v_i and v_j are adjacent in G 0 else Let d_i be the degree of the vertex v_j
Randić signless Laplacian matrix	matrix , $Q_\alpha = D^{2*\alpha+1} + R_\alpha$ where α is a nonzero real number and D is the degree diagonal matrix of G
General Randić energy	Energy RE_α is the sum of absolute values of the eigenvalues of R_α .
General Randić incidence matrix	Matrix $B_{R_\alpha} = \left[B_{R_\alpha} \right]_{ij nm}$ of a graph G is defined by $B_{R_\alpha\ ij} = d_1^{\alpha}$ if v_i is incident to e_j 0 else
General Randić incidence energy	BE_α : sum of singular values of B_{R_α}
	Lower and upper bounds on general Randić incidence energy.

KB

Generalized Randić index $R_p(G) = \sum_{uv \in E(G)} d_u * d_v^{\ p}$ is a suitable measure for branching iff $p \in \left[\lambda, 0 \bigcup 0, \lambda' \right]$;

λ: solution of equation $2^x + 6^x + \frac{1}{2} * 12^x + \frac{1}{4} * 16^x - \frac{11}{4} * 4^x = 0$ in the interval $-0.793, -0.792$

λ' : positive solution of the equation $3 * 3^x - 2 * 2^x - 4^x = 0$

Zagreb index: It is the sum of squares of number of connections to other non-hydrogen atoms irrespective of bond order

Descriptor	Zagrebindex	
Input	Distance matrix	

$$zagreb(1) = \sum_{i=1}^{Natom} \sum_{j=1}^{Natom} distMat_{i,j}^{2}$$

$$zagreb(2) = \sum_{i=1}^{Natom} \sum_{j=1}^{Natom} distMat_{i} * distMat_{j}$$

Zagreb (1) =	Also called Gutman index	
	Zagrb(1) = F + 2(A-1) = ***2 *** (N2 +***A*** - 1);	
	Related to zero order connectivity index + Increase with size – Increases with complexity	F *Platt number* N2 *connection number*
Zagreb (2) =	Related to first order connectivity index + Increases with size – Increases with complexity	

First and second Zagreb index	m = 1 to 2
First Zagreb index by valence vertex degrees	m = 1 to 2
Quadratic index	

Def: Sum of squares of number of connections to other non-hydrogen atoms irrespective of bond order

🕮 Kier and Hall connectivity index

Kier and Hall inspired by Huckel MO calculations and graph theory put forward this concept of connectivity. It is calculated as the sum of reciprocal of geometric mean of δ terms. It is known that geometric mean is more unique than arithmetic mean. The basis of the model, features, limitations and remedial measures are described. The χ index was extended to compounds with hetero atoms, multiple bonds and cyclic structures. The physic-chemical interpretation of terms in connectivity index based on simple and valence connected molecular descriptors, the algorithm, limitations and misuse of this connectivity index through light on the intensive research on the molecular descriptors.

🕮 Kier and Hall χ indices of multiple orders

Kier and Hall later introduced χs of different order (m=0 to 6) to alleviate the limited structural characteristics of the first order Randic index.They encode counts and weights of structural features like atoms, bonds and topological assemblies (cluster and path-cluster). The connectivity index (χ) is limited to normal alkanes.On the other hand, vertex valence (χ^{v}) is a term used by mathematical graphical theoreticians and accounts for heteroatoms, multiple bonds and cyclic structures. Thus, χ and χ^{v} have same numerical magnitude for alkanes. Molecular connectivity indices (${}^{m}\chi$ and ${}^{m}\chi^{v}$) are of the fixed type. Some of the variable connectivity indicesput forward are variable Balaban index and path variable weights.Molecular orbital and graph theoretical principles enable to represent the structure of a molecule by molecular graph reflecting atom-to-atom chemical connectivity. In this back drop, Kier and Hall introduced connectivity indices of different (zero to sixth) order (χS) (Formulas: zzz). They alleviate limited structural information of Randic index.

Descriptor	Kier and Hall connectivity indices
Encodes	Counts and weights of structural features Atoms, bonds Topological assemblies (path, path cluster) Encodes specific fragment Bond additivity
Independent of	Conformers 3D-Geometry
Advantage	+ Alleviates limited structural information Randic index
Limitations	– Decreases with complexity – Intermolecular forces (polarizability, volume) are not considered – Does not take into account heteroatoms and multiple bonds (=, ≡) i.e. χ is same for ▪ CH3.NH2 ▪ CH3-CH3 and ▪ CH2=CH2
Misuse	– Employing of χ for o-, p- and m-substituted compounds where polarity operates • **Reason:** X is designed for bond-additivity

KB 3.1:Choice of χs based on bonding

If σ bonds are needed & only alkanes
Then $^{1}\chi$ is adequate

If Multiple bonds or hetero atoms or acylic compounds
Then $^{1}\chi^{v}$ is used

If π and lone pair of electrons are needed
Then Valence state electro-negativity δ ratio is used

If Boiling point
Then Non-covalent interactions considered

If Nucleophilic attack
Then Atom based descriptors are used.

Descriptor	CRI indices
Basis	Connectivity matrix of valence vertex
Formulae	eValues = eigen value (ConnectMat_Val) eValues_positive = (eVlaues > 0) CRI = sum (eValues_positive)
Encodes	Information of all connectivities in H-depleted molecular graph
Discrimination	Sensitive to presence of hetero atoms

Connectivity

Degeneracy of χ and remedial measure

A molecular descriptor say χ distinguishes the connectivity with increasing carbon atoms. However, degeneracy develops for some conformers. This renders the descriptor inadequate to reflect the chemical information. This is the reason why descriptor is modified. χis corrected for double bonds, hetero atoms cyclic rings etc. (KB 3.1). Even then, the problem of degeneracy is eliminated and so 3D Weiner index is introduced. Newer descriptors are based on different basis of looking at the chemical information in the molecule.

Linear combination of χs

Linear combination(difference and addition) of χ and χ^{v} have been found to be potential descriptors encoding information not covered by either of the two (chart 3.5). The difference ($^{m}\chi$ - $^{m}\chi^{v}$) reflects electronic information with regio-specific properties.

Correlation between linearly combined descriptors

The derived descriptors by linear combination of connectivity indices have new type of information. However, their use along with others deserves attention from statistical/chance/chemical partial/completecorrelation (chart 3.6).

Chart 3.5: Linear combination of χ and χv SXR

Bio concentration factor	Chlorinated hydrocarbons Benzene Biphenyl Diphenyl oxides	$\chi^0 + {}^0\chi^v$	$\chi^0 + {}^0\chi^v{}^2$
		$\chi^0 + {}^0\chi^2$	$\chi^0 + {}^0\chi^2{}^2$
Ionization potential	Alkyl amine Alcohol Ether	$\Delta^0\chi$	$\Delta^1\chi, \Delta^2\chi$

Chart 3.6: Limitations and remedial measures of Topological descriptors

Limitation	Remedy
– Do not encode 3D-infomaion of molecule	ॐ Geometric ॐ WHIM ॐ QC ॐ Topo-electric
– No electronic characteristics of atom	ॐ QC
– Statistical linear correlation	ॐ Orthogonalization
– Higher order connectivity indices no physico-chemical significance	---

Tensorial classification of descriptors

First-order descriptors: Zagreb, Randic, Kier and Hall indices are vectors of descriptors per moleculeand thus referred as first order category. It no way reflects the order (0 to 6D) of representation of the molecule.

Correlation of connectivity indices

Connectivity indices of different order form a first order (vector) descriptor. The information in successive elements of the vector is overlapping resulting in correlated quantities. The extent of overlapping information is not transparent to a naked eye. But, the use of all these (statistically linear) correlated descriptor sets results in ill conditioning of the design matrix of the cause-effect model.

Remedy to correlation hurdle: Randic orthogonalized second, third etc. indices with respect to first connectivity index resulting in reliable model free parameters. The fringe benefit is separating the overlapping information present in the raw data matrix. Here, the number of orthogonal descriptors is equal to the number of correlated ones and thus there is no dimension reduction.

PC and PLSC analysis: The less the number of explanatory variables, the better is the model. Rotation of χspace into PC or PLSC space is preferable. PC and PLSCanalysis followed by regression is a sought after in drug research as superior models.

ॐ Environment descriptors

It describes the fragment surrounding an atom as a single parameter. It is calculated from its first and second nearest neighbors and their bonds. The algorithms of BED and AED are inchart 3.7 and alg. 3.1.

Chart 3.7: Fragment descriptors

Environment Descriptor	Significance	Abbreviation
Bond	Number of bonds used	BED
Weighted	Type of bond contains information about isolated atom fragments	WED
Augmented	Uses both type of bond of atom	AED

Alg. 3.1a: Bond Environment Descriptor (BED)

Step:0 **Input:** 1D-strucuture of molecule

Step:1 Hydrogen implied (supressed) graph

Step:2 Locate fragment of interest (eg. C=O)

Step:3 First nearest neighbor

No	Atom	Bond	
1	C	Single	1
1	O	Single	1
1	O	Double	2
	a		**4**

Step:4 Count number of non-hydrogen bonds

Step:5 Second nearest neighbor C... O-C

Alg. 3.1b: Augmented Environment Descriptor (AED)

Step:0 **Input:** 1D-strucuture of molecule

Atom	
C-C	1
C=O	4
C-O	2

Step:1 Formulate connection table

Step:2 **For** each nearest neighbour
Cal product of bond type and numeric value for each atom and bond
end

Step:3 Sum the product

Step:4 Repeat step 2 for nearest neighbor atom (y)

Step:5 AED = x + y

ॐ Electrostatic descriptors

The descriptors under this category reflect charge distribution and are calculated by non-quantum chemical methods. The important sub classes of electrostatic descriptors are based on partial charge, topo electronic measure and surface area (chart 3.8, Fig. 3.2).

Partial charge based electrostatic descriptors

The empirical partial charges ae calculated by Zefirov's approach. The molecular electronegativity is calculated as geometric mean of atomic electronegativities. The minimum and maximum of partial charges on the molecule/fragment/atom (C, O, P etc.) are a few typical descriptors. The derived ones are the difference of maximum and minimum charge and are equal to the polarity. The normalized parameter obtained by weighting with reciprocal of the squares of the distance between atom bearing maximum and minimum charge has been in vogue in descriptor research.

Molecular surface area

The calculation steps of Van der Waal's (polyhedral) surface area are in Alg. 3.2. Van der Waal radii, interaction energies, sum of atomic volumes etc. used in SPropR to account for cohesive forces in physical properties.

Chart 3.8 : Classification of Atom based topological descriptors

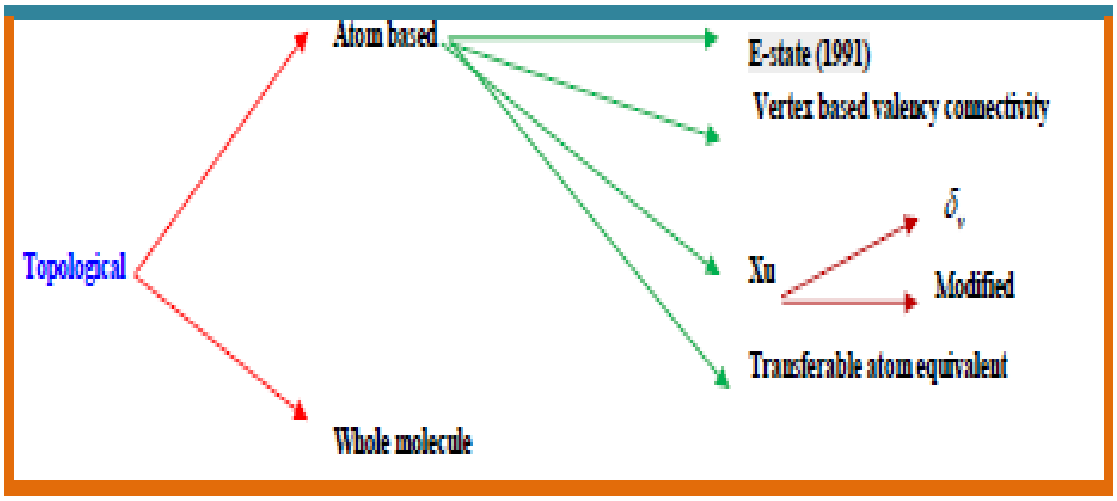
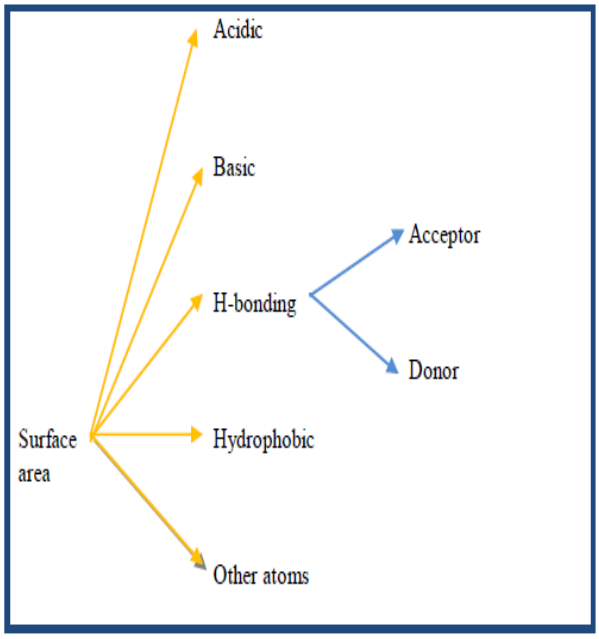


Fig 3.2: Classification of Surface area descriptors



Alg.3.2: Molecular surface area	
Step: 0	**Input:** step size dx
Step: 1	Divide molecule into number of slices (of dx) along X axis
Step: 2	**For** each slice Generate a set of circles corresponding to intersections of Van der Waals sphees and cutting planes **End**
Step: 3	**For** each circle Find lengths of non-occluded arcs using the step dl **End**
Step: 4	Surface area = sum (lengths of non-occluded arcs) * slice thickness (dx)

ॐ B-cut descriptors

Pearlman et al. used atomic charge, polarizability or H-bonding ability instead of atomic number of diagonal matrix of Burden matrix. The minimum and maximum Eigen values for the three matrices are calculated. It results in either a 2D- or 6D- descriptor space encoding chemical structural information. An open question posed was regarding the adequacy of 6D for representing the structure explaining all types of properties. Now, Eigen value descriptors in very high dimensional space (200 to 400) are available from theoretical UV-, IR-, NM-R spectra.

Molecular identification number

Burden proposed a descriptor combining connectivity information with atomic number. The Eigen values results in molecular identification number.

Descriptor	B-Cut
Authors	Frank Burden CAS University of Texas, (Bob Pearlman)
Basis	Burden identification number Cell based partitioning of chemical space
Encodes	Intermolecular interaction
Advantage	Non correlated space
Limit	Non intuitive nature
Application	Small molecule receptor (bi molecular) bonding

Alg. 3.3 B-cut descriptor	
Step 0	2D strucdture of compound
Step 1	Diagnonal matrix $Diag_{ii}$ = [atomic charge; Atomic polarisability; H-bonding ability]
Step 2	EV = eig(A=Diag)
Step 3	M1 2D space = min(min(ev1,ev2,ev3)) M2 = max (max(ev1,ev2,ev3))
Step 4	6D space [ev1,ev2,ev3,....]
Step 5	Output : 6 indices corresponding to six dimenstions.

BCUT_$$$	Diagonal elements correspond to	BCUT_$$$	
Mass	Atomic masses	highest	[1:8]
		lowest	[1:8]
Electronegativity	Sanderson Electronegativity (SE) of the atoms	highest	[1:8]
		lowest	[1:8]
Vdw radius	Van der Waals radii of the atoms	highest	[1:8]
		lowest	[1:8]
Atomic charges	Atomic charges	highest	[1:8]
		lowest	[1:8]
ALOGP	ALOGP of the atoms	highest	[1:8]
		lowest	[1:8]
E-State	Electrotopological state values of the atoms	highest	[1:8]
		lowest	[1:8]
Polarizability	Polarizability of the atoms	highest	[1:8]
		lowest	[1:8]

Descriptor	**Randic Molecular ID**
Formula	Square roots of reciprocals of the first nine prime numbers for nine edges commonly encountered in chemical graphs + Prime numbers instead atom valences are used

Descriptor	Molecular identification number
Authors	Burden
Input	2D-structure
Basis	Hybrid molecular connectivity and molecular path count
Formula	$D^2 = \sum (ai - bi)$
	a_i, b_i : Molecular path length i of two structures
Encodes	Measures dissimilarity
Similar to	Euclidean measure of dissimilarity

Burden, F.R J. Chem.Inf. Comput.Sci.,

29, **1989**, 225-227
Molecular Identification Number for Substructure Searches

Descriptor	Burden matrix
Input	**H**-depleted molecular graph
Output	Burden matrix

Bond	$=0.1*\pi^*$
Single	0.1
Double	0.2
Triple	0.3
Aromatic	15

- Diagonal element B_{ii} : Atomic numbers Zi of the atoms
- Off-diagonal elements Bij: two bonded atoms i and j are $=0.1*\pi^*$

Descriptor	Eigen values_Burden modified
Def	largest absolute eigenvalues of Burden matrix
Formula	[Evec,eigVal_Burden] = eig(BurdenMat)
	L : Maximum length 15; user defined
	BME : $BME \equiv \langle \lambda_1, \lambda_2, ..., \lambda_L \rangle$

ॐ Geometric descriptors

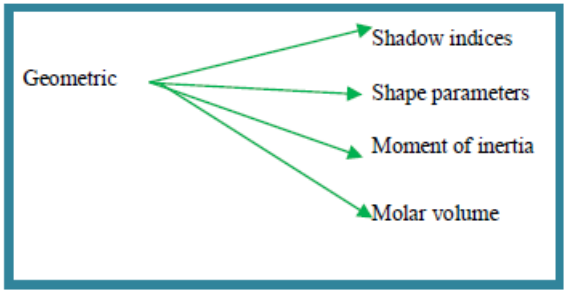
The descriptors calculated from optimized 3D-coordinates of fragment/atoms in a molecule are under geometric category (Chart 3.9). They reflect molecular size and shape which play a pivotal role in correlating biological response (BR) and molecular properties. The quality of geometric descriptors does depend upon the quality of geometric optimization, convergence and algorithm of CQC. Although Cartesian coordinates are obtainable from X-ray crystallographic data, they are generally derived from optimised geometry of the molecule by quantum chemical methods. A multistage refinement of the geometry starting with molecular mechanics followed by AM1, or PM3 and finally with ab initio methods is rarely taken up. Some of the important geometric descriptors widely used in structure property relationships (SPR) are Wiener_3D, Randic_3D, WHIM, shadow indices, shape parameters, moment of inertia, molar volume etc.

Moment of inertia

It signifies distribution of mass along three principle moment of inertia axes. These are calculated by rigid rotator approximation. The descriptor is calculated with and without hydrogen suppressed graph concept. Not only the total and components of moment of inertia along the three axes but also the ratios have been reported to be prospective explanatory variables.

Chart 3.9 : Geometric descriptors

Input	XYZ coordinates, Z matrix
Basis	Electronic potential energy
Chemical significance	Minimum PE configuration is optimum geometric structure
Invariance to	External refernce frame



Descriptor	**3D-Wiener index**	**Topological**
Input : 3D optimized geometry		
$^{3D}Wiener_H = \frac{1}{2} * \sum_{i=1}^{A} \sum_{j=1}^{A} r_{ij}$	***rij*** : Interatomic distance between the ith and jth atoms G Optimized 3D-geometry	

Knowledge bits

- Largest values corresponding to the most extended conformations
- Smallest to the most compact conformations
- \+ Index is obviously more discriminating the 2D-Wiener index and shows different values for different molecular conformations
- 3D-Geometry is also calculated from H-atom depleted structure

Descriptor	**3D-Balaban_ Diudea**	**Topological**
Input : 3D optimized geometry		

$$^{3D}J = \mathcal{IB}(G) = \frac{B}{C+1} \cdot \sum_b \left({}^{G}\sigma_i \cdot {}^{G}\sigma_j \right)_b^{-1/2}$$

$$DJ = \sum_{i=1}^{A} dj_i = \sum_{i=1}^{A} \sum_{j \in V_{i1}} \left(\frac{\sigma_i}{w_i(1+f_i)} \cdot \frac{\sigma_j}{w_j(1+f_j)} \right)^{-1/2}$$

- Inner sum runs over all vertices *j* at distance 1 from the ith atom, i.e. vertices bonded to i^{th} atom

A	Atom number
f	Multigraph factor
w	Weighting factor heteroatoms Accounted
dj	Local vertex invariants Accounts for heteroatoms and bond multiplicity

If	t factor w is equal to one & multigraph factor is equal to zero
Then	index DJ is related to the Balaban $DJ = 2 * J * \frac{C+1}{B}$

Balaban, A.T. and Diudea, M.V.	J. Chem. Inf. Comput. Sci., 1993, 33,421-428.

Real Number Vertex Invariants: Regressive Distance sums and Related Topological Indices

Shape index

Shape index represents the area of the rectangle enveloping the molecule (Alg.3.4)

Shadow index

It reflects the size of the natural shadow of the molecule. The X-coordinate of the molecule is aligned along the main axis of inertia. The shadow indices and the normalized ones are calculated as descried in (Alg.3.5). The geometric distances between the atoms (N, O, S, P, F, CL, Br, I) are outputted by many packages and account for processes involving interatomic distance.

Alg. 3.4 : Shape index

Step: **One**	Draw a rectangle enveloping molecule with minimum area
Step: **Two**	Measure longer (long) and shorter(short) sides of rectangle

Alg. 3.5: Shadow index

Step: **One**	X-coordinate of molecule is along main axis of inertia
Step: **Two**	Orientation of molecule in space along axes of inertia

Step: **Three** Shape index = $\frac{long}{short}$

Modified shape index

Step: **One** Draw rectangle such that length to brreadth ratio is maximized

Step: **Three** Calculate areas of shadows (S_1, S_2, S_3)

Significance Size (natural shadow indices) of molecule

Area	Plane of molecular shadow
$\frac{S_{XY}}{X_{max} * Y_{max}}$	X_{max} Maximum dimensions of molecule along corresponding axes Y_{max} Z_{max}
$\frac{S_{XY}}{Y_{max} * Z_{max}}$	
$\frac{S_{ZX}}{Z_{max} * X_{max}}$	

Interatomic distances

[X] = [N, O, S, P, F, Cl, Br, I]

[Y] = [N, O, S, P, F, Cl, Br, I]

[X..Y] = [N..N, N..O, N..S, N..P, N..F, N..Cl N..Br, N..I, O..O, O..S, O..P, O..F, O..CL, O..Br, O..I, S..S, S..P, S..F, S..Cl, S..Br, S..I, P..P, P..F, P..Cl, P..Br, P..I, F..F, F..Cl, F..Br, F..I, Cl..Cl, Cl..Br, Cl..I, Br..Br, Br..I I..I]

Shadow indices

Area	Plane of molecular shadow
S_{XY}	XY
S_{YZ}	YZ
S_{ZX}	ZX

L.B. Kier Acta.pharm.Jugosl.36(1986) 171-188
Indexes of molecular shape from chemical graphs

L.B. Kier Ed. D. H. Rouvray Nova science publishers, N Y 1990 (151-174)
Indexes of molecular shape from chemical graphs computational chemical graph theory

ॐ RDF (Radial Distribution Function) descriptors

The distribution of distances in 3D-geometric coordinate frame of a molecule is the basis of RDF category. The RDF of an ensemble of A atoms reflects the probability distribution of picking up an atom in a volume of sphere of radius R. Further, by including atomic properties as weighting factors, more useful information for regression and interpretation is generated.

Descriptor RDF RDF

$$Gauss(R) = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} w_i * w_j * e^{-\beta * \left(R - r_{ij}\right)^2}$$

$Gauss(R)$: Probability distribution of the individual interatomic distance

w	Atomic properties of i^{th} and j^{th} atom
r^{ij}	Interatomic distances between i^{th} and j^{th} atom
β	temperature factor referring to movement of atoms

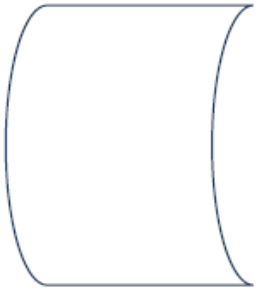
Inf.Bits

+ Gauss(R) meets all requirements for a 3D structure descriptor
+ Independent of number of atoms, i.e., size of a molecule
+ Unique regarding 3D- arrangement of atoms

- Invariant against translation and rotation of entire molecule
- Describes steric hindrance or structure/activity properties of a molecule about bond distances, ring types, planar and non-planar systems and atom types

3D-MoRSE (3D-Molecule Representation of Structures based on Electron diffraction) descriptors

A transform technique is employed in electron diffraction to prepare theoretical scattering curves. It is the basis of 3D-Morse descriptors generated from optimized 3D-optimized XYZ coordinates of atoms of a molecule. A 3D-MORSE code consisting of a 100-dimensional binary vector is generally used.

Descriptor	3D-MoRSE			
RDF_Wtd		Unity	signal	[1:30]
		Mass	signal	[1:30]
		van der Waals volume	signal	[1:30]
		Sanderson electronegativity	signal	[1:30]
		polarizability	signal	[1:30]
		Ionization potential	signal	[1:30]
3D-MoRSE		I-state	signal	[1:30]

$$G(\mathbf{s}) = \sum_{i=1}^{A} f_i \cdot \exp(2\pi i \cdot \mathbf{r}_i \cdot \mathbf{s})$$

$$s = 4 * \pi * sin\frac{\vartheta / 2}{\lambda}$$

$$I(s) = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} w_i \cdot w_j \cdot \frac{\sin(s \cdot r_{ij})}{s \cdot r_{ij}}$$

I(s)	Scattered electron intensity
w	Atomic property
rii	Interatomic distances between the i^{th} and j^{th} atoms
s	Scattering in various directions by a collection of A atoms located at points r_i
h	Wavelength of the electron beam
fi	Factor accounting for direction dependence of Scattering from a spherical body of finite size
ϑ	Scattering angle
λ	Wavelength of electron beam

Quantum chemical descriptors

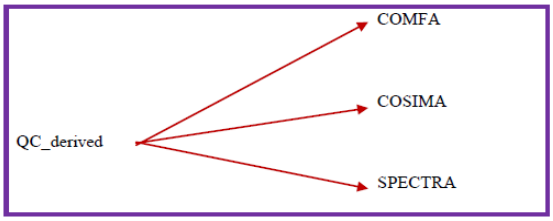
Quantum chemistry provides information that is not easily obtainable at molecular level. QC packages [AMPAC, MOPAC, HYPERCHEM, Schrodinger, GAUSSIAN 09, GAMESS etc.] output a large amount of geometric, electrostatic and electronic information for a chemical moiety. The parameters include optimised geometry, charge distribution at atom level, occupancy of electrons in each molecular orbital (MO), different types of charges, total energy and its components like core-core repulsion and electron–electron repulsion/attraction. HOMO (highest occupied molecular orbital), LUMO (lowest unoccupied molecular orbital), frontier orbital, total electron density and transition state energies are typical descriptors not amenable by experimental and other mathematical procedures. These parameters are instrumental in calculating physically significant quantities like dipole moment, ionisation potential and electron affinity.

Electrophilic and nucleophilic reactivities of the atom derived from quantum mechanical parameters predict the chemical interaction before performing the reaction. The other descriptors of chemical significance are Mullikan ($\sigma-\sigma$, $\pi-\pi$ and $\pi-\sigma$) bond orders, which explain the shifts in UV-visible spectra. The total surface area and volume obtained from optimized geometry are employed in arriving at atom, positive/negative solvent accessible areas in simple and normalized forms. The descriptors of Jurs are successful in explaining structure property relationships for compounds of diverse structures.

The computation of quantum chemical parameters for a whole molecule, fragment or atom was confined earlier to gas phase and only to small molecules with a few atoms in the periodic table. Due to concerted efforts in extension of theoretical results and development of fool proof commercial (with direct participation of expert quantum chemists in academia) packages, now it is a child play to derive QC parameters for large molecules consists of even hundred atoms and many of the elements in the periodic table, almost in all the three states of matter. The CQC derived spectral and molecular field are given chart 3.10. A general tendency is to limit first to as gas phase, followed by recalculation in solution (aqueous or non-aqueous) phase employing various solvation models like Born, Kirkwood, Poisson, Boltzman, Polarisable Continuum model (PCM) and Langevin-Dipole (LD) method.

Descriptor	Quantum-chemical descriptors
Input	3-D structure
Method	Semi empirical or ab initio or DFT
Definition	Molecular descriptors calculated based on data obtained from QC principles
Alg	• Semi empirical • AM1 • PM3 • Ab initio • DFT

Chart 3.10 Descriptors derived from CQC spectra



Charge based descriptors

Many of the physico-chemical processes like protonation, ion exchange and electrophoresis have a bearing on the charge on the species. Alg. 3.6 lists total, maximum, minimum and relative charge of different nature viz. positive, negative and absolute. The other important subclasses are based on polarity, partial and square of the charge. The H-bond based charges have significant role in biomolecules and their interaction with drugs, foods and small ionic species.

Alg. 3.6 QC derived charges

$Q^{-}_{\max} = \max \lvert qneg \rvert$	Maximum negative charges
$Q^{+}_{\max} = \max \ qplus$	Maximum positive charges
$\Delta = \lvert Q^{+}_{\max} - Q^{-}_{\max} \rvert$	Submolecular Polarity Parameter
$Q^{+} = \sum_{i=1} qplus_i$	Total positive atomic charge
$Q^{-} = \sum_{i=1} qneg_i$	Total negative atomic charge
$Q = \sum_{i=1}^{Natom} \lvert q_i \rvert$	Total absolute atomic charge

Descriptor	Extreme (minimum and maximum) values of the partial charges
Moiety	Atoms (N) of a given chemical species (adenosine) in the molecule (ATP)
Example	δH (min) is the minimum (negative)
Basis	MO LCAO theory Mulliken definition for atomic charge
Definition	On an hydrogen atom in the given molecule considered in the calculation

$Qsquare = \sum_{i=1}^{Natom} q_i^2$	Total squared atomic charge

	$Qpolariz = \frac{\sum_{i=1}^{Natom} q_i}{Natom}$	Charge Polarization
	$TopElecInd = \sum_{i=1}^{Natom-1} \sum_{i=1}^{Natom} \frac{\lvert q_i - q_j \rvert}{r_{ij}}$	Topological electronic index

```
Function chargeDes(qpos,qneg)
MaxNegCharge = max(abs(qneg))
MaxPosCharge = max(qpos)
TotPosCharge = sum(qpos)
TotNegCharge = sum(qneg)
TotNetCharge =  TotPosCharge + TotNegCharge
TotAbsCharge =  TotPosCharge + abs(TotNegCharge)
```

```
>> qneg  =  [ -1   -2   -4]
>> qpos  =  [ 2    3    6]

>>charges(qpos,qneg)
```

```
MaxNegCharge      4
MaxPosCharge      6
TotAbsCharge     18
TotNegCharge     -7
TotNetCharge      4
TotPosCharge     11
 TotSqCharge     70
```

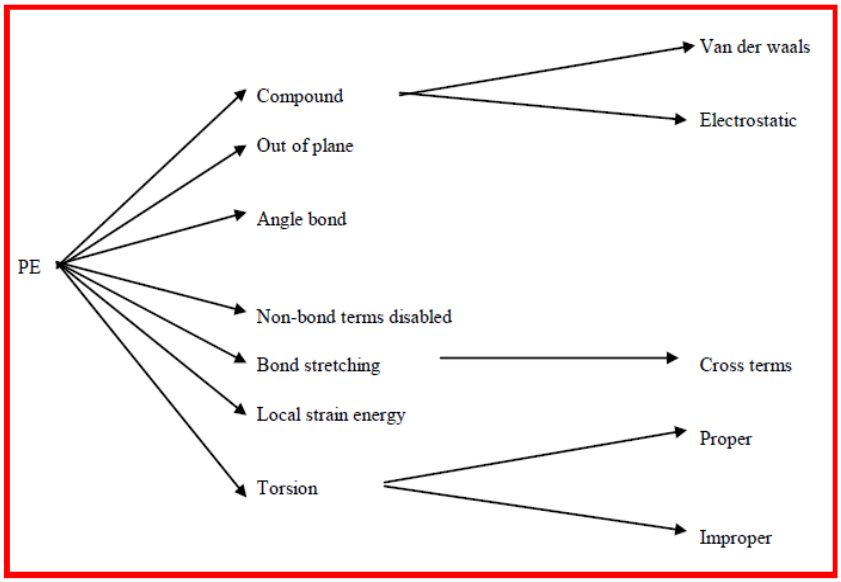
```
%
%  charge99.m  (R S Rao) 29-5-15
%
clc, clear

qpos9 = [2 3 6]', qneg9 = [-1  -2 -4]'
q = [qpos9' qneg9']'

tol = eps,p = 0;n = 0;length(q)
for i = 1: length(q)
  if q(i,1) >= tol
    p = p+1
    qpos(p ) = q(i,1)
  else
    n = n+1
    qneg(n ) = q(i,1)
  end
end
qpos, qneg
clc, clear
```

QC energy based descriptors

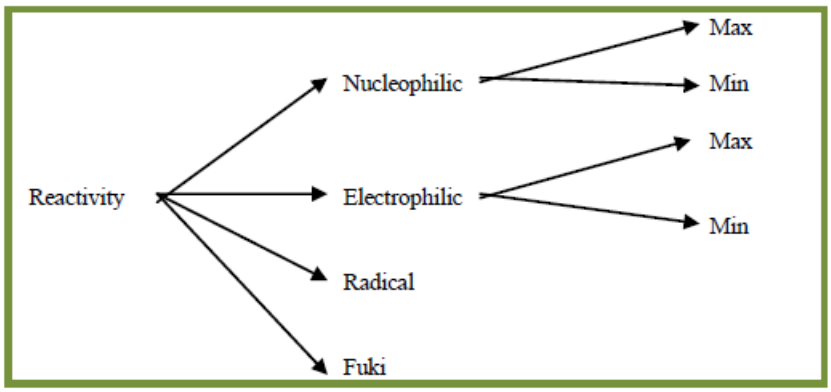
The total energy computed from computational chemical methods depends upon the rigor of the model and it is one of the parameters of utmost significance. The standard scale in QC calculation is considering the isolated electrons and nuclei at 0°K. In chart 3.11, the potential energy component for bond angle, bond stretch, out-of-plane torsion are described An inspection of its components down to the level of electron and nucleus gives an opportunity to relate micro processes in chemical interaction. The magnitudes of QC descriptors vary significantly as they are developed in different energy scales based on a variety of partition theories.

Chart 3.11: Components of QC energy

Quantum mechanical molecular rotational-vibrational descriptors
Extreme (maximum or minimum) values of atomic harmonic force constants
Extreme (maximum or minimum) values of normal mode vibrational frequencies
Extreme (maximum or



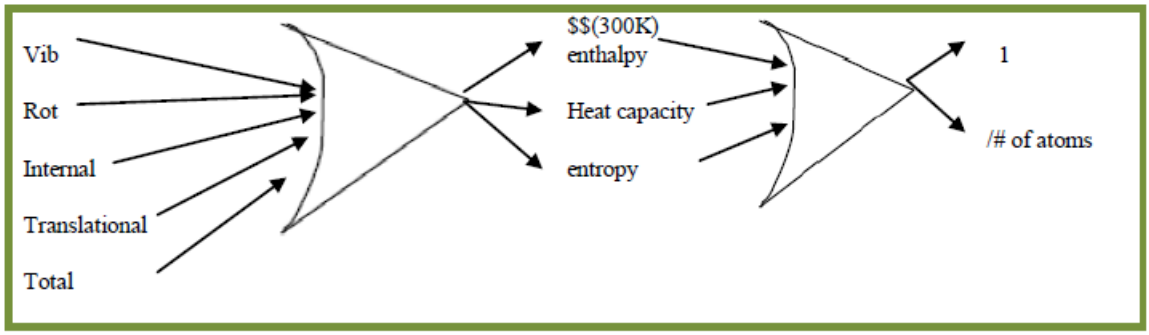
minimum) values of normal mode vibration transition dipoles
Zero-point vibrational (ZPV) energy



ॐ Thermodynamic descriptors

Theoretically calculated thermodynamic descriptors find place in explaining some of the bulk properties like B.P., M.P., solubility and flash point. Translational, vibrational and rotational components of enthalpy, entropy and heat capacity are the major components of thermodynamic descriptors. Normalized and total numerical values have better explainability compared to the component ones. Chart 3.12 lists some of the typical thermodynamic descriptors. Recently, probability based entropy descriptors have been applied by Jurs and Katrizsky.

Chart **3.12: Thermodynamic descriptor**



Orthogonal descriptors

The insignificant (<0.05) correlation between two descriptors is rare. The overlapping information, descriptors calculated from overlapping concepts, congeneric compounds etc. result in correlated descriptor vectors. In order to circumvent the limitations of hard modelling techniques, orthogonalisng the columns of variable descriptor matrix or development of orthogonal descriptors themselves are the alternatives. In PCA, linear combination of variables is used to calculate orthogonal PCs. Linear combination of chemically significant variables does not transparently indicate the individual chemical information and thus PC analysis is criticized. The molecular descriptors from optimized 3D geometry are known as geometric descriptors viz. moment of inertia, shadow index etc. They are also correlated and the same measures of remedy are in use. The basis of the sets of a large number of QC descriptors is the optimized geometry. Again QC descriptors (mostly zero order tensors) are correlated and PCA is in practice.

The well thought dream of creating orthogonal descriptors with chemical significance is realized in a way in 90's. Todeschini proposed orthogonal descriptors basing on finding Eigen values of a modified form of optimized Cartesian coordinate matrix of the molecule. These descriptors, being simple functions of Eigen values, encode size, shape, symmetry and atom distribution in the molecule. GWHIM considering a grid reflects interaction field dimensions. GETAWAY based on hat matrix encodes atomic information, mutual accessibility etc. Weighting with several electrostatic and quantum chemical quantities, WHIM descriptors grew from 11 to more than 40.

ॐ Eigen value descriptors

The Eigen analysis of multivariate (2way or 3-way) explanatory variables or response tensorial data is a sought after approach in Chemometrics (chart 3. 13). This versatile (SVD) method has been used at different stages in descriptor research. The first category is Eigen analysis after calculating descriptors. Distance matrix is calculated from adjacency (connection) 2way-tensor. A weighting scheme (with respect to atomic number, mass, van der waals volume, electro negativity or polarizability) is operated on the distance matrix. Scalar (zero) and first order (topological) descriptors are derived from distance matrix. The sum and sum of absolute Eigen values are one group of descriptors. The other category is from Eigen vector coefficient and comprises of sum, average, Randic and average Randic type. The Eigen vector analysis for geometric and QC descriptors also is beneficial. Alg. 3.7 and Alg. 3.8 describes electronic eigen value descriptor calculation.

Chart 3.13. : Eigen value analysis in chemical research

- Elimination of redundant or linearly dependent variables
- Arrive at hidden roots i.e. indistinguishable species in chemical
- equilibrium or independent reaction paths of dynamic (space or time) systems
- Feasibility studies of hypothetical reaction from reaction matrix

Descriptor	**Electronic Eigen value descriptor (Ele.EV)**
Significance	Mass distribution in molecule
Basis	Shannon information
Formula	$EleEV\ \ x = -\sum_{i=1} \frac{1}{\sigma * \sqrt{2 * \pi}} * exp\left[\frac{-\ x - E_i\ ^2}{2 * \sigma^2} \right.$
Reflects	Combined effect of sub molecular polarity and molecular topography
Expert opinion	Better structural descriptor compared to electronic excess energy on individual atoms

Alg. 3.7 : Electronic Eigen value descriptor (Ele.EV))

Step: Zero — **Input:** Structure of moleccule

Step: One — Choice of CQC [SEMO, CQC, DFT]

Step: Two — Scaling of energies between -45 and 10ev

Step: Three — Gaussian kernel $\sigma = 0.1ev$

Step: Four — Summation of overlaid kernel $\frac{\sigma}{2} = 0.1ev$

Alg. 3.8 : Electronic Eigen value descriptor (Ele.EV)

Step: Zero — Molecule is freely and separately imbedded in centre of grid (XG)

Step: One — Calculate scalar field values (W)

Step: Two — W is the weight for grid point co-ordinates

Radius of gyration

The Eigen values of gyration matrix of the molecule signify the three principle radii (alg. 3.9).

Alg. 3.9 : Radius of gyration

Step: **One** — Locate center of mass

Step: **Two** — Develop tensor for gyration matrix

Step: **Three** — Cal Eigen values

Significance — Eigen values correspond to three principle radii of molecule

\- **Limitation** — Orientation of original molecule in space is random

▶ **Remedy** — Sum of atomic van der Waals volumes (scaled on Carbon atom)

Descriptor	Leading eigenvalue of the distance matrix	Eigen value
Basis	Distance matrix	
Formula	Max(eigenvalue(DistMat))	
Reflects	Good discriminant of increasing size for a series of compounds	

Descriptor	**VAD indices _Balaban**	Eigen value
Basis	Distance matrix	
Formulae	eValues = eigen value (DistMat) VAD1 = Unnique_ negative (eValues)	
	$VAD2 = \frac{VAD1}{A}$;	

Descriptor	**VED indices**	Eigen value
Basis	Distance matrix	
Formulae	[eValues,eVec]= eigen value (DistMat) $VED1=\sum_{i=1}^{A} l_{iA}$ $VED2=\frac{VED1}{A}$ $VED3=\frac{A}{10}*\log(VED1)$	

A	Number of molecular graph vertices
l_{iA}	Loadings (i.e. coefficients) of the eigenvector associated with the largest negative eigenvalue

Descriptor	**VRD indices**	Eigen value
Basis	Randic operator for loadings of l_{iA} $RandOP=\sum_{b}\sqrt{a_i * a_j}_{b}$	
Formulae	$VRD1=\sum_{b} l_{iA} * l_{jA}{}_{b}^{-1/2}$ $VRD2=\frac{VRD1}{A}$ $VRD3=\frac{A}{10}*\log(VRD1)$	

b	All bonds in molecular graph
l_{iA}, l_{jA}	LOIs of two vertices for bond

The negative Eigen values of distance matrix are made use in VAD_Balaban, VED, VRD and leading eigenvalue of the distance matrix descriptors.

Eigen values of CQC Spectra as molecular descriptors

The theoretical IR, NMR, UV-VIS spectra computed from quantum chemical calculations are highly correlated, but reflect structure to a greater extent compared to other simple probes. Turner et al. used as many as 300 and 400 Eigen values as orthogonal descriptors. But, here again the chemical significance could not be explicitly transparent.

The hybrid nature of topological and electrostatic characteristics is achieved by taking quotient of shape (WHIM) descriptor with Polarizability or electro-negativity. Mostly, in descriptor research, weighting in terms of division by a descriptor value of other categories are used to have holistic information of as many types as possible.

ॐ WHIM descriptors

Todeschini put forward a novel set of weighted holistic invariant molecular (WHIM) descriptors viewing the geometry of the molecules in the Eigen framework. The algorithm for computation of WHIM descriptors are given in Alg. 3.10.

Alg. 3.10: WHIM

Phase I — Score matrix of atomic Cartesian coordinate matrix

Input: X, Y, Z co-ordinates of atomic positions for a conformer

Step: Three — W: diagonal weight matrix; Weights are along main diagonal

Step: Four — Weighted center of each coordinate

Step: Five — Centre molecular for each coordinate

Step: Six — Weighted covariance matrix

Step: Seven — Three Eigen values and loading matrix of Eigen vectors

λ_1	λ_2	λ_3

Eight — Score matrix (T) in PC space (Project atom coordinates on to three principal axes

Nine — Weighted statistical parameters along principal axes preserving invariance to roto-translation

Invriance to rotation is guaranteed by uniqueness of PC solution

Ten —
For i = 1 : weighting schemes
  For j = 1: 3
    Calculate descriptors (i,j)
  end j
end i

Descriptor	$$$ directional			weighted by $$
d-directional WHIM index	Size		[1:3]	Unweighted
	Accessibility		[1:3]	Atomic masses
	Symmetry		[1:3]	Atomic van der Waals volumes
	Shape		[1:2]	Atomic Sanderson electronegativities
				Atomic polarizabilities
				Atomic electrotopological states

Unique features of WHIM descriptors

- Statistical
  - Low degress of correlation
- Chemical
  - Roto-translation invariance
- Holistic view of molecules
- \+ Conformer dependent

The scope of information of WHIM descriptors is enhanced by weighting each component by mass, van der Waals volume (geometric), polarizability/electro-negativity (electrostatic). They contain 3D information regarding molecular size, shape, symmetry and atom distribution.

GWHIM descriptors

Todeschini extended the basic framework of WHIM descriptors in Grid-WHIM taking into consideration of interaction scalar fields. In G-WHIM, electrostatic and H-boning interaction energies non-bonding (Lenard-Jones) at each of the grid points is calculated using classical potentials (chart 3. 14). The conformer of a molecule is considered to be placed in a cubic box of large number of grid points. The interaction energy, calculated by conceiving that a probe (H^+, H_2O, or methyl) has moved all the grid points, is used as the weighting factor. The Eigen values of covariance of weighted grid pints gives GWHIM descriptors. The KB for choosing the grid points is given KB. The differences between GWHIM and WHIM are noteworthy. Ideally the information at each grid point in the field interaction should be the same independent of the orientation. It is achievable only for an infinite isotropic dense grid and scalar field calculated should reflect all structural aspects. However, in any theoretical or experimental method probe, mathematical function, grid size and shape play a role which sometimes is misunderstood as the inadequacy of the descriptor itself.

Electrostatic	H^+

Chart 3. 14: Probes to monitor Interaction field

Interaction	Probe
Lennard-Jones non-bonding	Methyl like
Electrostatic	H^+
Hydrogen bond	H_2O

Chart 3.14b: Interaction energy as a function of distance

Non-bonding	r^{-6}	Van der waals surface
Electrostatic	r^{-1}	Large grid
Range of points	0 or r	
	R limit where interaction is	

ॐ GETAWAY (Geometry, Topology and Atom_Weights Assembly)

In 2002, Todeschini proposed a new set of descriptors combining geometry and distance matrix (which is a key topological tuple). The hat matrix is calculated to arrive at a nascent class of GETAWAY (chart 3.15).

Chart 3.15: List of different GETAWAY descriptors

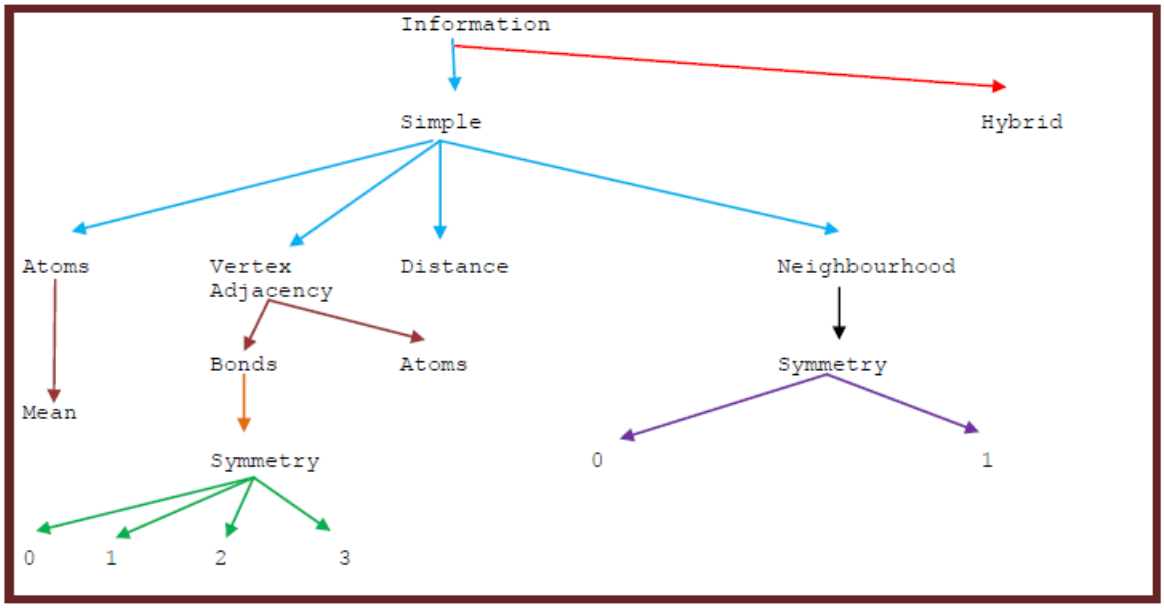
GETAWAY$$$	$$$_ [0:8]
$$$	
H autocorrelation	unweighted
Leverage-weighted autocorrelation	Atomic masses mass
	vdw volume
R autocorrelation	atomic Sanderson electronegativity
R maximal autocorrelation	atomic polarizability
H total index	▶ unweighted ▶ atomic masses vdw volume ▶ atomic Sanderson electronegativity ▶ atomic polarizability
R maximal index	
R total index	
Randic-type R matrix connectivity	
R matrix average row sum	
First eigenvalue of the R matrix	

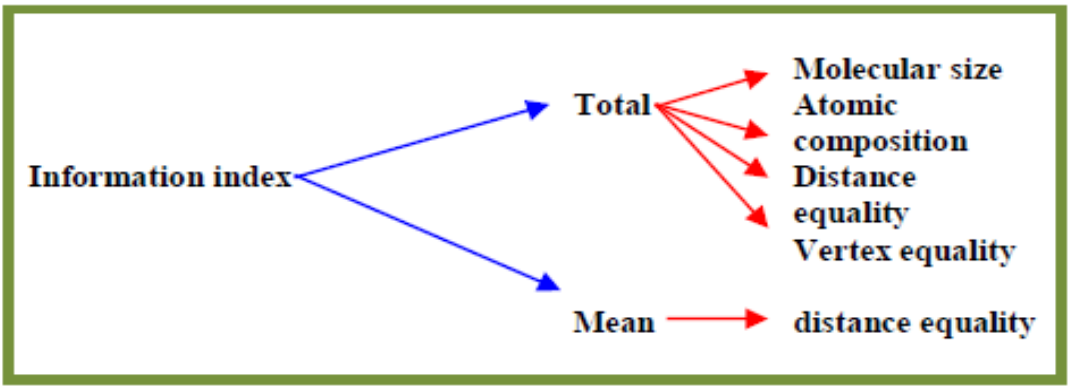
ॐ Information based descriptors

The information theory widely employed in computer science started in 1960s. However, it was imbedded in biology since times immemorial (38 million years). Biochemists, molecular biologists, chemical geneticists nowadays base their arguments within the frame of 3D-structures of conformations in presence of solvent/cell environment. The success of prediction and explanation of several phenomena enthused computational chemists to adopt information and coding theory right from the bonds to the whole molecule.

Shannon information derived descriptors, no doubt, encodes distribution of mass in the molecule but reflects diversity of atoms/bonds. Information based descriptors are also classified as simple and hybrid (chart 3. 16) and illustrates the neighborhood information descriptors of zero and first order symmetry.

Chart 3.16: Classification of information based descriptors





Descriptor	Information index
Significance	Mass distribution in molecule
Basis	Shannon information
Formula	$IC = -\sum_{i=1} \frac{n_v}{n} * \log_2 \frac{n_v}{n}$
Encodes	Size and shape of molecule
Reflects	Diversity of atoms/bonds in molecule
Application	Solvation studies of non-polar solutes

Descriptor	Average Information Content (AIC)
Basis:	Shannon Information Theory
Indicator:	Number of atom with similar connectivity in the molecule
Output:	Arranges molecules in order of rising chain length & No. of substituents of aliphatic alkanes
Formula:	$AIC_{(K)} = -\sum \log_2 \frac{n_i}{n}$ n_i: No. of atoms in ith class n: total number of atoms k: valence shell ; indicates order

```
Desc_Inf =

structural    information order [0:6]
complimentary information order [0:6]
              information order [0:6]
```

```
clean
S1= ['structural  ';
     'complimentary';
     '           ']
 orderz = 'order [0:6]'


 n = 0
 for i = 1:3
    n = n +1;
    Desc_Inf(n,:) = [S1(i,:), '
information ' orderz ]
 end

 k = 0;
 for i = 1:3
    for j = 1: 7
     k = k+1;
      desc(k,:)= [S1(i,:), ' information
order (', int2str(j-1), ')'];
    end
 end
 desc,k
```

Atom total information content	a_IC = n* a
Atom mean Information content	
Vertex adjacency information (equality) n : number of heavy atoms m : number of heavy-heavy bonds	$f=(n^2-m)/n^2$ VadjEq = - (1-f)$\log_2$(1-f)- f$\log_2$f interval (0,1) If f is not in the open Then 0 is returned
Vertex adjacency information	VadjMa =1+$\log_2$m If m is zero Then 0 is returned

Radial centric	information index	
Mean	Mean information content on	distance equality
Mean	Mean information content on	distance magnitude
Mean	Mean information content on	distance degree equality
Mean	Mean information content on	distance degree magnitude
Total	information content on	distance equality
Total	information content on	distance magnitude
Mean	information content vertex	degree equality
Mean	information content vertex	degree magnitude
total	information content on	leverage equality
Standardized	information content on	the leverage equality
mean	information content on	leverage magnitude
geometric	mean on	leverage magnitude

	Information index on molecular size		
Total	information index of	atomic composition	
mean	information index	on atomic composition	
Bond	information content neighbourhood symmetry of 0-order		M = 0 to 5
	Information content index (neighbourhood symmetry of 1-order)		M = 1 to 5
	Total information content index (neighbourhood symmetry of 1-order)		M = 1 to 5

ॐ Hybrid descriptors

The word hybrid refers to combination of two or more terms, basis sets, philosophies or disciplines. The extent of combination (complete amalgamation, loose combination, sequence operation) at component level pushes towards emergence of a new concept for interpretation. Many a times, it is next level of development and sought after over individual components.

Randic index (χ) is based on graph theoretical principles while, Valence (ν) has been proved in its greater explainability of many physico-chemical processes. χ^{υ} is not popularized as a hybrid descriptor, although it combined two different disciplines. Another instance is proposal of modified Balaban index, wherein graph theoretical distance was weighted with physico-chemical quantities like electronegativity, polarizability etc. In molecular descriptor research, hybrid descriptors are calculated using the information from any two or more of the categories viz. topological, geometric and electrostatic/electronic.

▶ Binary hybrid descriptors

Binary, ternary and quaternary combinations of four major unary descriptor categories are in chart 3. 17.

Chart 3. 17: Hybrid systems with different classes of descriptors

Unary	Binary	Ternary	Quaternary
T(opological)	T,G		
	T,E	TGE	
G(eometric)	T,Q		
		TGQ	TGEQ
E(lectrostatic)	G,E		
	G,Q	GEQ	
Q(uaternary)			
	E,Q		

Maximal electrotopological negative variation
Maximal electrotopological positive variation
Molecular electrotopological variation

🔔 Topo-electronic indices

Kalizan proposed two descriptors one for all pairs of atoms and the other for bonded pairs only. The topo-electronic index (Alg. 3.11) is for both connected and disconnected atoms in the molecule. It is an electronic analogy to the valence connectivity index and reflects the electronic structure of the molecule.

Alg. 3.11: Topological electronic index

Step:	**One**	Cal electronic charge distribution of a solute by CNDO/2 MO method
Step:	**Two**	Cartesian co-ordinates for each atom of solute
Step:	**Three**	Excess electronic charge is assigned to each vertex atom

Step:	**Four**	**For** each pair of vertices Find absolute value of excess charge difference Square of inter atomic distance **end**
Step:	**Five**	Summation
Encodes		Charges along with inter atomic distances

🕮 Electro-topological state atom (E-state) indices

In 1990s, Kier and Hall proposed electro-topological state atom (E-state) indices and are equal to numerical values computed for each atom in a molecule. The algorithm of computation is same as that of molecular connectivity index except that partial charges are employed instead of vertex/valence values. The hybrid descriptor is denoted as $ET_{order\ m}$ (order-m, m= 0 to 3). The sum of the electro-topological state values over all atoms in the molecule is represented as a single numerical value. It inspired many applications in SXR over the last quarter century.

Electro-topological descriptor **Hybrid**

$$ET1 = \sum^{N1} \left(\frac{1}{\sqrt{q_i * q_j}} \right)$$

$$ET2 = \sum^{N2} \frac{ABS\ \ q_i * q_j}{r_{ij}^2}$$

$$ET3 = \sum^{N3} \frac{ABS\ \ q_i * q_j}{r_{ij}^2}$$

q_i, q_j	Partial charges on atoms i, j
N1	Set of all bonds including hydrogen
N2	All pairs of atoms both connected and disconnected
N3	All pairs of bonded atoms

Atom based E-state descriptor

$$l_i = \frac{vi * s_i^2}{\sum vi * s_i}$$

$$L_i = \sum l_i$$

$$Xu = \sqrt{n} * \log L$$

$$AI_i = I + L_i$$

L: Valence average topological distance in G valence average length of distance sum s_i

$$\begin{bmatrix} 1 \\ 2 \\ \dots \\ \#row \end{bmatrix} \xrightarrow{Sum} 1$$

E_state $sum\ \ E\_state$

🕮 Topo-Geometric descriptor

Broto and Moreau et al. proposed several sets of binary hybrid descriptors.

🕮 Topo-QC descriptor

The topological descriptor in combination with QC derived charge, polarizability or electro-negativity constitutes first order hybrid descriptor. In this set, there is a scalar descriptor encoding global information of charge as well as topology.

Descriptor	Complimentary information content (zeroth order) + gravitation index
Encodes	Size and shape of molecule

Appllication	Solvation studies of non-polar solutes

🕮 Information content + Gravitational index:

Combining the complimentary information content of zeroth order with gravitation index, a new hybrid descriptor is generated.

Hybrid descriptor	Grav itational Index + CIC-zero
Encodes:	Size & Shape information about molecule
Application:	• Solvation of non-polar solute in water ○ Represents dispersion &cavity formation effects

Descriptor	Complementary Information Content zero order (CIC0)
Encodes	Degree of branching of a hydrocarbon molecule
Formula	$\log_2 n - OIC$ $OIC = -\sum_i \frac{n_i}{n} \log_2 \frac{n_i}{n}$ n_i: no. of atoms in ith class n: total number atoms in the molecule

🕉 Physical_chemical_biological descriptors

Lipinski's alert index along with new drug-like and lead-like indices increase search efficiency in picking for right molecule for a prospective drug. The bulk properties like dielectric constant, refractive index, molar refractivity also form a class of molecular descriptors.

- 🕮 Ghose-Viswanadhan-Wendoloski antineoplastic-like index at 80%
- 🔔 Ghose-Crippen molar refractivity
  - ❖ Antiinflammatory
  - ❖ Antidepressant
  - ❖ Antipsychotic
  - ❖ Antihypertensive
  - ❖ Hypnotic
  - ❖ Antineoplastic-
  - ❖ Antiinfective
- 🔔 Lipinski Alert index
- 🔔 Moriguchi octanol-water partition coeff. (logP)
- 🔔 Squared Moriguchi octanol-water partition coeff. (logP^2)
- 🔔 Ghose-Crippen octanol-water partition coeff. (logP)
- 🔔 Squared Ghose-Crippen octanol-water partition coeff. (logP^2)

Descriptor	**Molar refractivity**
Input	Distance matrix

$$Mol\ \text{Refract} = \frac{\text{Ref}Ind^2 - 1}{\text{Ref}Ind^2 + 2} * .$$

RefInd	Refractive index
Mol Wt	Molecular weight

Inf.Bits_**Molar refractivity**

- ○ Accounts for polarisability of molecule
- ○ Does not vary much

- ❖ Additive model with atomic contributions hold like for logP

Molar Refractivity

From measured refractive index, molar refractivity is calculated.

Correlation of connectivity indices with macroscopic properties

The classical physico chemical properties and connectivity indices or their functions are linearly correlated in the statistical sense (table 3.1), warning modelers to be cautious in using bulk properties along with connectivity indices in hard models like MLR.

▶ Recent Descriptors

Topological electronic index for GLC: Ośmiałowski et al. [136] introduced a new topological electronic index and applied for GLC response variation using total energy and electro-topological descriptors.

Descriptor	Electro-topological index- Ośmiałowski

CNDO/2 Descriptors

🔔 Total energy 🔔 Topological electronic index

For all pairs of atoms
Term1 = Cal absolute differences in electronic excess charges on each atom pairs
Term2 = Cal squares of interatomic distances
Term3 = Divide Term1 by Term2
End
ETI_Osmialowski = sum(Term3)

Molecules	🔔 Aliphatic, heterocyclic amines
Response (Measured)	🔔 GLC
Stationary phase	🔔 Three

Table 3.1: Correlation between macroscopic properties and connectivity indices

Physico chemical properties	χ1s
Polarizability	χ1
Vander waal a, b	χ1, χv
Resonance	1χ-1χv
Charge	χ1
Molecular mass	0χ to 1χ

Hong et al. [53] found Mold derived 2D-molecular descriptors have similar information as those of Cerius, Dragon, and Molconn-Z based on Shannon entropy analysis.

Inf.Bits_Mold

🔔 Shannon entropy analysis demonstrated that Mold^2 descriptors convey a similar amount of information.
\+ low computing cost for Mold^2
\+ Large databases in virtual screening.
\+ High reproducibility and reliability are expected because Mold^{2}does not require 3D structures.
\+ Mold^2 is freely available to the public

Descriptor	Molconn-Z Estate molecular descriptors
Encodes	🕮 topological environment of that atom 🕮 electronic interactions due to all other atoms in the molecule

Semi empirical Topological Descriptor: Zhou et al. [141] proposed a new topological descriptor using SEMO paradigm and found its high predictability of thermodynamic properties of an extensive set of organic molecules (chart 3.18).

Chart 3.18: Aliphatic, heterocyclic amines Electro-topological index- Ośmiałowski

Descriptors
- Path numbers
- P2 and P3
- SEMO_**topological**
- Equilibrium electronegativity

Molecules: Aliphatic, heterocyclic amines
Response (Measured): GLC
Stationary phase: Three

Descriptor: Equilibrium electronegativity
- Basis: SEMO ; Topological concept
- Cal Group electronegativities (A) taking into consideration of chemical environment
- Cal relative bond lengths(B), equilibrium electronegativity (C)
- SEMO_TOPOL_Desc = fn(A,B,C)

SThermPropR
- Organic compounds
correlation coefficient : high

MEDV-13 descriptor: A new set of 91 descriptors called 'Molecular electronegativity distance vector' based on 13 atomic types predicts biological activities (chart 3.19). It includes relative electro-negativity of non-hydrogen atoms and a topological distance for relative distance. Some of the elements C, H, O, N, S, Cl, Br, I and P are considered.

Chart 3.19: SBindingR — SActR

Descriptors	MEDV-13
Mod_E-state index	
Molecular electronegativity distance vector based on 13 atomic types	

Data set	571 *Compounds*	r^2	Standard error
Training	80%	0.852	0.469 log units
Test set	20%		

Response (Measured):
- Corticosteroid-binding globulin (CBG) binding affinity of the steroids
- Activity inhibiting angiotensin-converting enzyme (ACE) of dipeptides

Task: Prediction

Models	
PCR	
PLS	Linear
	Non-linear
Continuum Regression	

Inf.Bits.
2D-Topological MEDV-13 model = 3D-CoMFA

Atomic property weighted radial distribution function (AP-RDF) descriptor: Fernandez et al. [94] proposed the descriptor for large scale SPropR of gas absorption. The PC transformation of AP-RDF descriptors has good discriminating properties.

Count based molecular structure descriptor based on the Distribution of Molecular features: Baumann introduced a descriptor based on histograms of types of atoms and bonds in the molecule (chart 3.20). This descriptor was weighted by topological distance counts with geometric distance.

Chart 3.20: Count based molecular structure descriptor

- Encodes
  - Distribution of atoms and bonds types

\+ It is invariant to both translation and rotation.
\+ Computationally easier.
\+ Descriptor >> CoMFA orEVA descriptor

it does not require the alignment of the structures under study.

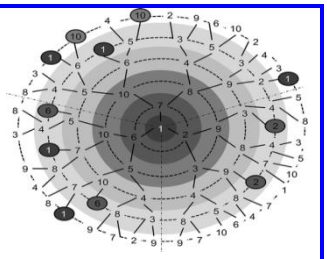
Extended Count based descriptor

- Encodes
  - Geometric features

Descriptors based on orthogonality measure: Nowik et al. [104] introduced a new measure of orthogonality from distances between nearest neighbors (NND) in the apex plot. New descriptors calculated from NND values reflect peak distribution (arithmetic mean) and homogeneity (geometric/ harmonic mean). This new descriptor is compared with fractal dimensionality. A stereo-signature descriptor based on their earlier signature_moleuclar_descriptor reported.

Electronegativity based descriptors: Liu et al. [72] found a new set of descriptors based on electronegativity are equivalent to 3D-based CoMFA.

Descriptor	Stereo Signature Molecular Descriptor
Basis	Signature molecular descriptor
Input	Stereochemistry of molecule
Output	Descriptor compliant with Cahn–Ingold–Prelog priority rules Augmented to a chiral molecular graph



4. Software

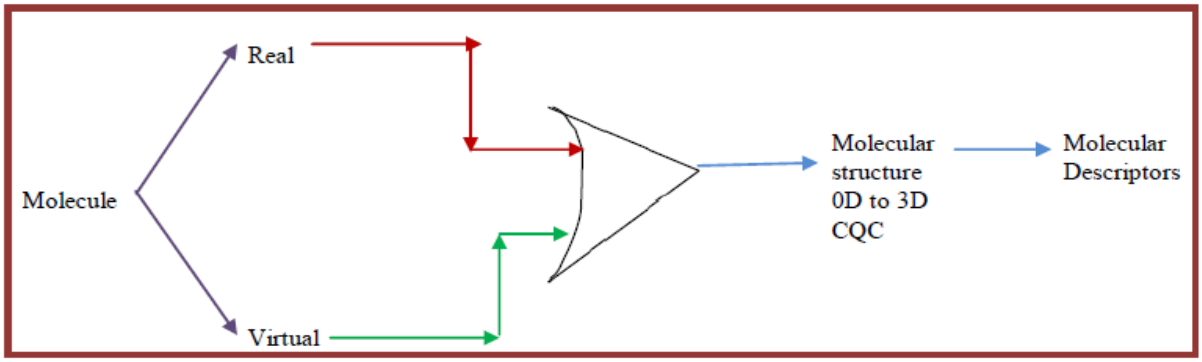
Liu et al. [84] developed PowerMV, a tool with target researches by biologists and statisticians. It has modules for visual appreciation of molecules, descriptor calculation, model development and arriving at optimum best set of models along with evaluation tools of high priority hits. The pros and cons of SXR even with 3D- optimized geometry based descriptors are dilated. The software Codessa Pro (Comprehensive Descriptors for Structural and Statistical Analysis) by Katrzsky eta al. is one of sought after packages for SXR. A perusal of typical software packages, types of descriptors are described in chart 4.1. The preliminary perspectives of molecules in nature, real/virtual molecules and modeling/inverse models are briefed in chart 4.2, Fig. 4.1 and Fig. 4.2.

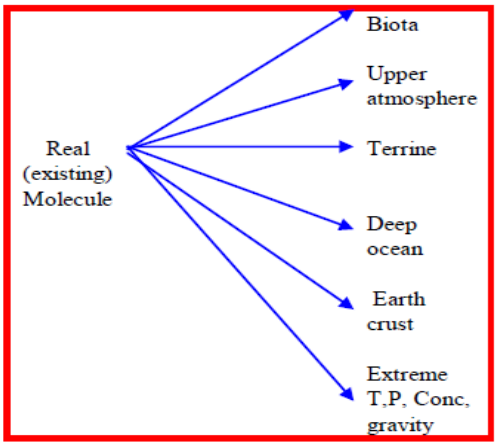
Chart 4.1: Typical packages of molecular descriptors

Package	Type	#MolDes
DRAGON	Constitutional, topological, 2D-autocorrelations, geometrical, WHIM, GETAWAY, RDF, functional groups, properties, 2D binary and 2D frequency fingerprints, etc.	4885
CODESSA	Constitutional, topological, geometrical, charge-related, semi-empirical, thermodynamical	1500

ADRIANA	Global physicochemical descriptors, atom property-weighted 2D- and 3D-autocorrelations, RDF, surface property-weighted autocorrelations	1244
Sarchitect	Constitutional, 2D and 3D descriptors	1084
PowerMV	Constitutional, atom pairs, fingerprints, BCUT, etc.	1000
PreADMET	Constitutional, topological, geometrical, physicochemical, etc.	955
PaDEL-Descriptor	Constitutional, WHIM, topological, fingerprints 729 1D, 2D descriptors and 134 3D descriptors	863
MOLGEN-QSPR	Constitutional, topological, geometrical, etc	707
MOE	Topological, physical properties, structural keys, etc.	300
ADMET	Constitutional, functional group counts, topological, E-state, Moriguchi descriptors, Meylan flags, molecular patterns, electronic properties, 3D descriptors, hydrogen bonding, acid-base ionization, empirical estimates of quantum descriptors. 266 2D and 31 3D	297
ADAPT	Topological, geometrical, electronic, physicochemical.	260
MARVIN Beans	Topological, geometrical, fingerprints, physico-chemical, etc.	
JOELib	Counting, topological, geometrical, properties, etc.	40
MOLCONN-Z	topological	40

Chart 4.2: Real/virtual molecules, SXR (X-y) and inverse models





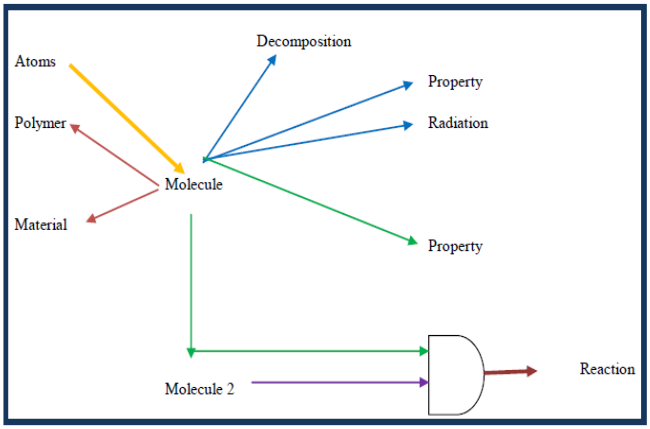


Chart 4.2b: Soft and hard model strategies

- *Ordinary Least Squares regression (OLS)*

Soft regression

- Principal Component Regression (PCR)
- Partial Least Squares regression (PLS)
- Ridge Regression (RR)
- Continuum Regression (CR)
- Step Wise Regression (SWR)

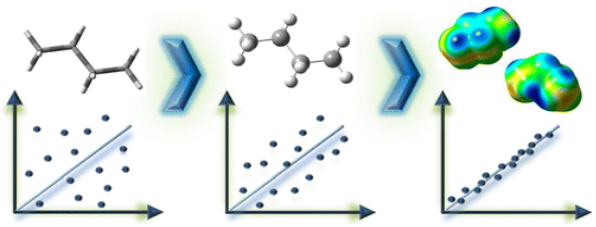
Non-linear methods

- Nonlinear least squares regression
- Polynomial regression
- Nonlinear PLS method
- Alternating Conditional Expectations(ACE)
- SMART
- MARS
- NN

Molecular structure

- CORINA
- Molecular mechanics
- SEMO
- Hybrid functionals
- Hybrid paradigms + Empirical correction for dispersion, van der Waal interactions

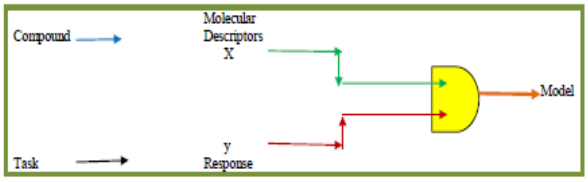
M. Hechinger, K. Leonhard, and W. Marquardt — *J. Chem. Inf. Model.*, **2012**, *52* (8), 1984–1993

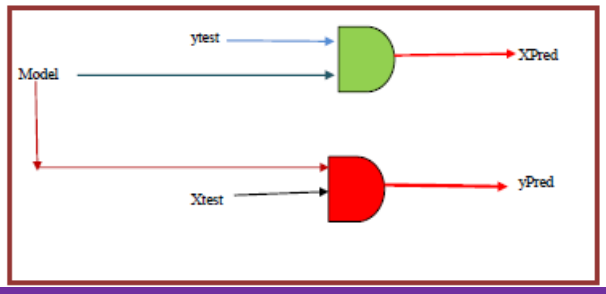


What is Wrong with Quantitative Structure–Property Relations Models Based on Three-Dimensional Descriptors?

Fig. 4.1 Improvement in SXR through multi-dimensional structure of molecules

Fig. 4.2: Model and inverse model





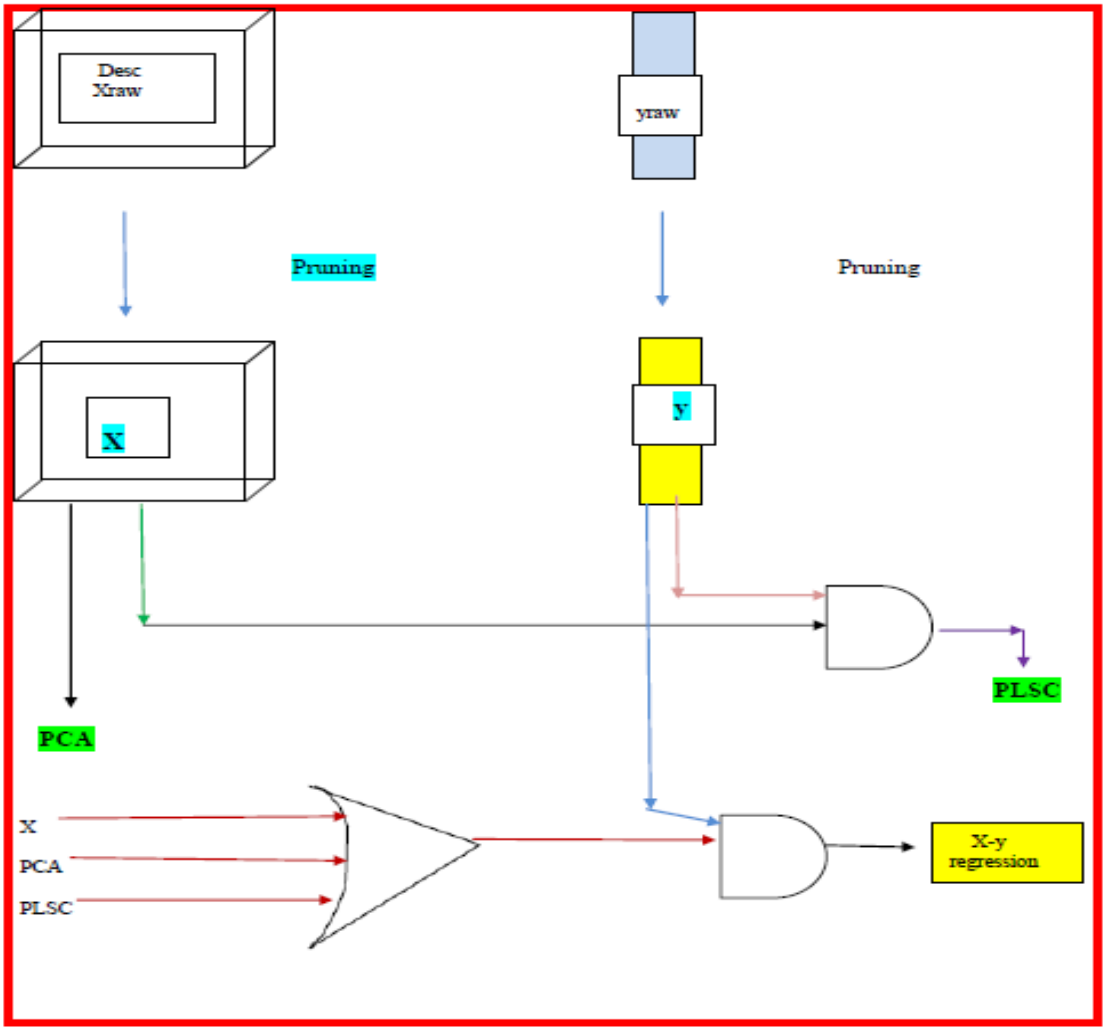


Fig 4.2b: A bird's eye view of SXR models

5. State-of-knowledge of molecular descriptors

Yester years' beliefs/myths: Due to the sizes (number of atoms) of molecules and CPU time for CQC, lower level geometric optimizations were in use for the follow up descriptor calculation. The later jargon of variable selection, modeling method, validation and best model selection are from standard practices in other disciplines.

Today' reality

Now graph-theoretical, quantum chemical, electrostatic, spectroscopic, Eigen approaches are separate streams with a consequence of exponential growth of descriptors of chemical moieties. Molecular structure describes constituent atoms and connections between them from chemical bond stand point. The count of each atom, bonds (double, triple, partial) and chosen fragments form a numerical representation of a molecule in terms of simple molecular descriptors. It does not involve any calculation and thus algorithm. The next stage is to invoke structure as a 2D-molecular graph and the corresponding connectivity matrix is binary matrix of zeros and ones. Various scalar descriptors are calculated by simple matrix operations and simple number crunching. This also does not require any intricate chemical knowledge and thus the descriptors also do not contain all information for diverse structures. The 3D-structure of molecule derived from X-ray, NMR or CQC is a quantum leap, but with higher information content. The first and foremost set is geometric descriptors with straight algebra resulting in shape and size based knowledge.

The binary and ternary hybridization of Topo, geo, electrostatic, QC categories brought forth more explainability/correlation/explaining variance of response of diverse sets of compounds. The number (statistical) correlation between sets of descriptors for given molecules is a stumble block in hard-regression. Although, orthogonolization (PCA, PLSA) surmounts this hurdle, the chemical space is now mapped on to non-orthogonal mathematical space. Here, the axes of mathematical space are linear combination of chemical space. But, chemical interpretability is lost. The fruits of search for orthogonal descriptor research are WHIM by Todeshini.

It is now a common sense that high quality geometry optimization methods are indispensable in descriptor based models research. Further, 3D-molecular descriptors are to be averaged for different conformational structures to be nearer to real time interactions of bio-/pharmaco-interactions.

Rational designs in drug leads or electronic/industrial material templates are target oriented research pursuits, ideally involving biologists, material scientists, chemists, and information/ omni-metric scientists. In the recent past, chemical genetics and high content screening depend upon screening in cells or animals. The limitation is that the biological target is mostly unknown. Software and hardware are fast life enabling tools. But, in practice, sub goals are investigated by technocrats with a little knowledge of many of the said fields. The consequence is like searching reliable bits from an ocean bed of admixture of data/information/knowledge. The simile goes as a costly venture of high tech search for a needle in a hay stack. Thus, the modelling and prediction with molecular descriptors is a multi-disciplinary research area especially for critical issues. Chemometricians are comfortable in trying with advanced mathematical and statistical tools for optimization/ classification. Quantum chemists trained in CQC pickup appropriate ab-initio or DFT procedure for 3D-structure optimization and chemically significant quantum descriptors. Mathematical chemists, with expertise in graph theoretical principles, look at molecules from pure connectionist point of view. But, later the descriptors are hybridized with quantum chemical, electrostatic, topological and geometric perceptive. Traditional synthetic organic chemists toil to produce larger sets newer analogues of a chosen moiety. Combinatorial chemists work hard to bring large number of chemicals in one pot synthesis and so on. Computational chemists work out virtual libraries, HTS and SXRs to short list sets of moieties without performing conventional experiments. Each of these approaches has pros and cons, yet yield unique, complementary, supplementary, redundant, conflicting information. From of a pool of these, a best model, information is to be sought. But, in routine research a trodden path is followed to bring out presentable/publishable results in a tiny way. Coveted research/commercial centres perform the jargon in a mega way in an effort to bring out most trust worthy outcome. At a higher level, the confidence on each of these paradigms and standard protocols available restrict the final say.

Rules for leads of fungicides: Liu et al. [39] proposed rules for leads of fungicides, insecticides, and herbicides (chart 5.1) as Lipinski-Rule-of-5 was unsuitable, although it was widely employed in drug discovery efforts. Around 400 descriptors were calculated by CODESSA and screening test was based on statistics (chart 5.2).

Chart 5.1 Rules for leads of fungicides, insecticides, herbicides

$$\log P = \log\left(\frac{Solute_{Octanol}}{Solute_{Water}}\right)$$

- Number of hydrogen bond donors
- Polar surface area
- Carcinogenic toxicity
- Mutagenic toxicity

Chart 5.2: Statistics

- ANOVA
- Kolmogorov–Smirnov test
- Kruskal–Wallis test
- Pearson product-moment Correlation

The definition base (DEF.MolDes) contains object oriented strings amenable for intelligent computer augmented instruction (ICAI). The necessary conditions, formulae, limitations and remedial measures of soft (LMS, PCR, PLSR) and hard (LS, MLR) modelling approaches are available. The knowledge base is in the form of IF Then rules (Horn clauses of first order predicate logic) is the heart of the knowledge based expert system. The use and misuse of supervised classification, clustering and discrimination (LDA, QDA, NN) with different measures of similarity, diversity, residual statistics and error distribution assumptions are coded as knowledge bits.

6. Exploratory super-highway for future MolDes research

The future course of research in better physico-chemical theories and new line of descriptors especially for solid/semi-solid/ biological cell states of molecular systems along with higher level of accuracy of experimental probes [66] (chart 6.1) improve this predictive science.

Fourth wave: In the coming years, new descriptors evolve with a reduced degeneracy, high invariance, maximum information, less correlation and increased discrimination to probe into real life point blank tasks. The base is however simple molecular descriptors derived by counting some atom-types or structural fragments in the molecule, others calculated from algorithms applied to a topological 2D-molecular graph and descriptors derived from a 3D-optimized geometry with follow up involved computations. The, less degenerative, , process oriented descriptors will evolve with better perspective.

Chart 6.1: SPMeltPointR — SPropR

Descriptors	DRAGON
E-state	Sub-structural molecular fragments
Descriptors from Counts of atoms determined for E-state atom types	

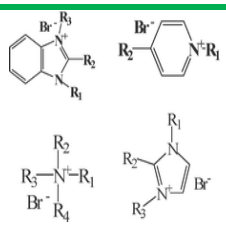
Machine Learning models	
MLR	Associative NN
MLP_BP	RBF_NN
SVR	Modified PLSR
K_NN	

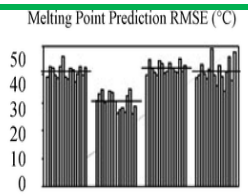
Molecules (ionic liquids)
717: Bromides nitrogen-containing organic cations

Pyridinium bromides	126
Imidazolium /benzoimidazolium bromides	384
Quaternary ammonium bromides	207

Response (Measured)	Melting point
Task	Prediction

Software	ISIDA, NASAWIN VCCLAB
Model_Statistics	RMSE 5-fold-CV





If Two MolDesV do not duplicate chemical
Then Information Vectors are orthogonal & MLR results in unbiased results with good predictive capability

If Two MolDesV have duplicate information
Then Two vectors are collinearity & MLR fails

If Topological index is same for two structures
Then graph invariant

If Isomorphic graphs
Then graph invariant

If isomorphic compounds
Then isomorphic graphs
MolDesV : Two column vectors of molecular descriptors for a series of compounds

Inf.Bits.

Predictive models

- Non-linear [SVM ; Associative NN; MLP_BP] models are superior
  - Moderate predictive quality of models
    - Low quality of experimental data
    - Poor representation of structural features in ILs in solid state due to glass formation, polymorphic effects, eutectics etc.

Acknowledgements

We admire Dr. Anantha Ramam, committed to probe into cause of visible/invisible phenomena one experiences by searching for truth through chemistry. Prof Ramam was deeply immersed in final scrutiny of manuscripts on neural networks up to ten days before of his last breath.

We thank Dr M Suryanarayana, scientist (DRDE, Gwalior), alumni of Andhra University for a startup session for the project 'Molecular structural information for nerve agents'. RSR lectured on 'Impact of Chemometrics in structure response relationships' on invitation of Rama Chandran director of DRDE. RSR expresses his gratitude to Dr Swamy, director general for his keen interest in high tech applications in organic chemistry. During period 1999 to 2001, a technical report of various categories of molecular descriptors was prepared after experiencing structure-response relationships using of CODESSA software. This is the core of third section of the review. The encouragement of editors (Dr Somasekahara Rao and Prof Suresh) to cross a little of conventional boundaries in documenting references in our earlier publications brought a positive feedback to pick up sets of sought after literature on subjects prime concern. It also favors to browse through breadthwise results to find gaps for further research at a glance.

I was fond of explicitly expressing the assumptions in any theory or derivation. I tried to write down for thermodynamics like that on the left side page of my notes in M Sc. I was selected for an officer post in BARC in 1972 just after three months of joining research, but remained in Andhra University as a research scholar and faculty since 1974, of course, up to my retirement in 2008. My career was along the stream and never an upstream swim. I started computer programming on IBM 1130 with ten days reading of FORTRAN book by Ratnam and Raju. Myself along with Satyanarayana, a research scholar learned statement by statement by writing programs of course for calculations in complex equilibria. Those formulae and solution of equations were already done with FX8 calculator (which does not have even logarithm or exponential functions) during 1974 to 1976. With, in a couple of years, we started running the

then standard programs like SCOGS, POT3 and MINIQUD-74. With satisfactory results for chemical tasks oriented to Ph.D. programs, MINIQUAD series (MINIQUAD 75, SUPERQUAD and HYPERQUAD) was continued in our school for the last three and half decades. During the last two and half decades the high computing power, multi-dimensional graphics, data structures, and object oriented patterns of MATLAB have been extensively applied in research programs, apart from DBASE III+ and TurboProlog software.

The author's (RSR) first chemical education article with Bapanaiah, a co-research scholar was published four decades ago in Science Reporter (12(1975)315-318). The impetus was the language barrier between statisticians and chemists in that decade at least here. Although now trivial, the first question statisticians use to ask was 'Is your data adhere to normal distribution of errors?' The chemist uses to say ' My data is accurate as we have taken care at all stages of experiment. Our desire is to apply statistical analysis to obtain more accurate result and eliminate errors unknown. Neither the chemist nor statistician could understand the other's mind (or science in it). With the help of four figure logarithmic tables or calculating machine (which does not give decimal point finally), the averages to coefficient of variance were calculated and both groups call it day with utmost satisfaction and contentment. Again for a query of statistician, 'is there data for repeated experiments', and the reply of chemist were replicating measurements of absorbance readings in spectrophotometric analysis were available. The implicit fun is the experiments like preparation solutions for measurements were not repeated, but only with solution, the measurements were made six or eight times. Naturally the readings were always within the readability of instruments. Thus, it is not standard deviation of chemical experiment, but only repeatability of instruments within a matter of five minutes. These lone stories or many others do not have long shelf life with the progress of one's own experiences, unless one takes care of probing into details of interdisciplinary research. The outliers, robust statistics, distribution free statistics, fuzzy/rough sets, possibility, robotic-experimental operations, hyphenated (second and third order) instruments, were all later add-ons in chemistry, mostly under Chemometrics, matured discipline. Its off-spring and cousins are -mics, -omics, and/or -metrics are tertiary or quaternary cross-disciplinary research areas, challenging even an expert in one or two disciplines to feel and have vision when closed his eyes.

REFERENCES

Monographs in Molecular descriptors

[1] R. Todeschini, Handbook Of Molecular Descriptors, Wiley-VCH, New York, **2000.**
[2] P. Duchowicz, E. Castro, The Order Theory in QSPR - QSAR Studies, ISBN: 978-86-81829-90-5, **2008.**
[3] R. Todeschini, Handbook Of Molecular Descriptors, **2008.**
[4] R. Todeschini, Molecular Descriptors For Chemoinformatics, John Wiley & Sons, **2009.**
[5] R. Todeschini, Molecular Descriptors For Chemoinformatics, Wiley-VCH Verlag GmbH, **2009.**
[6] I. Gutman (Editor Boris Furtula), Distance in Molecular Graph - Applications; Distance in Molecular Graph - Theory, **2012.**
[7] M. Dehmer, Statistical Modelling Of Molecular Descriptors, John Wiley & Sons, **2012.**

Bcut descriptors

[8] M. P. González, C. Terán, M. Teijeira, P. Besada, M. J. González-Moa**,** *Bioorganic & Medicinal Chemistry Letters*, **2005,** 15, 15, 3491-3495.
BCUT descriptors to predicting affinity toward A3 adenosine receptors
[9] J.S Mason, B. R Beno, *J. Molecular Graphics and Modelling*, **2000,** 18, 4–5, 438-451.
Library design using BCUT chemistry-space descriptors and multiple four-point pharmacophore fingerprints: simultaneous optimization and structure-based diversity

[10] Ijjaali, F. Petitet, E. Dubus, O. Barberan, R. Michel, *Bioorganic & Medicinal Chemistry,* **2007,** 15, 12, 4256-4264.
Assessing potency of c-Jun N-terminal kinase 3 (JNK3) inhibitors using 2D molecular descriptors and binary QSAR methodology

[11] H. Kim, Y. S. Cho, H. Y. Koh, J. Y. Kong, K. T. No, A. N. Pae, *Bioorganic & Medicinal Chemistry,* **2006,** 14, 5, 1454-1461.
Classification of dopamine antagonists using functional feature hypothesis and topological descriptors

[12] M. P. González, P. L. Suárez, Y. Fall, G. Gómez, *Bioorganic & Medicinal Chemistry Letters,* **2005,** 15, 23, 5165-5169.
Quantitative structure–activity relationship studies of vitamin D receptor affinity for analogues of 1α,25-dihydroxyvitamin D3. 1: WHIM descriptors

[13] L. Saíz-Urra, M. Pérez González, M. Teijeira, *Bioorganic & Medicinal Chemistry,* **2007,** 15, 10, 3565-3571.
2D-autocorrelation descriptors for predicting cytotoxicity of naphthoquinone ester derivatives against oral human epidermoid carcinoma

[14] M. Pérez González, C. Terán, M. Teijeira, A. Morales Helguera, *European J. Medicinal Chemistry,* **2006,** 41, 1, 56-62.
Radial distribution function descriptors: an alternative for predicting A2 A adenosine receptors agonists

[15] L. Saíz-Urra, M. Pérez González, M. Teijeira, *Bioorganic & Medicinal Chemistry,* **2006,** 14, 21, 7347-7358.
QSAR studies about cytotoxicity of benzophenazines with dual inhibition toward both topoisomerases I and II: 3D-MoRSE descriptors and statistical considerations about variable selection

[16] S. Sharma, B. K. Sharma, S. K. Sharma, P. Singh, Y. S. Prabhakar, *European J. Medicinal Chemistry,* **2009,** 44, 4, 1377-1382.
Topological descriptors in modeling the agonistic activity of human A3 adenosine receptor ligands: The derivatives of 2-chloro-N6-substituted-4′-thioadenosine-5′-uronamide

[17] M. G Ford, W. R Pitt, D. C Whitley, *J. Molecular Graphics and Modelling,* **2004,** 22, 6, 467-472.
Selecting compounds for focused screening using linear discriminant analysis and artificial neural networks

[18] M. Rybka, G. Mercader, E. A. Castro, *Chemometrics and Intelligent Laboratory Systems,* **2014,** 132, 18-29.
Predictive QSAR study of chalcone derivatives cytotoxicity activity against HT-29 human colon adenocarcinoma cell lines

[19] M. P. González, M. del Carmen, T. Moldes, *Bioorganic & Medicinal Chemistry,* **2004,** 12, 11, 2985-2993.
A TOPS-MODE approach to predict affinity for A1 adenosine receptors. 2-(Arylamino) adenosine analogues

[20] M. P. González, C. Terán, Y. Fall, M. Teijeira, P. Besada, *Bioorganic & Medicinal Chemistry,* **2005,** 13, 3, 601-608.
A radial distribution function approach to predict A2B agonist effect of adenosine analogues

[21] S. Sharma, Y. S. Prabhakar, P. Singh, B. Kishore Sharma, *European J. Medicinal Chemistry*, **2008,** 43, 11, 2354-2360.
QSAR study about ATP-sensitive potassium channel activation of cromakalim analogues using CP- MLR approach

[22] M. Pérez González, M. Carmen, T. Moldes, *Bulletin of Mathematical Biology*, **2004,** 66, 4, 907-920.
QSAR study of N6-(substituted-phenylcarbamoyl) adenosine-5′-uronamides as agonist for A1 adenosine receptors

[23] M. Pérez González, A. M. Helguera, H.G. Díaz, *Polymer*, **2004,** 45, 6, 2073-2079.
A TOPS-MODE approach to predict permeability coefficients

[24] M. P. González, M. del Carmen, T. Moldes, *Bioorganic & Medicinal Chemistry Letters*, **2004,** 14, 12, 3077-3079.
A TOPS-MODE approach to predict adenosine kinase inhibition

CODESSA package

[25] M. Randić, M. Pompe, *J.Chem. Inf. Comput. Sci.*, **2001,** 41, (3), 631–638.
The Variable Connectivity Index 1χf versus the Traditional Molecular Descriptors: A Comparative Study of 1χf against Descriptors of CODESSA

[26] A. R. Katritzky, I. B. Stoyanova-Slavova, K. Tämm, T. Tamm, M. Karelson, *J.Phys. Chem. A*, **2011,** 115, (15), 3475–3479.
Application of the QSPR Approach to the Boiling Points of Azeotropes

[27] A. R. Katritzky, D. A. Dobchev, D. C. Fara, E. Hür, K. Tämm, L. Kurunczi, M. Karelson, A. Varnek, V. P. Solov'ev, *J.Med. Chem.*, **2006,** 49, (11), 3305–3314.
Skin Permeation Rate as a Function of Chemical Structure

[28] A. R. Katritzky, S. Perumal, R. Petrukhin, *J.Org. Chem.*, **2001,** 66, (11), 4036–4040.
A QSRR Treatment of Solvent Effects on the Decarboxylation of 6-Nitrobenzisoxazole-3-carboxylates Employing Molecular Descriptors

[29] A. R. Katritzky, D. C. Fara, H. Yang, M. Karelson, T. Suzuki, V. P. Solovev, A. Varnek, *J.Chem. Inf. Comput. Sci.*, **2004,** 44, (2), 529–541.
Quantitative Structure−Property Relationship Modeling of β-Cyclodextrin Complexation Free Energies

[30] A. R. Katritzky, A. A. Oliferenko, P. V. Oliferenko, R. Petrukhin, D. B. Tatham, U. M. Lomaka, W. E. Acree, Jr., *J.Chem. Inf. Comput. Sci.*, **2003,** 43, (6), 1806–1814.
A General Treatment of Solubility. 2. QSPR Prediction of Free Energies of Solvation of Specified Solutes in Ranges of Solvents

[31] A. R. Katritzky, D. B. Tatham, U. Maran, *J.Chem. Inf. Comput. Sci.*, **2001,** 41, (5), 1162–1176.
Theoretical Descriptors for the Correlation of Aquatic Toxicity of Environmental Pollutants by Quantitative Structure-Toxicity Relationships

[32] A. R. Katritzky, R. Lomaka, R. Petrukhin, R. Jain, M. Karelson, A. E. Visser, R. D. Rogers, *J.Chem. Inf. Comput. Sci.*, **2002,** 42, (1), 71–74.
QSPR Correlation of the Melting Point for Pyridinium Bromides, Potential Ionic Liquids

[33] R. Hiob, M. Karelson, *J.Chem. Inf. Comput. Sci.*, **2000,** 40, (4), 1062–1071.
Quantitative Relationship between Rate Constants of the Gas-Phase Homolysis of C−X Bonds and Molecular Descriptors

[34] H. Liu, X. Yao, R. Zhang, M. Liu, Z. Hu, B. Fan, *J.Phys. Chem. B*, **2005,** 109, (43), 20565–20571.

Accurate Quantitative Structure−Property Relationship Model To Predict the Solubility of C60 in Various Solvents Based on a Novel Approach Using a Least-Squares Support Vector Machine

[35] A. Varnek, D. Fourches, V. P. Solov'e vV. E. Baulin, A. N. Turanov, V. K. Karashev, D. FaraA. R. Katritzky, *J.Chem. Inf. Comput. Sci.,* **2004,** 44, (4), 1365–1382.
"In Silico" Design of New Uranyl Extractants Based on Phosphoryl-Containing Podands: QSPR Studies, Generation and Screening of Virtual Combinatorial Library, and Experimental Tests

[36] K. Tämm, D. C. Fara, A. R. Katritzky, P. Burk, M. Karelson, *J.Phys. Chem. A,* **2004,** 108, (21), 4812–4818.
A Quantitative Structure−Property Relationship Study of Lithium Cation Basicities

[37] A. R. Katritzky, A. A. Oliferenko, P. V. Oliferenko, R. Petrukhin, D. B. Tatham, U. M. Lomaka, W. E. Acree, Jr., *J.Chem. Inf. Comput. Sci.,* **2003,** 43, (6), 1794–1805.
A General Treatment of Solubility. 1. The QSPR Correlation of Solvation Free Energies of Single Solutes in Series of Solvents

[38] A. R. Katritzky, S. Perumal, R. Petrukhin, E. Kleinpeter, *J.Chem. Inf. Comput. Sci.*, **2001,** 41, (3), 569–574.
CODESSA-Based Theoretical QSPR Model for Hydantoin HPLC-RT Lipophilicities

[39] B. Liu, F. Zhu, Y. Huang, Y. Wang, F. Yu, B. FanJianhua Yao, *J.Agric. Food Chem.*, **2010,** 58, (5), 2673–2684.
Screening Rules for Leads of Fungicides, Herbicides, and Insecticides

[40] A. R. Katritzky, L. Pacureanu, D. Dobchev, M. Karelson, *J.Chem. Inf. Model.*, **2007,** 47, (3), 782–793.
QSPR Study of Critical Micelle Concentration of Anionic Surfactants Using Computational Molecular Descriptors

[41] A. R. Katritzky, R. Jain, R. Lomaka, R. Petrukhin, M. Karelson, A. E. Visser, R. D. Rogers, *J.Chem. Inf. Comput. Sci.*, **2002,** 42, (2), 225–231. Correlation of the Melting Points of Potential Ionic Liquids (Imidazolium Bromides and Benzimidazolium Bromides) Using theCODESSA Program

[42] M. Oblak, M. Ric, T. Solmajer, *J.Chem. Inf. Comput. Sci.*, **2000,** 40, (4), 994–1001.
Quantitative Structure−Activity Relationship of Flavonoid Analogues. 3. Inhibition of p56lck Protein Tyrosine Kinase

Dragon software

[43] B. Bordás, I. Bélai, T. Kőmíves**,** *J.Agric. Food Chem.*, **2011,** *59*, (7), 2863–2869.
Theoretical Molecular Descriptors Relevant to the Uptake of Persistent Organic Pollutants from Soil by Zucchini. A QSAR Study

[44] A. Rizzi, A. Fioni**,** *J.Chem. Inf. Model.*, **2008,** 48, (8), 1686–1692.
Virtual Screening Using PLS Discriminant Analysis and ROC Curve Approach: An Application Study on PDE4 Inhibitors

[45] B. Robson**,** *J.Proteome Res.*, **2004,** 3, (6), 1113–1119.
The Dragon on the Gold: Myths and Realities for Data Mining in Biomedicine and Biotechnology Using Digital and Molecular Libraries

[46] F. Pereira, D. A. R. S. Latino, J.Aires-de-Sousa**,** *J.Org. Chem.*, 2011, *76*, (22), 9312–9319.
Estimation of Mayr Electrophilicity with a Quantitative Structure–Property Relationship Approach Using Empirical and DFT Descriptors

[47] P. L. Andersson, U. Maran, D. Fara, M. Karelson, J.L. M. Hermens, *J.Chem. Inf. Comput. Sci.*, **2002,** 42, (6), 1450–1459**.**
General and Class Specific Models for Prediction of Soil Sorption Using Various Physicochemical Descriptors

[48] R. Lowe, R. C. Glen, J.B. O. Mitchell, *Mol. Pharmaceutics*, **2010,** 7, (5), 1708–1714.
Predicting Phospholipidosis Using Machine Learning

[49] G. Pasquale, G. P. Romanelli, J.C. Autino, J.García, E. V. Ortiz, P. R. Duchowicz, *J.Agric. Food Chem.*, **2012,** 60, (2), 692–697.
Quantitative Structure–Activity Relationships of Mosquito Larvicidal Chalcone Derivatives

[50] M. Goodarzi, M. P. Freitas Richard Jensen, *J.Chem. Inf. Model.*, **2009,** 49, (4), 824–832.
Feature Selection and Linear/Nonlinear Regression Methods for the Accurate Prediction of Glycogen Synthase Kinase-3β Inhibitory Activities

[51] R. Solimeo, J. Zhang, M. Kim, A. Sedykh, H. Zhu, *Chem. Res. Toxicol.*, **2012,** 25, (12), 2763–2769.
Predicting Chemical Ocular Toxicity Using a Combinatorial QSAR Approach

[52] S. Novotarskyi, I. Sushko, R. Körner, A. Kumar Pey, I. V. Tetko, *J.Chem. Inf. Model.*, **2011**, 51, (6), 1271–1280.
A comparison of different QSAR approaches to modeling CYP450 1A2 inhibition

[53] H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. P. Tong, *J.Chem. Inf. Model.*, **2008,** 48, (7), 1337–1344.
Mold^2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics

[54] X. S. Wang, H. Tang, A. Golbraikh, A. Tropsha, *J.Chem. Inf. Model.*, **2008**, *48,* (5), 997–1013.
Combinatorial QSAR Modeling of Specificity and Subtype Selectivity of Ligands Binding to Serotonin Receptors 5HT1E and 5HT1F

[55] H. Pham-The, T. Garrigues, M. Bermejo, I. González-Álvarez, M. Cruz Monteagudo, M. Ángel Cabrera-Pérez, *Mol. Pharmaceutics*, **2013**, 10, (6), 2445–2461.
Provisional Classification and *in Silico* Study of Biopharmaceutical System Based on Caco-2 Cell Permeability and Dose Number

[56] A. Kovatcheva, A. Golbraikh, S. Oloff, Y. Xiao, W. Zheng, P. Wolschann, G. Buchbauer, A. Tropsha, *J.Chem. Inf. Comput. Sci.*, **2004**, 44, (2), 582–595.
Combinatorial QSAR of Ambergris Fragrance Compounds

E-state descriptors

[57] H. H. Maw, L. H. Hall **,** *J.Chem. Inf. Comput. Sci.*, **2001,** 41, (5), 1248–1254.
E-State Modeling of Corticosteroids Binding AffinityValidation of Model for Small Data Set

[58] I. V. Tetko, V. P. Solov'ev, A. V. Antonov, X. Yao, J.P. Doucet, B. Fan, F. Hoonakker, D. Fourches, P. Jost, N. Lachiche, A. Varnek, *J.Chem. Inf. Model.*, **2006,** 46, (2), 808–819.
Benchmarking of Linear and Nonlinear Approaches for Quantitative Structure−Property Relationship Studies of Metal Complexation with Ionophores

[59] I. V. Tetko, V. Yu. Tanchuk, T. N. Kasheva, A. E. P. Villa, *J.Chem. Inf. Comput. Sci.,* **2001,** 41, (6), 1488–1493.
Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices

[60] E. S. Wisniewski, A. W. Castleman, Jr., *J.Phys. Chem. A,* **2002,** 106, (45), 10843–10848.
Femtosecond Photodissociation Dynamics of Excited-State SO2

[61] H. H. Maw, L. H. Hall, *J.Chem. Inf. Comput. Sci.*, **2002,** 42, (2), 290–298.

E-State Modeling of HIV-1 Protease Inhibitor Binding Independent of 3D Information
[62] R. L. McNaughton, J.Min Chin, W. W. Weare, R. R. Schrock, B. M. Hoffman ;, *J.Am. Chem. Soc.,* **2007,** 129, (12), 3480–348.
EPR Study of the Low-Spin [d3; S =1/2], Jahn−Teller-Active, Dinitrogen Complex of a Molybdenum Trisamidoamine
[63] K. L. Knappenberger, Jr. A. W. Castleman, Jr., *J.Phys. Chem. A,* **2004,** 108, (1), 9–14.
Photodissociation of Sulfur Dioxide: The Ẽ State Revisited
[64] H. Qian, P. Shi, *J.Phys. Chem. B,* **2009,** 113, (8), 2225–2230.
Fluctuating Enzyme and Its Biological Functions: Positive Cooperativity without Multiple States
[65] T. Chiu, Sung-Sau So, *J.Chem. Inf. Comput. Sci.,* **2004,** 44, (1), 147–153.
Development of Neural Network QSPR Models for Hansch Substituent Constants. 1. Method and Validations
[66] A. Varnek, N. Kireeva, I. V. Tetko, I. I. Baskin, V. P. Solov'ev, *J.Chem. Inf. Model.,* **2007,** 47, (3), 1111–1122.
Exhaustive QSPR Studies of a Large Diverse Set of Ionic Liquids: How Accurately Can We Predict Melting Points?
[67] Q. Hu, Y. Liang, H. Yin, X. Peng, K. Fang, *J.Chem. Inf. Comput. Sci.,* **2004,** 44, (4), 1193–1201.
Structural Interpretation of the Topological Index. 2. The Molecular Connectivity Index, the Kappa Index, and the Atom-type E-State Index
[68] K. Rose, L. H. Hall, L. B. Kier, *J.Chem. Inf. Comput. Sci.,* **2002,** 42, (3), 651–666.
Modeling Blood-Brain Barrier Partitioning Using the Electrotopological State
[69] J.A. Grodnitzky, J.R. Coats, *J.Agric. Food Chem.,* **2002,** 50, (16), 4576–4580.
QSAR Evaluation of Monoterpenoids' Insecticidal Activity
[70] C. Catana, H. Gao, C. Orrenius, P. F. W. Stouten, *J.Chem. Inf. Model.,* **2005,** 45, (1), 170–176.
Linear and Nonlinear Methods in Modeling the Aqueous Solubility of Organic Compounds
[71] G. Schuurmann, R. Ebert, R. Kuhne, *Environ. Sci. Technol.,* **2006,** 40, (22), 7005–7011.
Prediction of the Sorption of Organic Compounds into Soil Organic Matter from Molecular Structure
[72] S. Liu, C. Yin, Z. Li, S. Cai, *J.Chem. Inf. Comput. Sci.,* **2001,** 41, (2), 321–329.
QSAR Study of Steroid Benchmark and Dipeptides Based on MEDV-13
[73] D. R. Albaugh, L. Mark Hall, D. W. Hill, T. M. Kertesz, M.Parham, L. H. Hall, David F. Grant, *J.Chem. Inf. Model.,* **2009,** 49, (4), 788–799.
Prediction of HPLC Retention Index Using Artificial Neural Networks and IGroup E-State Indices
[74] L. M. Hall, L. H. Hall, L. B. Kier, *J.Chem. Inf. Comput. Sci.,* **2003,** 43, (6), 2120–2128.
Modeling Drug Albumin Binding Affinity with E-State Topological Structure Representation
[75] R. D. Beger, D. A. Buzatu, J.G. Wilkes, O. Lay, Jr., *J.Chem. Inf. Comput. Sci.,* **2002,** 42, (5), 1123–1131.
Comparative Structural Connectivity Spectra Analysis (CoSCoSA) Models of Steroid Binding to the Corticosteroid Binding Globulin
[76] I. V. Tetko, V. Yu. Tanchuk, T. N. Kasheva, A. E. P. Villa, *J.Chem. Inf. Comput. Sci.,* **2001,** 41, (5), 1407–1421.
Prediction of n-Octanol/Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-State Indices
[77] R. K, Mitra I., *Curr Comput Aided Drug Des.,* **2012,** 8, (2), 135-58.

Electrotopological state atom (E-state) index in drug design, QSAR, property prediction and toxicity assessment.

[78] J.Huuskonen, *J.Chem. Inf. Comput. Sci.*, **2001,** 41, (2), 425–429.
QSAR Modeling with the Electrotopological State: TIBO Derivatives

[79] M. J.Tubergen, R. J.Lavrich, D. F. Plusquellic, R. D. Suenram, *J.Phys. Chem. A*, **2006,** 110, (49), 13188–13194.
Rotational Spectra and Conformational Structures of 1-Phenyl-2-propanol, Methamphetamine, and 1-Phenyl-2-propanone

[80] D. Cao, Y. Liang, J.Yan, G. Tan, Q. Xu, S. Liu, *J.Chem. Inf. Model.*, **2013,** 53, (11), 3086–3096.
PyDPI: Freely Available Python Package for Chemoinformatics, Bioinformatics, and Chemogenomics Studies

[81] D. R. Albaugh, L. Hall, D. W. Hill, T. M. Kertesz, M. Parham, L. H. Hall, D. F. Grant, *J.Chem. Inf. Model.*, **2009,** 49, (4), 788–799.
Prediction of HPLC Retention Index Using Artificial Neural Networks and IGroup E-State Indices

[82] M.C. Sharma, D.V. Kohli, J.Saudi Chemical Society, 18(1), **2014**, 35-45.
Insight into the structural requirement of substituted quinazolinone biphenyl acylsulfonamides derivatives as Angiotensin II AT_1 receptor antagonist: 2D and 3D QSAR approach

Geom descriptors

[83] L. Franke, O. Schwarz, L. Müller-Kuhrt, C. Hoernig, L. Fischer, S. George, Y.Tanrikulu, P. Schneider, O. Werz, D. Steinhilber, G. Schneider , *J.Med. Chem.*, **2007,** 50, (11), 2640–2646.
Identification of Natural-Product-Derived Inhibitors of 5-Lipoxygenase Activity by Ligand-Based Virtual Screening

[84] K. Liu, J.Feng, S. S. Young , *J.Chem. Inf. Model.*, **2005,** 45, (2), 515–522.
PowerMV: A Software Environment for Molecular Viewing, DescriptorGeneration, Data Analysis and Hit Evaluation

[85] C. Ma, J. S. LazoXiang-Qun Xie, *ACS Comb. Sci.*, **2011,** 13, (3), 223–231.
Compound Acquisition and Prioritization Algorithm for Constructing Structurally Diverse Compound Libraries

[86] Y. A. Abramov, *Mol. Pharmaceutics*, ASAP (**2015**).
Major Source of Error in QSPR Prediction of Intrinsic Thermodynamic Solubility of Drugs: Solid vs Nonsolid State Contributions?

[87] S. S. Young, M. Wang, F. Gu**,** *J.Chem. Inf. Comput. Sci.*, **2003,** 43, (6), 1916–1921.
Design of Diverse and Focused Combinatorial Libraries Using an Alternating Algorithm

[88] K. P. Singh, S. Gupta, A. Kumar, D. Mohan, *Chem. Res. Toxicol.*, **2014,** *27, (5),* 741–753.
Multispecies QSAR Modeling for Predicting the Aquatic Toxicity of Diverse Organic Chemicals for Regulatory Toxicology

[89] S. A. Mirkhani, F. Gharagheizi, *Ind. Eng. Chem. Res.*, **2012,** 51, (5), 2470–2477.
Predictive Quantitative Structure–Property Relationship Model for the Estimation of Ionic Liquid Viscosity

[90] J.J.Panek, A. Jezierska, M. Vračko, *J.Chem. Inf. Model.*, **2005,** 45, (2), 264–272.
Kohonen Network Study of Aromatic Compounds Based on Electronic and Nonelectronic Structure Descriptors

[91] Yi Wang, Y. Shao, Y. Wang, L. Fan, X. Yu, X. Zhi, C. Yang, H. Qu, X. Yao, H. Xu, *J.Agric. Food Chem.*, **2012,** *60, (34),* 8435–8443.
Synthesis and Quantitative Structure–Activity Relationship (QSAR) Study of Novel Isoxazoline and Oxime Derivatives of Podophyllotoxin as Insecticidal Agents

[92] J.K. Wegner, H. Fröhlich, R. Zell, *J.Chem. Inf. Comput. Sci.*, **2004,** 44, (3), 931–939.
Feature Selection for Descriptor Based Classification Models. 2. Human Intestinal Absorption (HIA)

[93] Z. Che, S. Zhang, Y. Shao, L. Fan, H. Xu, X. Yu, X. Zhi, X. Yao, R. Zhang, *J.Agric. Food Chem.*, **2013,** 61, (24), 5696–5705.
Synthesis and Quantitative Structure–Activity Relationship (QSAR) Study of Novel N-Arylsulfonyl-3-acylindole Arylcarbonyl Hydrazone Derivatives as Nematicidal Agents

[94] M. Fernandez, N. R. Trefiak, T. K. Woo, *J.Phys. Chem. C*, **2013,** 117, (27), 14095–14105.
Atomic Property Weighted Radial Distribution Functions Descriptors of Metal–Organic Frameworks for the Prediction of Gas Uptake Capacity

[95] K. Baumann, *J.Chem. Inf. Comput. Sci.*, **2002,** 42, (1), 26–35.
An Alignment-Independent Versatile Structure Descriptor for QSAR and QSPR Based on the Distribution of Molecular Features

[96] A. Amadasi, J.Surface, F. Spyrakis, P. Cozzini, M. G. E. Kellogg, *J.Med. Chem.*, **2008,** 51, (4), 1063–1067.
Robust Classification of "Relevant" Water Molecules in Putative Protein Binding Sites

[97] M. Dervarics, F. Ötvös, T. A. Martinek, *J.Chem. Inf. Model.*, **2006,** 46, (3), 1431–1438.
Development of a Chirality-Sensitive Flexibility Descriptor for 3+3D-QSAR

[98] B. E. Mattioni, P. C. Jurs, *J.Chem. Inf. Comput. Sci.*, **2002,** *42, (2),* 232–240.
Prediction of Glass Transition Temperatures from Monomer and Repeat Unit Structure Using Computational Neural Networks

[99] M. S. Costanza-Robinson, K. H. Harrold, R. M. Lieb-Lappen, *Environ. Sci. Technol.*, **2008,** 42, (8), 2949–2956.
X-ray Microtomography Determination of Air−Water Interfacial Area−Water Saturation Relationships in Sandy Porous Media

[100] B. E. Mattioni, P. C. Jurs, *J.Chem. Inf. Comput. Sci.*, **2002,** 42, (1), 94–102.
Development of Quantitative Structure−Activity Relationship and Classification Models for a Set of Carbonic Anhydrase Inhibitors

[101] Z. Boudene, T. De Bruin, H. Toulhoat, P. Raybaud, *Organometallics*, **2012,** 31, (23), 8312–8322.
A QSPR Investigation of Thermal Stability of [Al(CH3)O]n Oligomers in Methylaluminoxane Solution: The Identification of a Geometry-Based Descriptor

[102] A. Cammarata, W. Zhang, P. Shiv Halasyamani, J. M. Rondinelli, *Chem. Mater.*, **2014,** 26, (19), 5773–5781.
Microscopic Origins of Optical Second Harmonic Generation in Noncentrosymmetric–Nonpolar Materials

[103] S. J.Patankar, P. C. Jurs**,** *J.Chem. Inf. Comput. Sci.*, **2002,** 42, (5), 1053–1068.
Prediction of Glycine/NMDA Receptor Antagonist Inhibition from Molecular Structure

[104] W. Nowik, S. Héron, M. Bonose, M. Nowik, A. Tchapla, *Anal. Chem.*, **2013,** 85, (20), 9449–9458.
Assessment of Two-Dimensional Separative Systems Using Nearest-Neighbor Distances Approach. Part 1: Orthogonality Aspects

[105] F. J.Martín-Martínez, S. Fias, B. Hajgató, G. Van Lier, F. De Proft, P. Geerlings, *J.Phys. Chem. C,* **2013,** 117, (49), 26371–26384.
Inducing Aromaticity Patterns and Tuning the Electronic Transport of Zigzag Graphene Nanoribbons via Edge Design

[106] F. De Vleeschouwer, A. Toro-Labbé, S. Gutiérrez-Oliva, V. Van Speybroeck, M. Waroquier, P. Geerlings, F. De Proft, *J.Phys. Chem. A,* **2009,** 113, (27), 7899–7908.
Reversibility from DFT-Based Reactivity Indices: Intramolecular Side Reactions in the Polymerization of Poly (vinyl chloride)

[107] A. Rolo-Naranjo, E. Codorniu-Hernández Noel Ferro, *J.Chem. Inf. Model.,* **2010,** 50, (5), 924–933.
Quantum Chemical Associations Ligand−Residue: Their Role to Predict Flavonoid Binding Sites in Proteins

[108] N. R. McElroy, E. D. Thompson, P. C. Jurs, *J.Chem. Inf. Comput. Sci.,* **2003,** *43, (6),* 2111–2119.
Classification of Diverse Organic Compounds That Induce Chromosomal Aberrations in Chinese Hamster Cells

[109] C. Nativi, M. Cacciarini, O. Francesconi, A. Vacca, G. Moneti, R. Ienco, S. Roelens, *J.Am. Chem. Soc.,* **2007,** 129, (14), 4377–4385.
Pyrrolic Tripodal Receptors Effectively Recognizing Monosaccharides. Affinity Assessment through a Generalized Binding Descriptor

[110] K. Yeturu, N. Chandra, *J.Chem. Inf. Model.,* **2011,** 51, (7), 1725–1736.
PocketAlign A Novel Algorithm for Aligning Binding Sites in Protein Structures

[111] A. B. Camargo, E. Marchevsky, J.M. Luco **,** *J.Agric. Food Chem.*, 2007, 55, (8), 3096–3103.
QSAR Study for the Soybean 15-Lipoxygenase Inhibitory Activity of Organosulfur Compounds Derived from the Essential Oil of Garlic

[112] J.Desaphy, E. Raimbaud, P. Ducrot, D. Rognan, *J.Chem. Inf. Model.,* **2013,** 53, (3), 623–637.
Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs

[113] J.L. Hodgson, C. Y. LinMichelle, L. Coote, S. R. A. Marque, K. Matyjaszewski, *Macromolecules,* **2010,** 43, (8), 3728–3743.
Linear Free-Energy Relationships for the Alkyl Radical Affinities of Nitroxides: A Theoretical Study

Randic indices

[114] F. C. G. Manso, H. S. Júnior, R. E. Bruns, A. F. Rubira, E. C. Muniz, *J. Molecular Liquids,* **2012,** 165, 125-132.
Development of a new topological index for the prediction of normal boiling point temperatures of hydrocarbons: The Fiindex

[115] M. Randić, J.Zupan, *Topology in Chemistry,* **2002,** 249-291.
On the Structural Interpretation of Topological Indices

[116] E. Estrada, *Chemical Physics Letters,* **2001,** 336, 3–4, 248-252.
Generalization of topological indices

[117] H. Wang, *Discrete Mathematics,* **2008,** 308, 15, 3407-3411.
Extremal trees with given degree sequence for the Randić index

[118] C. Delorme, O. Favaron, D. Rautenbach, *Discrete Mathematics*, **2002,** 257, 1, 29-38.
On the Randić index

[119] I. Gutman, M. Robbiano, B. S. Martín, *Linear Algebra and its Applications,* **2015,** 478, 241-255.
Upper bound on Randić energy of some graphs

[120] R. Liu, W. Chee Shiu, *Discrete Applied Mathematics,* **2015,** 186, 168-175.
General Randić matrix and general Randić incidence matrix
[121] J. Rada, C. Uzcátegui, *Discrete Applied Mathematics,* **2005,** 150, 1–3, 232-250.
Randić ordering of chemical trees
[122] D. Vukičević, *Discrete Applied Mathematics,* **2010,** 158, 18, 2056-2065.
Which generalized Randić indices are suitable measures of molecular branching?
[123] W. Chee Shiu, L. Zhang, *Discrete Mathematics,* **2009,** 309, 13, 4409-4416.
The maximum Randić index of chemical trees with pendants
[124] M. Knor, B. Lužar, R. Škrekovski, *Discrete Applied Mathematics,* **2015,** 181, 160-166.
Sandwiching the (generalized) Randić index
[125] B. Wu, J. Yan, X. Yang, *Discrete Applied Mathematics,* **2014,** 178, 163-165.
Randić index and coloring number of a graph
[126] A. Taherpour, F. Shafiei, *J. Molecular Structure: THEOCHEM,* **2005,** 726, 1–3, 183-188.
The structural relationship between Randić indices, adjacency matrixes, distance matrixes and maximum wave length of linear simple conjugated polyene compounds
[127] J. Liu, *Discrete Applied Mathematics,* **2013,** 161, 16–17, 2544-2548.
On a conjecture of the Randić index and the minimum degree of graphs
[128] T. R. Divnić, L. R. Pavlović, *Discrete Applied Mathematics,* **2013,** 161, 7–8, 953-960.
Proof of the first part of the conjecture of Aouchiche and Hansen about theRandić index
[129] W. Linert, K. Kleestorfer, F. Renz, I. Lukovits, *J. Molecular Structure: THEOCHEM,* **1995,** 337, 2, 121-127.
Description of cyclic and branched-acyclic hydrocarbons by variants of the hyper-Wiener index
[130] A. Behtoei, M. Jannesari, B. Taeri, *Applied Mathematics Letters,* **2009,** 22, 10, 1571-1576.
Maximum Zagreb index, minimum hyper-Wiener index and graph connectivity

TOPOlogical descriptors

[131] B. Ren, *Chemometrics and Intelligent Laboratory Systems,* **2003,** 66, 1, 29-39. Atom-type-based I topological descriptors for quantitative structure–retention index correlations of aldehydes and ketones
[132] F. Liu, Y. Liang, C. Cao, N. Zhou, *Talanta,* **2007,** 72, 4, 1307-1315. QSPR study of GC retention indices for saturated esters on seven stationary phases based on novel topologicalindices
[133] C. Rojas, P. R. Duchowicz, P. Tripaldi, R. Pis Die, *Chemometrics and Intelligent Laboratory Systems,* **2015,** 140, 126-132.
QSPR analysis for the retention index of flavors and fragrances on a OV-101 column
[134] J.C. Drosos, M. Viola-Rhenals, R. Vivas-Reyes, *J. Chromatography A,* **2010,** 1217, 26, 4411-4421.
Quantitative structure–retention relationships of polycyclic aromatic hydrocarbons gas-chromatographic retention indices
[135] M. Jalali-Heravi, M. H. Fatemi, *J. Chromatography A,* **2001,** 915, 1–2, 177-183.
Artificial neural network modeling of Kováts retention indices for noncyclic and monocyclic terpenes
[136] K. Ośmiałowski, J.Halkiewicz, R. Kaliszan, *J. Chromatography A,* **1986,** 361, 63-69.
Quantum chemical parameters in correlation analysis of gas-liquid chromatographic retention indices of amines
[137] O. Farkas, I. G. Zenkevich, F. Stout, J.H. Kalivas, K. Héberger, *J. Chromatography A,* **2008,** 1198–1199, 188-195.

Prediction of retention indices for identification of fatty acid methyl esters
[138] B. Hemmateenejad, K. Javadnia, M. Elyasi, *Analytica Chimica Acta*, **2007,** 592, 1, 72-81.
Quantitative structure–retention relationship for the Kovats retention indices of a large set of terpenes: A combined data splitting-feature selection strategy
[139] J. M. Sutter, T. A. Peterson, P. C. Jurs, *Analytica Chimica Acta*, **1997,** 342, 2–3, 113-122.
Prediction of gas chromatographic retention indices of alkylbenzenes
[140] S. Sekušak, A. Sabljić, *J. Chromatography A*, **1993,** 628, 1, 69-79.
Calculation of retention indices by molecular topology: III. Chlorinated dibenzodioxins
[141] C. Zhou, X. Chu, C. Nie, *J.Phys. Chem. B*, **2007,** 111, (34), 10174–10179.
Predicting Thermodynamic Properties with a Novel Semiempirical Topological Descriptor and Path Numbers
[142] T. Etienne, X. Assfeld, A. Monari, *J.Chem. Theory Comput.*, **2014,** 10, (9), 3896–3905.
Toward a Quantitative Assessment of Electronic Transitions' Charge-Transfer Character
[143] R. Argüello-García, O. N. Medina-Campos, N. Pérez-Hernández, J.Pedraza-Chaverrí, G. Ortega-Pierres, *J.Agric. Food Chem.*, **2010,** 58, (21), 11226–11233.
Hypochlorous Acid Scavenging Activities of Thioallyl Compounds from Garlic
[144] J.M. Quigley, S. M. Naughton, *J.Chem. Inf. Comput. Sci.*, **2002,** 42, (4), 976–982.
The Interrelation of Physicochemical Parameters and Topological Descriptors for a Series of β-Blocking Agents
[145] Y. S. Kim, J.H. Kim, J.S. Kim, K. T. No, *J.Chem. Inf. Comput. Sci.*, **2002,** 42, (1), 75–81.
Prediction of Glass Transition Temperature (Tg) of Some Compounds in Organic Electroluminescent Devices with Their Molecular Properties

WHIM descriptors

[146] F. Alves de Lima Ribeiro, M. Miguel Castro Ferreir, *J. Molecular Structure: THEOCHEM*, **2003,** 663, 1–3, 109-126.
QSPR models of boiling point, octanol–water partition coefficient and retention time index of polycyclic aromatic hydrocarbons
[147] M. D. Ertürk, M. T. Saçan, *Ecotoxicology and Environmental Safety*, **2013,** 90, 61-68.
Assessment and modeling of the novel toxicity data set of phenols to Chlorella vulgaris
[148] M. D'Hondt, B. Gevaert, S. Stalmans, S. Van Dorpe, E. Wynendaele, K. Peremans, C. Burvenich, B. De Spiegeleer, *J. Pharmaceutical Analysis*, **2013,** 3, 2, 93-101.
Reversed-phase fused-core HPLC modeling of peptides
[149] Y. W. Li, Bo Li, J.He, P. Qian, *J. Molecular Structure*, **2011,** 998, 1–3, 53-61.
Quantitative structure–activity relationship study of antioxidative peptide by using different sets of amino acids descriptors
[150] A. Antonio D'Archivio, M. Anna Maggi, P. Mazzeo, F. Ruggieri, *Analytica Chimica Acta*, **2008,** 628, 2, 162-172.
Quantitative structure–retention relationships of pesticides in reversed-phase high-performance liquid chromatography based on WHIM and GETAWAY molecular descriptors
[151] A. Pérez-Garrido, M. Pérez González, A. Garrido Escudero, *Bioorganic & Medicinal Chemistry*, **2008,** 16, 10, 5720-5732.
Halogenated derivatives QSAR model using spectral moments to predict haloacetic acids (HAA) mutagenicity

[152] J. Xu, B. Chen, W. Xu, S. Zhao, C. Yi, W. Cui, *Chemometrics and Intelligent Laboratory Systems*, **2007,** 87, 2, 275-280.
3D-QSPR modeling and prediction of nonlinear optical responses of organic chromophores
[153] F. Zheng, G. Zheng, A. Gabriela Deaciuc, C. Zhan, L. P. Dwoskin, P. A. Crooks, *Bioorganic & Medicinal Chemistry*, **2007,** 15, 8, 2975-2992. Computational neural network analysis of the affinity of lobeline and tetrabenazine analogs for the vesicular monoamine transporter-2
[154] L. Saíz-Urra, M. Pérez González, Y. Fall, G. Gómez, *European J. Medicinal Chemistry*, **2007,** 42, 1, 64-70.
Quantitative structure–activity relationship studies of HIV-1 integrase inhibition. 1. GETAWAY descriptors
[155] F. Zheng, E. Bayram, S. P. Sumithran, J. T. Ayers, C. Zhan, J. D. Schmitt, L. P. Dwoskin, P. A. Crooks, *Bioorganic & Medicinal Chemistry*, **2006,** 14, 9, 3017-3037.
QSAR modeling of mono- and bis-quaternary ammonium salts that act as antagonists at neuronal nicotinic acetylcholine receptors mediating dopamine release
[156] E. Van Gyseghem, A. Elkihel, M. Jimidar, R. Sneyers, Y. Ver Heyden, *Analytica Chimica Acta*, **2006,** 562, 1, 73-84.
Chemometric selection of a small set of pharmaceutical active substances used in determining the orthogonality and similarity of chromatographic systems
[157] C. W. Yap, Z. R. Li, Y. Z. Chen, *J. Molecular Graphics and Modelling*, **2006,** 24, 5, 383-395.
Quantitative structure–pharmacokinetic relationships for drug clearance by using statistical learning methods
[158] M. P. González, C. Terán, M. Teijeira, M. J.González-Moa, *European J. Medicinal Chemistry*, **2005,** 40, 11, 1080-1086.
GETAWAY descriptors to predicting A2A adenosine receptors agonists
[159] M. P. González, C. Terán, M. Teijeira, P. Besada, M. J. González-Moa, *Bioorganic & Medicinal Chemistry Letters*, **2005,** 15, 10, 2641-2645.
Geometry, topology, and atom-weights assembly descriptors to predicting A1 adenosine receptors agonists
[160] Z. Jelcic, *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, **2004,** 242, 1–3, 159-166.
Solvent molecular descriptors on poly(D, L-lactide-co-glycolide) particle size in emulsification–diffusion process
[161] O. Farkas, K. Héberger, I. G. Zenkevich, *Chemometrics and Intelligent Laboratory Systems*, **2004,** 72, 2, 173-184.
Quantitative structure–retention relationships XIV: Prediction of gas chromatographic retention indices for saturated O-, N-, and S-heterocyclic compounds
[162] R. Carotti, C. Altomare, L. Savini, L. Chiasserini, C. Pellerano, M. P. Mascia, E. Maciocco, F. Busonero, M. Mameli, G. Biggio, E. Sanna, *Bioorganic & Medicinal Chemistry*, **2003,** 11, 23, 5259-5272.
High affinity central benzodiazepine receptor ligands. Part 3: insights into the pharmacophore and pattern recognition study of intrinsic activities of pyrazolo[4,3-c]quinolin-3-ones
[163] J.C Pinheiro, M. M. C Ferreira, O. A. S Romero, *J. Molecular Structure: THEOCHEM*, **2001,** 572, 1–3, 35-44.

Antimalarial activity of dihydroartemisinin derivatives against P. falciparum resistant to mefloquine: a quantum chemical and multivariate study

[164] W. D. Marzio, S Galassi, R Todeschini, F Consolaro, *Chemosphere,* **2001,** 44, 3, 401-406.
Traditional versus WHIM molecular descriptors in QSAR approaches applied to fish toxicity studies

Wiener descriptors

[165] S. Klavžar, I. Gutman, *Chemical Physics Letters,* **2003,** 373, 3–4, 328-332.
Relation between Wiener-type topological indices of benzenoid molecules

[166] A. Voelkel, T. Kopczyński, R. Gawałek, *J. Molecular Structure: THEOCHEM,* **1993,** 279, 217-222.
Structural descriptors in organic chemistry. Topological indices of isoquinoline derivatives

[167] E. Brendlé, E. Papirer, *J. Colloid and Interface Science,* **1997,** 194, 1, 217-224.
A New Topological Index for Molecular Probes Used in Inverse Gas Chromatography

[168] V. S. Raman, C. D. Maranas, *Computers & Chemical Engineering,* **1998,** 22, 6, 747-763.
Optimization in product design with properties correlated with topological indices

[169] Muhuo Liu, Bolian Liu, *Discrete Applied Mathematics,* **2010,** 158, 6, 699-705.
On the th smallest and th greatest modified Wiener indices of trees

[170] E. Estrada, *Topology in Chemistry,* **2002,** 185-207.
The Wiener Number in the Context of Generalized Topological Indices

[171] L. Chen, X. Li, M. Liu, *European J. Combinatorics,* **2014,** 36, 237-246.
The (revised) Szeged index and the Wiener index of a nonbipartite graph

[172] W. Yan, B. Yang, Y. Yeh, *Applied Mathematics Letters,* **2007,** 20, 3, 290-295.
The behavior of Wiener indices and polynomials of graphs under five graph decorations

[173] X. Li, A. F. Jalbout, *J. Molecular Structure: THEOCHEM,* **2003,** 634, 1–3, 121-125.
Bond order weighted hyper-Wiener index

[174] S. Nikolić, N. Trinajstić, M. Rić, *Chemical Physics Letters,* **2001,** 333, 3–4, 319-321.
Wiener index revisited

[175] A. Hamzeh, S. Hossein-Zadeh, A. R. Ashrafi, *Applied Mathematics Letters,* **2011,** 24, 7, 1099-1104.
Wiener index of composite graphs

[176] V. Sills, H. Wang, *Discrete Applied Mathematics,* **2012,** 160, 10–11, 1615-1623.
On the maximal Wiener index and related questions

[177] A. Behmaram, H. Yousefi-Azari, A. R. Ashrafi, *Applied Mathematics Letters,* **2012,** 25, 10, 1510-1513.
Wiener polarity index of fullerenes and hexagonal systems

[178] X. Qi, B. Zhou, *Computers & Mathematics with Applications,* **2011,** 62, 1, 523-531.
Extremal properties of reciprocal complementary Wiener number of trees

[179] X. Deng, J. Zhang, *Applied Mathematics Letters,* **2012,** 25, 3, 580-585.
Equiseparability on terminal Wiener index

[180] M. Knor, P. Potočnik, R. Škrekovski, *Discrete Applied Mathematics,* **2012,** 160, 15, 2234-2245.
The Wiener index in iterated line graphs

[181] J. Fink, B. Lužar, R. Škrekovski, *Discrete Applied Mathematics,* **2012,** 160, 12, 1851-1858.
Some remarks on inverse Wiener index problem

[182] M. J.Nadjafi-Arani, H. Khodashenas, A. R. Ashrafi, *Mathematical and Computer Modelling*, **2012,** 55, 3–4, 1644-1648.
Graphs whose Szeged and Wiener numbers differ by 4 and 5
[183] J. Liu, I. Gutman, Z. Mu, Y. Zhang, *Applied Mathematics and Computation*, **2012,** 218, 18, 9528-9535.
q-Wiener index of some compound trees
[184] H. Yousefi-Azari, M. H. Khalifeh, A. R. Ashrafi, *J. Computational and Applied Mathematics*, **2011,** 235, 16, 4866-4870.
Calculating the edge Wiener and edge Szeged indices of graphs
[185] M. J. Nadjafi-Arani, H. Khodashenas, A. R. Ashrafi, *Discrete Mathematics*, **2011,** 311, *20*, 2233-2237.
On the differences between Szeged and Wiener indices of graphs
[186] A. Karbasioun, *Scientia Iranica*, **2011,** 18, 6, 1623-1626.
The Wiener index of a new class of carbon nanojunctions
[187] G. G. Cash, *Applied Mathematics Letters*, **2002,** 15, 7, 893-895.
Relationship between the Hosoya polynomial and the hyper-Wiener index
[188] M. Hino, *Bulletin des Sciences Mathématiques*, **2011,** 135, 6–7, 667-683.
Dirichlet spaces on H-convex sets in Wiener space

MOLGEN suit

[189] T. Grüner, A. Kerber, R. Laue, M. Meringer, K. Varmuza, W. Werther, *MATCH Commun. Math. Comput. Chem.*, **1998,** 38, 173-180.
MASSMOL
[190] A. Kerber, R. Laue, M. Meringer and K. Varmuza, *Advances in Mass Spectrometry*, **2001,** 15, 939-940.
MOLGEN-MS: Evaluation of Low Resolution Electron Impact Mass Spectra with MS Classification and Exhaustive Structure Generation
[191] J.Meiler, M. Meringer, *MATCH Commun. Math. Comput. Chem.*, **2002,** 45, 85-108.
Ranking MOLGEN Structure Proposals by 13C NMR Chemical Shift Prediction with ANALYZE
[192] A. Kerber, R. Laue, M. Meringer, C. Rücker, *MATCH Commun. Math. Comput. Chem.*, **2004,** 51, 187-204.
MOLGEN-QSPR, a Software Package for the Study of Quantitative Structure Property Relationships
[193] C. Rücker, M. Meringer, A. Kerber, *J.Chem. Inf. Comput. Sci.*, **2004,** 44, 2070-2076.
QSPR Using MOLGEN-QSPR: The Example of Haloalkane Boiling Points

Omnimetrics

[194] T. Grüner, A. Kerber, R. Laue, M. Liepelt, M. Meringer, K. Varmuza, W. Werther, *MATCH Commun. Math. Comput. Chem.*, **1998,** 37, 163-177. Bestimmung von Summenformeln aus Massenspektren durch Erkennung überlagerter Isotopenmuster
[195] K. RamaKrishna, R. Sambasiva Rao, *J. Applicable Chemistry*, 2015, 4 (2), 355-449.
Omnimetrics, Part II: Applications of neural networks (Ma_NN) in Enviromentrics
[196] K. RamaKrishna, R. Sambasiva Rao, *J. Applicable Chemistry*, 2015, 4 (1), 1-98.
OmniMetrics, Part I: Applications of neural networks (Ma_NN) in Medicinometrics and pharmacometrics

[197] K. RamaKrishna, V. A. Ramam, R. Sambasiva Rao, *J. Applicable Chemistry*, **2014,** 3 (5), 1807-1893.
Mathematical Neural Network (MaNN) Models, Part V: Radial basis function (RBF) neural networks (NNs) in Chemometrics, Envirometrics and Medicinometrics (ChEM)

[198] K. RamaKrishna, V. A. Ramam, R. Sambasiva Rao, *J. Applicable Chemistry*, **2014,** 3, 6, 2209-2311.
Mathematical Neural Network (MaNN) Models Part VI: Single-layer perceptron [SLP] and Multi-layer perceptron [MLP] Neural networks in ChEM- Lab

[199] K. RamaKrishna, Ch. V. Kameswara Rao, R. Sambasiva Rao, *J. Applicable Chemistry*, **2013,** 2 (4): 698-713.
Tutorial on gravitational algorithm in Structure activity relationships (SXR)

[200] I.Suryanarayana, A. Braibanti, R. Sambasiva Rao, V. Ananta Ramam, D. Sudarsan, G. Nageswara Rao, *Fisheries Research*, **2008,** 92,115-139.
Neural Networks in Fisheries Research

AUTHORS' ADDRESSES

1. **K RamaKrishna**
Department of Chemistry,
Gitam Institute of Science,
Gitam University,
Visakhapatnam, A.P
2. **Ch. V. Kameswara Rao**
Department of Chemistry,
Basic sciences and Humanities,
GMRIT, Rajam 532 127, AP
3. **R. Sambasiva Rao**
School of Chemistry,
Andhra University,
Visakhapatnam 530 003, A.P