Chapter 2
Molecular Descriptors in QSPR/QSAR Modeling

Shahin Ahmadi, Sepideh Ketabi, and Marjan Jebeli Javan

Abstract Molecular descriptors are mathematical representation of a molecule obtained by a well-specified algorithm applied to a defined molecular representation or a well-specified experimental procedure. The molecular descriptors as the core feature-independent parameters used to predict biological activity or molecular property of compounds in the quantitative structure property/activity relationship (QSPR/QSAR) models. Over the years, more than 5000 molecular descriptors have been introduced and calculated using different software. In this chapter, the main classes of theoretical molecular descriptors including 0D, 1D, 2D, 3D, and 4D-descriptors are described. The most significant progress over the last few years in chemometrics, cheminformatics, and bioinformatics has led to new strategies for finding new molecular descriptors. The different approaches for deriving molecular descriptors here reviewed, and some of the new important molecular descriptors and their applications are presented.

Keywords Molecular descriptors · QSAR · QSPR · Chemometrics · Chemoinformatic

Abbreviations

MoRSE	3D-Molecular Representation of Structures Based on Electron Diffraction
ACE	Angiotensin-Converting Enzymes
AFM	Atomic Force Microscopy
AZI	Augmented Zagreb Index
BET	Brunauer, Emmett, and Teller
CORAL	CORrelation And Logic

S. Ahmadi (✉) · S. Ketabi · M. Jebeli Javan
Department of Chemistry, Faculty of Pharmaceutical Chemistry, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran
e-mail: s.ahmadi@iautmu.ac.ir; ahmadi.chemometrics@gmail.com

A. P. Toropova and A. A. Toropov (eds.), *QSPR/QSAR Analysis Using SMILES and Quasi-SMILES*, Challenges and Advances in Computational Chemistry and Physics 33, https://doi.org/10.1007/978-3-031-28401-4_2

DHFR	Dihydrofolate Reductase
DLS	Dynamic Light Scattering
EM	Electronic Microscopy
EDX	Energy Dispersive X-ray Spectrometry
ESEM	Environmental Scanning Electron Microscopy
FFF	Field Flow Filtration
FMO	Frontier Molecular Orbital Theory
HOMO	Highest Occupied Molecular Orbital
WW	Hyper-Wiener Index
ICPOES	Inductively Coupled Plasma Emission Spectroscopy
ICP-MS	Inductively Coupled Plasma Mass Spectrometry
LC	Liquid Chromatography
LUMO	Lowest Unoccupied Molecular Orbital
MW	Molecular Weight
MVC	Multivariate Characterization
PCA	Principal Component Analyses
PPs	Principal Properties
QSAR	Quantitative Structure–Activity Relationship
QSPR	Quantitative Structure–Property Relationship
SMILES	Simplified Molecular Input Line Entry System
TMACC	Topological Maximum Cross Correlation
TEM	Transmission Electron Microscopy

2.1 Introduction

2.1.1 History

The history of molecular descriptors as a feature vector for each compound is closely related to the concept of molecular structure [1]. The years between 1860 and 1880 were marked by a strong disagreement about the theory of molecular structure, which arose from studies on substances showing optical isomerism and Kekulé's (1867–1861) studies on the structure of benzene [2].

Today, many chemical, physical, and biological characteristics of compounds rely on the principle that these parameters are effects of its structural descriptors.

In 1868, Crum-Brown and Fraser [3] introduced first formulation about relationship between the bioactivity/property of a chemical (Φ) and its chemical constitution (C), as the following equation:

$$\Phi = f(C) \tag{2.1}$$

Based on this concept, many studies were conducted on the relationship of molecular descriptors to observed properties, including the relationship between the anesthetic power of various aliphatic alcohols with chain length of carbon and molecular weight [4], between the color of disubstituted benzenes with various ortho-, meta-, and para-orienting [5], and between the narcotic toxicity and solubility in water [6].

One of the most attractive quantitative structure–activity relationship (QSAR) approach is the Hammett equation [7]. In 1973, he showed a linear relationship between the rate constants of a series of methyl ester reactions with $N(CH_3)_3$ and the ionization equilibrium constants of the related carboxylic acids in aqueous solution at ambient temperature. The linear relationship between the ionization constant of the ester containing a substituent X in the meta (m) or para (p) orientation (K_X) and the ionization constant of the unsubstituted ester (K_H) is defined by the following formula:

$$\log\left(\frac{K_X}{K_H}\right) = \rho \cdot \sigma_X, \tag{2.2}$$

where σ_X is the constant of the substituent in m or p position is indicated by σ_m or σ_p, respectively. The absolute value of σ, which varies for each substituent, refers to the measure of the global electronic effect exerted on the reaction center by the presence of substituent X. The sign of σ is positive for electron-withdrawer and negative for electron-donor substituent. The electronic induction effect and the electronic resonance effect denote by σ_I and σ_R, respectively; the constant for the unsubstituted aromatic ring as a reference represented by σ_R^0. Hammett's equation in this case defined by the following equation.

$$\log\left(\frac{K_X}{K_H}\right) = \rho_I \cdot \sigma_I + \rho_R \cdot \sigma_R^0 \tag{2.3}$$

2.1.2 QSPR/QSAR Modeling

In cheminformatics, a QSPR/QSAR model, either qualitative or quantitative, is a mathematical function that can be used to describe the connection between the molecular structures of a series of chemical compounds and their physicochemical properties/biological activities [8–14].

This field of knowledge assumes that the activity or property of a compound depends on its structural features, which affect its overall activities and properties [15–19].

Despite the formal differences between different methodologies, each QSPR/QSAR method is based on a QSPR/QSAR table that can be generalized as presented in Fig. 2.1 [20].

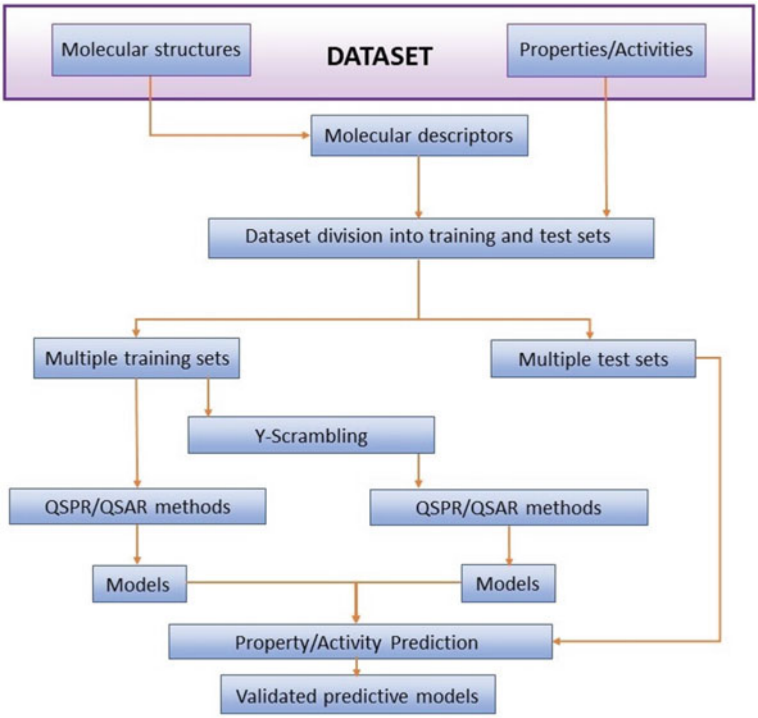


Fig. 2.1 Flowchart of the combinatorial QSAR methodology

The differences in various QSPR/QSAR studies can be explained in the following terms:

- Endpoint value
- Molecular descriptors
- Optimization algorithms.

Endpoint value as dependent variables can generally be of three types:

- Continuous

This endpoint is real values covering certain range, e.g., physicochemical properties of compounds such as boiling point and melting point. or IC_{50} values and binding constant.

- Categorical-related

This is classes of activities covering certain range of values, e.g., active and inactive compounds.

- Adjacent classes of metabolic stability

Adjacent classes of metabolic stability such as unstable, moderately stable, stable; and categorical-unrelated (i.e., classes of endpoints that do not relate to each other in any continuum, e.g., compounds that belong to different pharmacological categories, or compounds that are categorized as drugs vs. non-drugs).

Understanding this classification is indeed very important because the choice of descriptor types as well as modeling methods is often determined by the type of endpoints. Thus, in general the latter two types require classification modeling methods, whereas the former type of the target properties allows using linear regression modeling. Therefore, the latter two types require categorical modeling methods, generally while the former type of endpoint characteristics allows the use of linear regression modeling. Methods related to data analysis are called classification or continuous QSPR/QSAR.

2.1.3 Molecular Descriptors

Chemical descriptors as independent features in QSPR/QSAR modeling are usually classified into the following two types:

- Continuous

There are so many continuous descriptors such as molecular weight or many molecular connectivity indices.

- Categorical-related

The categorized descriptors such as counts of functional groups, binary descriptors indicating the presence or absence of a chemical functional group or an atom in a molecule.

2.1.3.1 Types of Molecular Descriptors

Molecular descriptors can be obtained from different representations of molecules. Knowing various types of descriptors is also critical for a fundamental understanding of QSPR/QSAR modeling because, as mentioned above, any modeling requires establishing a relationship between the chemical similarity of compounds and their target properties [21–24]. Chemical similarity is calculated in descriptor space using various similarity metrics [25]. For example, in the case of continuous molecular descriptors, the Euclidean distance in the descriptor space is an advisable choice of similarity metric, while in the case of binary descriptors metrics such as the Tanimoto coefficient or the Manhattan distance seem more appropriate.

The grade of the sufficiency of molecular structure samples differs from 0 to 4D demonstrations.

0D Descriptors

The 0D models contain the simplest molecule interpretation that does not hold any information about atom connections. Chemical formula, which organizes the atom types and their occurrences within a molecule, is independent of any information about the molecular structure. Therefore, molecular descriptors gained from the chemical formula stated as 0D descriptors. The most usual examples are atom type, number of atoms, molecular weight (MW), and any function of atomic properties.

1D Descriptors

Substructure list representation can be classified as a 1D description and contain of structural fragments of a molecule such as functional groups, bonds, rings, and substituents. Therefore, 1D descriptors do not involve a full information of molecular structure. These descriptors are inanimate to any conformation variation and, hence, do not recognize between isomers.

2D Descriptors

The 2D models include knowledge about the structure of the compound on the basis of its structural formula [26]. These patterns solely mirror the topology of the molecule. Such templates are highly common. The ability of such methods is that the topology model of the molecular structure includes information about the possible combinations of the molecule in virtual form.

Evaluation of the internal atomic arrangement of compounds is done by topological parameters [27]. They originated from the topological exhibition of molecules and can be measured as structure-manifest descriptors. These factors numerically code data related to molecular shape, size, branching, attendance of heteroatoms, and multifold bonds in numeric form. These topological parameters show the correlation of atoms by the characteristic of chemical bonds.

In modeling distinct biological, physicochemical, and pharmacokinetic properties, they have considerable performance. A topological display of the molecule is accessible as a molecular diagram. This diagram is defined in mathematical phrases as $G = (V, E)$, where V is a series of vertices corresponding to the atoms of the molecule and E is a series of elements that initiate a double connection between pairs of vertices.

These chemical diagrams illustrate a non-numerical figure of the molecular compound although a numeric interpretation of the diagram is crucial for computing topological parameters [28].

Some common 2D descriptors together with their description have been listed in the following.

Wiener (W) Index

The structure descriptor based on the classical molecular diagram is the Wiener index (W) which has become one of the most heavily applied descriptors in QSAR/QSPR approaches [29]. The descriptor is defined as the sum of edges on the shortest path in a chemical diagram.

Actually, the following equation denotes Wiener index $W(G)$ of the graph G (the graph G is a tree, T):

$$W(G) = \sum_{e \in E(G)} n_1(e|G) n_2(e|G) \tag{2.4}$$

$n_1(e|G)$ and $n_2(e|G)$ counts the vertices of G lying closer to the endpoints of the edge e than to its other endpoint

Hyper-Wiener Index (WW)

This index of a chemical tree T is defined as the sum of $n_1 n_2$ products over all pairs of u vertices of T [30]. In fact, WW is the path number, and it is defined as the sum of the distances between any two atoms in the molecule, in terms of atom-atom bonds. Actually, WW can be calculated by multiplying the number of atoms on one side of any path by those on the other side, and the sum of these values for all paths. Wiener index is restricted to bonds and in Hyper-Wiener index bond is replaced with path.

Modified Wiener Index (W*)

Bond contribution is determined by using the reciprocal of the number of atoms on each side of the bond [31].

Novel Wiener Index

It is obtained as an additive bond quantity, where the bond contribution is given as the product of the number of atoms close to each of the two points of each bond [32].

Connectivity Indices

It is structural invariant. Such indices are widely used in structure–property and structure–activity studies. These descriptors are on the basis of graph-theoretical constants that are presented to calculate the branching index of alkenes [33].

Kier and Hall extended these indices and intrinsic valence coupling indices to differentiate heteroatoms. Today, these phenomena have been optimized for a wide range of biological and physicochemical properties [34]. Randic [35] proposed some descriptors for topological indices: (i) they should be well-correlated with at least one feature; (ii) have structure commentary; (iii) be normal and self-determining; (iv) easily applied in a situational structure; (v) be free of empirical features; and (vi) be independent of other parameters.

Higher Order Connectivity

These indices are weight paths, where higher weight is given to terminal bonds and a lower weight to less exposed internal bonds [36].

Kier Shape

The descriptor defines shape indexes from molecular graphs. The shape of molecules is defined by the number of atoms and their bonding pattern which present in various orders [37].

Balaban Index

It is also one of the most distinctive molecular descriptors. Its value is independent of the molecular size or the number of rings [38].

Zagreb Indices

This descriptor is the first topological indices used for the total π-energy of conjugated molecules. The significant use of these indices is the distinction between the size of the molecules, flexibility, degree of branching, and entire shape [39].

Augmented Zagreb Index (AZI)

This index is based on the atom-bond connectivity (ABC index) used to obtain extreme values of AZI in chemical trees, and it can be used for upper and lower bonds' power of chemical trees [40].

Hosoya (Z)

It constructs QSAR/QSPAR models that describe the physical properties [41].

Modified Hosoya Index (Z*)

The frequency of occurrence of single CC bond in disjoint bond patterns is considered [42].

Autocorrelation Indices

This is a function of spatial separation and has particular advantageous for any QSAR/QSPAR study [43]

Szeged (SZ)

It is obtained as an additive bond quantity, where the bond contributions are given as the product of the number of atoms close to each of the two points of each bond [44].

Luckily, most of these parameters are identified in the topological descriptors. Therefore, they have been widely utilized in QSAR/QSPR simulation to determine the structural resemblance or disparity of chemical compounds.

Topological Maximum Cross Correlation (TMACC)

These descriptors generated from atom properties determined by molecular topology based on concepts derived from autocorrelation descriptors. In 2007, Topological Maximum Cross Correlation (TMACC) was developed through atomic features characterized by molecular topology [45]. These parameters are based on meanings derived from coefficient descriptors. The ability to decode TMACC descriptors using QSAR simulation of angiotensin-converting enzymes (ACE) and dihydrofolate reductase (DHFR) inhibitors was demonstrated by Spowage et al. [46]. Altogether, TMACC revealed specific properties for C domain-selective ACE inhibition, which was an improvement on prior QSAR studies [46].

The physical and chemical features of a molecule that are evaluated by examining its 2D structure are physicochemical descriptors. These features play a main role in characterizing the drug condensation in the body. The convenient characteristics of a drug can enhance its effect and thus its market value.

Therefore, investigating these features of a drug not only contributes to the general plan of drug safety but also plays a significant role in drug detection collaboration by optimizing the selected compounds. Thus, it is necessary to pay attention to properties like solubility, permeability, and lipophilicity that can warrant optimal power, as well as to select the volunteer compounds with proper physicochemical properties.

The lipophilicity of a drug is related to its dependence on a lipophilic surrounding. It is an essential feature in the movement of drugs in the body, which includes intestinal absorption, membrane penetrance, protein linkage, and dispensation among multiple tissues [47].

Generally, a drug exhibits negligible chemical absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties in the presence of low lipophilicity [48]. Many pieces of research have been conducted on in vitro cellular permeance, which have demonstrated its connection to lipophilicity with other parameters, like molecular size, hydrophilicity, hydrogen bonds, and degree of ionization. These factors are recognized to have a considerable role in the intestinal absorption of a molecule. Molecular size is the main operative influencing biological activity like intestinal absorption.

Hydrogen bond donors and lipophilicity play considerable roles in predicting human intestinal permeability [49]. MW is associated with reduced permeability. Solubility in water plays a significant role in the distribution of drugs and their permeance through biological membranes, and their redeploy and sorption.

3D Descriptors

The 3D QSAR models [50–53] provide complete structural data including composition, topology, and steric form of the molecule for only one conformer. These patterns are the most common. Geometrical descriptors are computed from the 3D correlations of atoms in a given molecule. These parameters are in contrast to topological descriptors in terms of data and distinction power for similar chemical structures and molecular compounds [54].

In addition, they also contain data procured from atomic van der Waals regions and their participation on the molecular surface. In spite of their high data quantity, these parameters normally have drawbacks.

Geometrical descriptors need geometry optimization and, thus, the overhead to compute them. Thus, new data are available and can be extracted for flexible molecules that can have different molecular compositions. However, this propels the complexity that can enhance considerably. In addition, most of these parameters (grid-based descriptors) require arrangement rules to accomplish molecule abduction. Different groups of descriptors can be recognized using the set of geometric descriptors [54].

A diversity of 3D descriptors is accessible, some of them are:

3D-Molecular Representation of Structures Based on Electron Diffraction (MoRSE)

MoRSE descriptors have been shown to have good modeling power for various biological and physicochemical properties and can also be used to simulate infrared spectra [55].

Weighted Holistic Invariant Molecular (WHIM)

WHIM descriptors are applied to obtain related 3D data about molecular shape, size, symmetry and atom dispensation and have been utilized to model several physicochemical and toxicology properties. At the minimum, ten distinct sorts of WHIM parameters with distinct molecular characteristics have been expanded [54].

3D Autocohesion

Using the autocohesion function, these parameters are computed at individual spots on molecular surface. For a specific geometry and sensitive conformational change, they are unique and are constant to rototranslation [56].

GEometry, Topology, Atom-Weights AssemblY (GETAWAY)

These parameters are on the basis of spatial coherence formula, which weights the atom to calculate van der Waals volume, atomic mass, and electronegativity alongside 3D data. According to data factors and the matrix operator, seven GETAWAY descriptors have been declared until now [54].

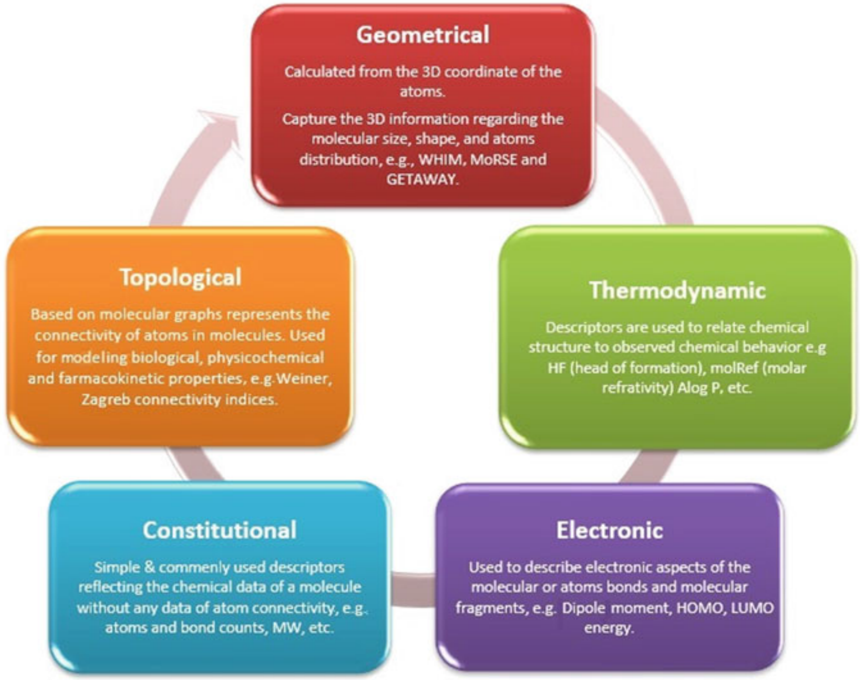
4D Descriptors

In 3D descriptors, the choice of the analyzed conformer is often random. The most adequate explanation of the molecular structure will be provided by 4D-QSAR patterns [57]. These models are similar to 3D models, but unlike them, structural data are discussed for a set of conformers (in essence, the fourth dimension), for a firm conformation.

Representation of molecular descriptors used in QSPR/QSAR modeling indicated in Fig. 2.2.

2.1.3.2 Molecular Descriptors' Resources

To get a considerable connection in QSAR studies, suitable descriptors must be used, whether they are empirical, theoretical, or derived from easily accessible experimental features of the molecules. Multiple descriptors mirror simple molecular features and thus can equip vision into the physicochemical characteristics of the property/activity under observation.



Calculated from the 3D coordinate of the atoms.

Capture the 3D information regarding the molecular size, shape, and atoms distribution, e.g., WHIM, MoRSE and GETAWAY.

Descriptors are used to relate chemical structure to observed chemical behavior e.g HF (head of formation), molRef (molar refrativity) Alog P, etc.

Used to describe electronic aspects of the molecular or atoms bonds and molecular fragments, e.g. Dipole moment, HOMO, LUMO energy.

Simple & commenly used descriptors reflecting the chemical data of a molecule without any data of atom connectivity, e.g. atoms and bond counts, MW, etc.

Based on molecular graphs represents the connectivity of atoms in molecules. Used for modeling biological, physicochemical and farmacokinetic properties, e.g Weiner, Zagreb connectivity indices.

Fig. 2.2 Representation of molecular descriptors used in QSPR/QSAR modeling

Quantum Chemical Descriptors. Quantum chemical computations are an important source of new molecular descriptors that can actually represent all electronic and geometrical properties of molecules and their interactions.

Quantum chemical and molecular modeling techniques provide the description of a large number of molecular and local values that determine the shape, reactivity, and binding characteristics of an entire molecule in addition to its molecular pieces and substituents.

In the last years, quantum chemical parameters have been significant in QSAR models helping researchers illustrate the biological activities and toxicity mechanisms of various chemicals. In the past decades, semiempirical calculations were the prior ways to generate descriptors owing to the restrictions of the software and applied systems. Recent advances in computational hardware and the expansion of effective algorithms have helped to expand molecular quantum mechanical computations. In particular, the parameters derived from density functional theory (DFT) and hybrid density functional calculations (mPW1PW91) have excellent potential through their better accuracy in contrast to the semiempirical procedure and have good efficiency to fit into the geometrical, electrostatic, and orbital energy calculations [58–61].

Since the context of large discrete physical data is encoded in a large number of theoretical descriptors, their usage in the scheme of instruction sets in QSAR studies offers two significant priorities: (a) molecules, their diverse parts, and their substitutions; can be instantly identified based on their molecular structure, and (b) the presented mechanism of action can be straight considered for the chemical reaction of the studied compounds [62]. As a result, the derived QSAR models contain data on the essence of the intermolecular interactions imported in specifying the biological or other properties of the investigated compounds. The most commonly used quantum chemical descriptors can be classified as follows:

Geometry Descriptors. The bond lengths, angles, and molecular dihedrals of the root segment should be the same for all molecules in the series.

Atomic Charges. In accordance with the classical theory of chemistry, all chemical interactions are either orbital (covalent) or electrostatic (polar) in nature. The electric charges in the molecule are clearly the order of the electrostatic interactions. Indeed, local electron density or charges have been shown to be momentous in a large number of physicochemical properties and chemical reactions of structures. Therefore, charge-based descriptors have been broadly utilized as indicators of chemical reactivity or as a measure of fragile intermolecular interactions. Numerous quantum chemical descriptors are derived from partial charge. Partial atomic charges are known as indicators of static chemical reactivity [63]. The computed σ- and π-electron densities on a specific atom determine the feasible direction of the chemical interactions and, hence, are often discussed as indices of directional reactivity. Unlike the total electron density, specific charges on atoms are observed as indicators of non-directional reactivity. Several sums of absolute or squared values of partial charges have also been used to characterize intermolecular interactions, e.g., solute–solvent interactions [64–66].

Molecular Orbital Energies. Highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies are very universal quantum chemical descriptors. It has been displayed [67] that these orbitals play an important role in controlling various chemical reactions and specifying electronic band gaps in solids. They are also in charge of the formation of several charge transfer complexes [63, 68]. Based on the frontier molecular orbital theory (FMO) of chemical reactivity, the organization of a transition state is owing to the interaction between the frontier orbitals (HOMO and LUMO) of the reacting fragments [69]. Therefore, the behavior of frontier molecular orbitals is distinct from others based on the general origins controlling the character of chemical reactions [69]. The HOMO energy is straightly connected to the ionization potential and characterizes the ability of the molecule to attack by electrophiles. The LUMO energy is straightly connected to the electron affinity and determines the readiness of the molecule against nucleophile attack. Both the HOMO and the LUMO energies are essential in radical reactions [70, 71]. The meaning of soft and hard nucleophiles and electrophiles is also connected to the relative energy of the HOMO/LUMO orbitals.

Soft nucleophiles have high-energy HOMOs. Hard nucleophiles have low-energy HOMOs. Soft electrophiles have low-energy LUMO, and hard electrophiles have high-energy LUMOs[72]. The HOMO–LUMO gap, i.e., the energy difference between HOMO and LUMO, is a major stability indicator [73].

$$E_{\text{gap}} = E_{\text{LUMO}} - E_{\text{HOMO}} \tag{2.5}$$

A large HOMO–LUMO gap indicates high resistance for the molecule by definition its less reactivity in chemical reactions [67]. The HOMO–LUMO gap has also been utilized as an estimate of the lowest stimulation energy of the molecule. However, this definition ignores electronic restructuring in the excited state and hence may mostly make incorrect theoretical results. The meaning of activation hardness (η) and softness (S) is also determined based on the HOMO–LUMO energy gap.

$$\eta = \frac{(E_{\text{LUMO}} - E_{\text{HOMO}})}{2} \tag{2.6}$$

$$S = \frac{1}{2\eta} \tag{2.7}$$

Activation hardness determines the rate of reaction at various sites of the molecule and is therefore related to anticipating direction effects [67]. The qualitative description of hardness is intimately connected to polarizability, as a reduction in the energy gap normally results in an easier polarization of the molecule [74].

Frontier Orbital Densities. Frontier orbital electron densities on atoms provide an effective alternative or accurate description of donor–acceptor interactions [71, 75]. Due to the theory of frontier electron reactivity, most chemical reactions happen in the location and direction where the overlap of the HOMO and LUMO of the respective reactants can be maximized [69].

In the matter of a donor molecule, both ionization potential (IE) and HOMO density (electrophilic electron density, f_r^E) are necessary to charge transfer:

$$f_r^E = \sum \left(C_{\text{HOMO},n}\right)^2; \quad C_{\text{HOMO},n} \text{ are atomic orbital factors in HOMO} \tag{2.8}$$

$$\text{IE} = -\text{EHOMO} \tag{2.9}$$

and in the terms of an acceptor molecule, LUMO density (nucleophilic electron density, f_r^N) and electron affinity (EA) are critical [63].

$$f_r^N = \sum \left(C_{\text{LUMO},n}\right)^2; \quad C_{\text{LUMO},n} \text{ are atomic orbital factors in LUMO} \tag{2.10}$$

$$\text{EA} = -E_{\text{LUMO}} \tag{2.11}$$

These descriptors have been applied in QSAR studies to characterize drug–receptor interaction sites. By comparing the relativities of different molecules, the frontier electron density should be normalized by the energy of the frontier molecular orbitals, and hence molecules with lower ionization potentials are predicted to be more reactive as nucleophiles. Absolute electronegativity index (χ), electron affinity (ω), and electron charge transfer (ΔN) are also determined based on ionization potential and electron affinity:

$$\chi = \frac{(I + A)}{2} \quad \text{absolute electronegativity} \tag{2.12}$$

$$\omega = \frac{\mu^2}{2\eta} \quad \text{electrophilicity index} \tag{2.13}$$

$$\Delta N = \frac{(\mu_B - \mu_A)}{2(\eta_A + \eta_B)} \quad \text{electron charge transfer} \tag{2.14}$$

Molecular Polarizability. The polarization of a molecule by an external electric [76] area is given by the potential tensors of order n of the molecular mass. The first-order term is used as polarizability (α):

$$\alpha = \frac{1}{3}\left(\alpha_{xx} + \alpha_{yy} + \alpha_{zz}\right) \tag{2.15}$$

The second-order term is mentioned in the first hyperpolarizability, etc. Therefore, the most considerable characteristic of molecular polarizability is binding to the molecular bulk or molar volume [73]. Polarizability values have been demonstrated to depend on hydrophobicity and other biological activities [77–79]. In addition, the electronic polarizability of the molecules contributes to the typical parameters of electrophilic super-delocalizability [80]. The first-order polarizability tensor includes data about feasible inductive interactions in the molecule [70, 73, 81, 82]. The total anisotropy of the polarizability (second-order term) determines the properties of a molecule as an electron acceptor:

$$\beta^2 = \frac{1}{2}[\left(\alpha_{xx} - \alpha_{yy}\right)^2 + \left(\alpha_{yy} - \alpha_{ZZ}\right)^2 + (\alpha_{ZZ} - \alpha_{xx})^2] \tag{2.16}$$

Dipole Moment and Polarity Indices. The polarity of a molecule is essential for several physicochemical properties. A large number of descriptors have been suggested to estimate the polarity effects. For instance, molecular polarity counts for chromatographic retention in a polar static phase [65, 83]. The dipole moment (μ) is the most obvious and is often used to explain the polarity of the molecule [64, 65, 70, 81, 84]. Difference between net charges on atoms (Δ) [68, 84], and topological electronic index (T_E) [68].

$$T_E = \sum_{ij, i \neq j} \frac{\left|q_i - q_j\right|}{r_{ij}^2} \tag{2.17}$$

The quadrupole moment tensor can also be applied as an index to characterize probable electrostatic interactions. However, such tensors belong to the selection of the coordinate system and thus the direction of the molecular root section must be the same for all molecules in the series [70].

Energy. The total energy computed by quantum mechanical methods has been presented as a good descriptor in several cases [64, 68, 85, 86].

In addition, thermodynamic parameters contain entropy ($S°$), internal energy (Eth), constant-enthalpy ($H°$), free energy ($G°$), zero-point vibrational energy (ZPE), and volume heat capacity (CV°) can be computed from frequency quantum mechanical calculations. Reaction enthalpy (ΔH), entropy (ΔS), and free energy (ΔG) can be calculated by the difference in heats of formation, entropy, and free energies of formation between reactants and products or between conjugate forms [87, 88]. The protonation energy, described as the difference between the total energy of the protonated and neutral forms of the molecule, can be discussed as a good scale of the power of hydrogen bonds (the higher the energy, the stronger the bond) and can be used to specify the correct position of the most desirable hydrogen bond acceptor [89].

The others. The descriptors considered above form the bulk of quantum chemical descriptors effectively used in QSAR/QSPR studies. Other descriptors have also been designed but do not fall into the categories mentioned above, such as frequency and NMR chemical shifts.

2.1.3.3 Empirical and Experimental Descriptors

Quantum chemical and molecular modeling techniques allow the description of many molecular and local values that determine the reactivity, binding features, and shape of a molecule in addition to molecular moieties and substituents. A principled combination of theoretical molecular descriptors with both empirical Hammett's substituent constants (σ_m and σ_p) [90, 91], Swain–Lupton's field and resonance constants (F and R) [92], hydrophobic constant (Π) [92], Taft's steric parameter (E_s) [92], Verloop's steric parameters [90, 91], etc., and experimental descriptors (substituent-induced chemical shifts, molecular weight and molecular refractivity (MR) [92]) are available. Table 2.1 shows the list of empirical and experimental descriptors.

The mentioned substituent descriptors can be categorized pursuant to three main cluster groups: (a) descriptors that capture the effects of the substituent on the aromatic ring (electronic charges on the ring carbon atoms, resonance and field substituent constants, and substituent-induced chemical shifts); (b) descriptors characterizing the properties of the majority of substituents (Verloop's steric parameters and the molecular refractivity) are clustered with theoretical descriptors describing the polarizability properties of the substituents, molecular polarizability anisotropy, dispersion interaction terms (IP*ANIS, IP*$\Sigma\Pi_{\text{mol}}$) and electrophilic super-delocalizability of the substituent.

Table 2.1 List of empirical and experimental descriptors

Descriptor	Definition	References
σ_x	Taft's substituent electronegativity effect parameter	[93]
σ_α	Taft's substituent polarizability effect parameter	[93]
σ_f	Taft's substituent field effect parameter	[93]
σ_r	Taft's substituent resonance effect parameter	[93]
C_0	^{13}C substituent chemical shift on the ortho-carbon atom	[94]
C_i	^{13}C substituent chemical shift on the ipso-carbon atom	[94]
C_m	^{13}C substituent chemical shift on the meta-carbon atom	[94]
C_p	^{13}C substituent chemical shift on the para-carbon atom	[94]
σ_m	Hammett's substituent constant for the meta position	[90, 91]
σ_p	Hammett's substituent constant for the para position	[90, 91]
F	Swain–Lupton's field constant	[92]
R	Swain–Lupton's resonance constant	[92]
Π	Π hydrophobic constant	[92]
MR	Molecular refractivity	[92]
E_s	Taft's steric parameter	[92]
H_a	Number of hydrogen bonds that the substituent can accept	[95]
H_d	Number of hydrogen bonds that the substituent can donate	[95]
L	Verloop multidimensional steric parameter	[90, 91]
B_1	Verloop multidimensional steric parameter	[90, 91]
B_2	Verloop multidimensional steric parameter	[90, 91]
B_3	Verloop multidimensional steric parameter	[90, 91]
B_4	Verloop multidimensional steric parameter	[90, 91]
μ_{ar}	Lien's group dipole moment for aromatic substituent	[22]
λ_{ar}	Testa's lipophobic constant for aromatic substituent	[95]

IP = ionization potential derived from the AM1 wave function.

ANIS = anisotropy of the molecular polarizability.

IP*ANIS = product of the molecular ionization potential and the anisotropy of the molecular polarizability.

IP*$\Sigma\Pi_{\text{mol}}$ = product of the molecular ionization potential and the sum of the self-atom polarizability over all the atoms of the molecule.

$\Sigma\Pi_{\text{XX}}$ = sum of the self-atom polarizability values of the substituent atoms.

$\Sigma\Pi_{\text{mol}}$ = sum of the self-atom polarizability over all the atoms of the molecule.

ΣS_X^H = sum of the electrophilic super-delocalizability on the substituent atoms.

$\Sigma S_{E,X}$ = sum of the electrophilic super-delocalizability (computed over all the occupied molecular orbitals) on the substituent atoms.

$\Sigma S_{N,X}$ = sum of the nucleophilic super-delocalizability (computed over all the unoccupied molecular orbitals) on the substituent atoms.

The hydrophobic parameter Π is near to this cluster and to the solvent hydrophobic available surface of the substituent and the electrophilic super-delocalizability with the polarizability of the benzene ring; (c) molecular dipole moments and their experimental and theoretical substituents and their square.

(a) Hammett substituent constants, substituent-induced chemical shifts, and Taft and Lupton's resonance constants are mapped by the first component, the major contribution of which is the electronic charges of the carbon atoms of the benzene ring, the super-electrophilic mobility of the benzene ring and the energy of frontier molecular orbitals; (b) Verloop steric descriptors and the molecular refraction along with substituent van der Waals volumes and molecular weight are mapped by the second principal component, which includes theoretical parameters described as polarizability ($\Sigma\Pi_{\mathrm{XX}}$, ANIS, $\Sigma\Pi_{\mathrm{mol}}$), dispersion forces (IP*$\Sigma\Pi_{\mathrm{mol}}$, IP*ANIS), and substituent reactivity indices (ΣS_X^H, $\Sigma S_{E,X}$, and $\Sigma S_{N,X}$). These recent cases perhaps indicate the portion of the molecular orbital development to molecular shape; (c) the third component models the lipophobic descriptor λ_{ar} and the lipophilic descriptor Π. The parameters that collaborate to this part are the dipole moments (consisting of the group dipole moment, μ_{ar}) and their square terms, the solvent available surfaces of the substituent, the energy difference between the HOMO and the LUMO (GAP), the Π-symmetry component of the electronic charges and the polarizability of the ring.

However, λ_{ar} and Π are not solely modeled by this section, as they also contribute significantly to the first and the third components, respectively. This suggests that more than one type of substituent effect specifies the values of these parameters. The same result is for the steric descriptors E_s modeled both by the first and the second components. These findings are similar to other research aimed at modeling Π [96] and E_s [97] and support the intricate character of these empirical parameters.

Empirical scales called principal properties (PPs) which define the physicochemical features of twenty naturally encoded amino acids were recently developed by Sjostrom and Wold [98].

Sjostrom et al. applied the PPs in the same way to categorize several types of signal peptides of different lengths [99]. Carlson and co-workers have reported principal component analyses (PCA) of multivariate characterization (MVC) characterize PPs, the physicochemical properties of organic solvents [100], Lewis acids in organic synthesis [101], amines in the Willgerodt Kindler reaction [102], and aldehyde/ketones [103].

These PPs are now heavily used in their laboratory to explore the realm and limits of new organic reactions. PPs of amino acids may be suitable for instance for screening of peptides [104]. The expansion of PPs for many aromatic substituents for subsequent uses has been the aim of researchers, and unfortunately, it is very difficult to find experimental information evaluated in a coordinated manner on a large number of substituents. Therefore, they should use the next best kind of data, famous and broadly used physicochemical parameters that are accessible for a large number of substituents.

The empirical parameter used to characterize a class of monosubstituted benzenes were Π, MR, σ_m, σ_p [92, 105], and the Verloop descriptors L and B_1–B_4 [106]. The Verloop parameters B_1–B_4, derived from STERIMOL calculations, are normally listed in order of magnitude improvement. Researchers attempt to choose the variables to define steric bulk (MR), hydrophobicity (Π), the shape of each substituent (Verloop parameters), and electronic properties (sigmas).

In this case, they knew that there are three groups of variables: hydrophobicity/bulk, electronic, and size.

From the numeric amounts of the loadings, it is shown that the first component is significantly connected to the steric bulk and hydrophobicity because the length, molecular refractivity, and Π have the largest contributions. The second component is dominated by the two electronic descriptors, σ_m and σ_p, while the third component is again mainly hydrophobicity (Π) but also shape since L and B_1–B_4 (Verloop parameters) [106] have relatively large contributions.

Since biological sieving of chemical substances is both expensive and time-consuming, it is essential to expand an instrument for the statistical design of the compounds in a filtering experiment. The main features are heavily appropriate for this purpose because they are few and orthogonal.

2.2 Descriptors for Nano-QSPR/QSAR

Over the past few decades, nano-based technology has become one of the top research areas in all fields of science and technology. A wide variety of consumer products are at the nanoscale, typically defined by all species having at least one diameter of 100 nm or less. Currently, nanotechnology has integrated various fields including biomedicine, pharmaceutical industry, food industry, environmental protection, solar batteries, energy, information and communication, heavy industry, consumer goods, and so on. However, it seems that we are only at the beginning of the “nano-industrial revolution.” Because of the unique electrical as well as optical, magnetic, thermal, and chemical properties of nanomaterials, the range of their possible applications is likely to expand rapidly.

Some recent papers report obvious evident toxicity of selected nanoparticles and highlight potential risk associated with the development of nano-engineering. Currently, there are many gaps in nanomaterial data. Predictive nano-QSAR/QSPR is one of the most promising methods used by chem informaticians to extrapolate the activity/property of nanomaterials. We believe that some of the missing data that are crucial for environmental risk assessment can be obtained using computational chemistry, saving the time and cost of conducting experiments. It is worth noting that the nano-QSPR/QSAR approach should be employed to predict not only activity responses (e.g., toxicity) but also many important physicochemical properties (e.g., water solubility, n-octanol/water partition coefficient, vapor pressure). These physicochemical properties affect the absorption, distribution, and metabolism of the compound in the organism, as well as environmental transport and the fate.

In nano-QSPR/QSAR modeling, one of the important parameters for building a validate model is suitable descriptors. In general, there are more than 5000 different descriptors for the characterization of molecular structure from zero to four dimensional (0D–4D). Only a few of traditional descriptors can characterize nanostructures. There are some reports that [107, 108] the existing descriptors are not enough to express the specific physical and chemical properties of nanoparticles. Therefore, new and more suitable types of descriptors for characterizing of nanoparticles should be developed.

Even though the computational features used for QSPR/QSAR modeling, experimentally derived features may also be employed as descriptors for nano-QSARs development (Fig. 2.3). The experimental descriptors seem to be especially useful for expressing size distribution, aggregation mode, shape, porosity, and surface disorder. Moreover, the combination of experimental results with a numerical approach can be used to define a new descriptor. For instance, images obtained by scanning electron microscopy (SEM), transmission electron microscopy (TEM), or atomic force microscopy (AFM) might be processed with new chemometric methods of image analysis. This means that first a series of pictures of different particles of a nanostructure should be taken. Then, the images must be numerically averaged and converted into a matrix containing numerical values that correspond to each pixel's grayscale intensity or red, green, and blue (RGB) color value. The other descriptors can be produced based on the matrix (i.e., the shape descriptor can be obtained as the sum of the nonzero elements in the matrix; the porosity as the sum of the relative differences between each pixel and its "neighbors," etc.) [109].

Undoubtedly, proper characterization of nanoparticle structure is currently one of the most challenging tasks in nano-QSAR. Although more than five thousand QSAR descriptors have been defined until now, they may be insufficient to express the supramolecular phenomena governing the unusual activity/property of nanomaterials. Consequently, much more effort is needed in this area.

2.3 SMILES and Quasi-SMILES Descriptors

The CORrelation And Logic (CORAL) software (http://www.insilico.eu/coral/) was developed by Alla Toropova and Andrey Toropov used to build up QSPR/QSAR models using Simplified Molecular Input Line Entry System (SMILES) [61, 111–116] and quasi-SMILES descriptors. SMILES is a chemical notation system designed by Weininger et al. [117, 118]. According to the principles of molecular graph theory, SMILES uses a very small, natural grammar to specify precise structural features. The SMILES symbol system is also suitable for high-speed machine processing [119, 120].

Over the last two decades, there have been numerous reports on the QSAR/QSPR modeling of nanomaterials and other compounds using CORAL software. This approach provides simple representation of molecular structures. There are defined equivalences between the representation of molecular structure using diagrams and

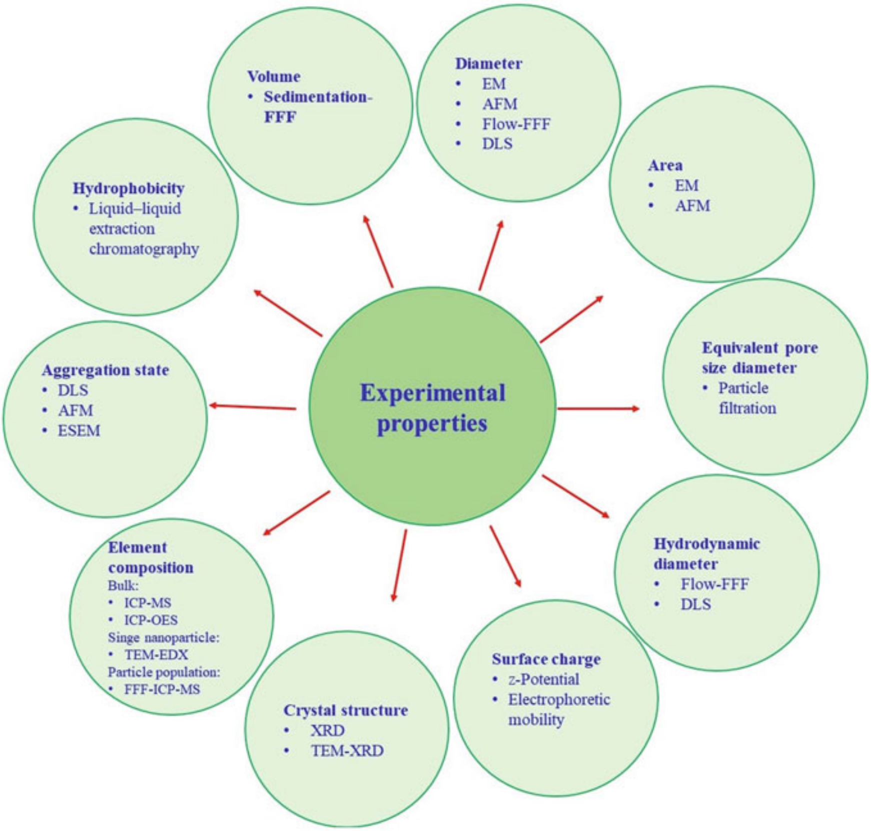


Fig. 2.3 Experimental characteristics as descriptors in nano-QSAR research [110]

the SMILES symbol. However, one should also be aware of their significant differences [121]. The SMILES can be produced by popular software such as ChemSketch, Biovia, and Chem Draw [122].

The prediction of activity/property of nanomaterials can be predicted by SMILES [123–125]. Quasi-SMILES is an alternative of SMILES-based optimal descriptors to build up predictive models for nanomaterials and other materials by consideration of the experimental conditions. Quasi-SMILES may be eclectic condition [126, 127] or combination of SMILES and eclectic conditions [128, 129]. The continuous eclectic conditions can be normalized by the following equation for assigning codes:

$$\mathrm{Norm}(P_i) = \frac{\min(P_i) + P_i}{\min(P_i) + \max(P_i)} \tag{2.18}$$

P_i is its value of physicochemical parameter P, $\min(P_i)$ is minimum value of P and $\max(P_i)$ indicates maximum value of P.

Table 2.2 Distinction of standardized physiochemical features into classes 1–9 according to its value

Norm value	Class
Norm(P) > 0.9	9
0.8 < Norm(P) < 0.9	8
0.7 < Norm(P) < 0.8	7
0.7 < Norm(P) < 0.6	6
0.6 < Norm(P) < 0.5	5
0.5 < Norm(P) < 0.4	4
0.4 < Norm(P) < 0.3	3
0.3 < Norm(P) < 0.2	2
0.2 < Norm(P) < 0.1	1
Norm(P) < 0.1	0

According to Table 2.2, the number of unique values in each parameter was less than 10; therefore, the quasi-SMILES descriptors representations could be coded by assigning a number between zero and nine in a single character.

2.3.1 Quasi-SMILES Examples in Peer-Reviewed Papers

Table 2.3 shows an example of the construction codes for the quasi-SMILES. Based on the data shown in Table 2.3, the quasi-SMILES can be generated, which can be used to build a model according to the optimal descriptors. Table 2.4 indicates some examples for quasi-SMILES generated by codes shown in Table 2.3.

The new reported QSPR analysis of MOFs by Ahmadi et al. is application of quasi-SMILES parameters including Brunauer, Emmett, and Teller (BET) specific surface area and pore volume, pressure, and temperature for prediction of CO_2 adsorption of MOFs [128]. Tables 2.5 and 2.6 show the eclectic data range and quasi-SMILES codes for them, respectively.

In the code-2019 of CORAL software for quasi-SMILES groups of symbols %10–%99 (reserved for representation of complex systems of rings for usual SMILES) were applied as codes for the quasi-SMILES (Table 2.6). The disadvantage of this version of quasi-SMILES is the difficulty of interpretation of results by a user.

Further development of the CORAL software (CORAL-2020) allows the display of experimental conditions through groups of symbols enclosed in parentheses. Table 2.7 shows the comparison codes in the last version (CORAL-2020) and old version of CORAL for creating quasi-SMILES in recently proposed models for the mutagenic potential. One can see codes-2020 are quite transparent and consequently are more convenient for a user. As is clearly evident, CORAL-2020 codes are quite transparent and thus more user-friendly.

Table 2.3 Codes used for the cell line, method, time exposition, concentration, size of nanoparticles, and type of metal oxide to convert various information of experimental data into quasi-SMILES [126]

Feature	Value or type	Code	Feature	Value or type	Code
Cell line	MCF-7	H	Normalized nanoparticles size	0.2 < Norm(size) $\le$ 0.3	P
	HT-1080	I		0.3 < Norm(size) $\le$ 0.4	Q
	HepG-2	J		0.4 < Norm(size) $\le$ 0.5	R
	HT-29	K		0.5 < Norm(size) $\le$ 0.6	S
	PC-12	L		0.9 < Norm(size) $\le$ 1.0	T
Method	MTT	M	Metal oxide type	SnO_2	1
	NRU	N		MnO_2	2
Time exposition	24	X		ZnO	3
	48	Y		Bi_2O_3	4
	72	Z		NiO	5
Concentration (μg mL^{-1})	5	A		CeO_2	6
	10	B		SiO_2	7
	25	C		TiO_2	8
	50	D			
	100	E			
	200	F			

Toropov et al. reported the model of toxicity examined based on four eclectic data including three possible forms of silver nanoparticles (bare, coat, cons), organisms (*Daphnia magna* or Zebrafish), size (nm), and zeta-potential (mV) [131], where "bare" characterizes nanoparticles without any coating, coat (coating) demonstrates nanoparticles with a shell, and "cons" defines nanoparticles including coating material descriptors (Table 2.8).

2.4 Software for Generation of Molecular Descriptors

Over the last two decades, the growing interest in property/activity prediction has led to the release of many software products to the market and open-source domains for scientists working in the field of QSPR/QSAR modeling. Table 2.9 shows some popular software for calculating molecular descriptors. In addition, some of them are complex packages that also include modules for QSPR/QSAR modeling, statistical analysis, and data visualization.

Table 2.4 Some examples for quasi-SMILES produced by codes indicated in Table 2.3

Cell line	Method	Time exposition (h)	Concentration (μg mL^{-1})	Normalized NPs size	Metal oxide type	Quasi-SMILES	Cell viability (%)
MCF-7	MTT	24	10	$0.2 < \text{Norm(size)} \leq 0.3$	SnO_2	HMXBP1	94.5
MCF-7	MTT	24	10	$0.2 < \text{Norm(size)} \leq 0.3$	MnO_2	HMXBP2	91.0
HT-1080	NRU	24	25	$0.2 < \text{Norm(size)} \leq 0.3$	MnO_2	INXCP2	79.0
MCF-7	MTT	48	50	$0.2 < \text{Norm(size)} \leq 0.3$	ZnO	HMYDP3	5.0
HepG-2	NRU	72	5	$0.2 < \text{Norm(size)} \leq 0.3$	SiO_2	JNZAP7	95.7
PC-12	MTT	48	50	$0.4 < \text{Norm(size)} \leq 0.5$	TiO_2	LMYDR8	52.0
HT-1080	MTT	24	25	$0.5 < \text{Norm(size)} \leq 0.6$	NiO	IMXCS5	77.0
HT-29	MTT	24	100	$0.3 < \text{Norm(size)} \leq 0.4$	CeO_2	KMXEQ6	88.5
MCF-7	MTT	24	100	$0.3 < \text{Norm(size)} \leq 0.4$	NiO	HMXEQ5	51.7

Table 2.5 Lower and high levels of CO_2 capture capacity, BET, pore volume, pressure (bar), and temperature (K) [128]

	CO_2 capture capacity (mol/kg)	BET	Pore volume (cm^3/g)	Pressure (bar)	Temperature (K)
Low level	0.1	0	0.035	0.01	195
High level	54.5	6240	7.5	55	318

Table 2.6 Defined quasi-SMILES codes for eclectic conditions (BET-normalized, normalized pore volume normalized, pressure-normalized, and temperature-normalized) of CO_2 capture capacity of MOFs [128]

Normalized range	BET	Code-2019 for pore volume	Code-2019 for pressure	Code-2019 for temperature
$0 < \text{BET} - \text{normalized} \leq 0.1$	%10	%20	%30	%40
$0.1 < \text{BET} - \text{normalized} \leq 0.2$	%11	%21	%31	%41
$0.2 < \text{BET} - \text{normalized} \leq 0.3$	%12	%22	%32	%42
$0.3 < \text{BET} - \text{normalized} \leq 0.4$	%13	%23	%33	%43
$0.4 < \text{BET} - \text{normalized} \leq 0.5$	%14	%24	%34	%44
$0.5 < \text{BET} - \text{normalized} \leq 0.6$	%15	%25	%35	%45
$0.6 < \text{BET} - \text{normalized} \leq 0.7$	%16	%26	%36	%46
$0.7 < \text{BET} - \text{normalized} \leq 0.8$	%17	%27	%37	%47
$0.8 < \text{BET} - \text{normalized} \leq 0.9$	%18	%28	%38	%48
$0.9 < \text{BET} - \text{normalized} \leq 1$	%19	%29	%39	%49

2.5 Conclusion and Future Direction

Molecular descriptors are a critical component of the methodological toolbox used to study quantitative structure–property/activity relationship (QSPR/QSAR) modeling and are widely used to describe the structures of chemical compounds for design of new compounds. The predictive and reliable QSPR/QSAR models depend on accurate descriptors, as accurate predictions can save the time and cost needed to design new compounds with the desired property/activity.

In this chapter, the main classes of theoretical molecular descriptors including 0D, 1D, 2D, 3D, and 4D descriptors are described. The most significant progress

Table 2.7 Definition of eclectic condition for the definition of quasi-SMILES [130]

	Condition	Code-2019	Code-2020
Coating	TA100	%10	[TA100]
	TA98	%11	[TA98]
	20-nm citrate	%12	[20cit]
	20-nm PVP	%13	[20PVP]
	50-nm citrate	%14	[50cit]
	50-nm PVP	%15	[50PVP]
	100-nm citrate	%16	[100cit]
Doses (µg/plate)	100-nm PVP	%17	[100PVP]
	0.0	%18	[d0.0]
	6.3	%19	[d6.3]
	12.5	%20	[d12.5]
	25	%21	[d25]
	50	%22	[d50]
	100	%23	[d100]

Table 2.8 Indicates some quasi-SMILES used to generate nano-QSAR model for pLC$_{50}$ [131]

Status of nanoparticles	Organisms	Size (nm)	Zeta-potential (mV)	Quasi-SMILES
nanoparticles without any coating	Daphnia magna	17.150–21.700	− 8.480 to − 5.050	[Bare][Daph][s%14][z%25]
NPs without any coating	Daphnia magna	12.600–17.150	− 25.630 to − 22.200	[Bare][Daph][s%13][z%20]
NPs with a shell	Daphnia magna	53.550–58.100	− 11.910 to − 8.480	[Daph][s%22][z%24]
NPs including coating material descriptors	Daphnia magna	21.700–26.250	− 11.910 to − 8.480	[Daph][s%15][z% 24]
NPs without any coating	Zebrafish	135.450–140.000	− 22.200 to − 18.770	[Bare][Fish][s%40][z%21]
NPs with a shell	Zebrafish	44.450–49.000	− 25.630 to − 22.200	[Fish][s%20][z%20]

over the last few years in chemometrics, cheminformatics, and bioinformatics has led to new strategies for finding new molecular descriptors. Here, some of the most common molecular descriptors and some new molecular descriptors especially for design and QSPR/QSAR modeling of nanocomposites have been highlighted.

In nano-QSPR/QSAR modeling, the data in many different publications are small and not ready enough for model building. In addition, nanomaterials exhibit high complexity and heterogeneity in their structures, which makes data collection and processing more challenging compared to traditional QSPR/QSAR. Quasi-SMILES descriptors are one of the solutions to this challenge and have been introduced as new descriptors combining SMILES and eclectic conditions. These novel descriptors provide transparent interpretation equation models with correlation weights calculated by Monte Carlo optimization using CORAL software.

Finally, a list of the most commonly used software packages for calculating molecular descriptors is reviewed here.

Table 2.9 List of software packages for the calculation of molecular descriptors

Name	Organization/institution	Availability	Descriptors	Platform/license
RDKit	GitHub	https://github.com/rdkit	> 200	Windows/Linux/Mac (freeware)
PaDELPy	University of Massachusetts Lowell	https://github.com/ecrl/padelpy	> 2500	Windows/Linux/Mac (freeware)
ADAPT	Pennsylvania State University	http://research.chem.psu.edu/pcjgroup/adapt.html	> 260	Unix/Linux (freeware)
ADMET	Simulations Plus, Inc	http://www.simulations-plus.com/	297	Windows (commercial)
Predictor™ CODESSA	Semichem	http://www.semichem.com/codessa/default.php	> 600	Windows/Linux (commercial)
DRAGON	Talete SRL	http://www.talete.mi.it/products/dragon_description.htm	4885	Windows/Linux (commercial)
EPISUITE™	EPA	http://www.epa.gov/opptintr/exposure/pubs/episuite.htm	20	Windows (freeware)
MOE	Chemical Computing Group	http://www.chemcomp.com/software-moe2009.htm	> 300	Windows/Linux/SGI/MAC/Sun (freeware)

(continued)

Table 2.9 (continued)

Name	Organization/institution	Availability	Descriptors	Platform/license
Molconn-Z™	EduSoft	http://www.edusoft-lc.com/molconn/	327	Windows/Unix/MAC (commercial)
MOLD	NCTR/FDA	http://www.fda.gov/ScienceResearch/BioinformaticsTools/Mold2/default.htm	777	Windows (freeware)
MOLGEN	University of Bayreuth	http://www.molgen.de/?src¼documents/molgenqspr.html	707	Windows (commercial
PowerMV	NISS	https://www.niss.org/research/software/powermv	> 1000	Windows (freeware)
Sarchitect™	Strand Life Sciences	http://www.strandls.com/sarchitect/index.html	1084	Windows/Linux (commercial)
SciQSAR™	SciMatics	http://www.scimatics.com/jsp/qsar/QSARIS.jsp	> 600	Windows (commercial)
Alvadesc	Alvascience	https://www.alvascience.com/alvadesc/	> 6000	Windows/Linux/MAC (commercial)
CORAL	Istituto di Ricerche Farmacologiche Mario Negri	http://www.insilico.eu/coral/SOFTWARECORAL.html	> 1000	Windows (freeware)

References

1. Rocke AJ (1981) Br J Hist Sci 14:27–57. https://doi.org/10.1017/S0007087400018276
2. Kekulé A (858) Liebigsder Chemie JA 106:129–159. https://doi.org/10.1002/jlac.18581060202
3. Brown AC, Fraser TR (1868) Eearth Environ Sci Trans Roy Soc 25:151–203. https://doi.org/10.1017/S0080456800028155
4. Richardson B (1869) Med Times Gazzette (ii), pp 703–706
5. Körner W (1874) Gazz Chim It 4:242
6. Richet M (1893) Compt Rend Soc Biol (Paris) 45:775–776
7. Hammett LP (1937) J Am Chem Soc 59:96–103. https://doi.org/10.1021/ja01280a022
8. Ghasemi J, Ahmadi S (2007) Ann Chim 97:69–83. DOI:https://doi.org/10.1002/adic.200690087
9. Ahmadi S, Mardinia F, Azimi N, Qomi M, Balali E (2019) J Mol Struct 1181:305–311. https://doi.org/10.1016/j.molstruc.2018.12.089
10. Ahmadi S, Ghanbari H, Lotfi S, Azimi N (2021) Mol Divers 25:87–97. https://doi.org/10.1007/s11030-019-10026-9
11. Javidfar M, Ahmadi S (2020) SAR QSAR Environ Res 31:717–739. https://doi.org/10.1080/1062936X.2020.1806922
12. Ahmadi S (2012) Macroheterocycles 5:23–31. https://doi.org/10.6060/mhc2012.110734a
13. Ahmadi S, Babaee E (2014) J Incl Phenom Macro 79:141–149. https://doi.org/10.1007/s10847-013-0337-7
14. Ahmadi S, Deligeorgiev T, Vasilev A, Kubista M (2012) Russ J Phys Chem A 86:1974–1981. https://doi.org/10.1134/S0036024412130201
15. Ghasemi JB, Ahmadi S, Brown S (2011) Environ Chem Lett 9:87–96. https://doi.org/10.1007/s10311-009-0251-9
16. Ahmadi S, Khazaei MR, Abdolmaleki A (2014) Med Chem Res 23:1148–1161. https://doi.org/10.1007/s00044-013-0716-z
17. Ahmadi S, Habibpour E (2017) Anti-Cancer Agent Med Chem 17:552–565. https://doi.org/10.2174/1871520611009010001
18. Ahmadi S (2012) J Incl Phenom Macro 74:57–66. https://doi.org/10.1007/s10847-010-9881-6
19. Ghasemi JB, Ahmadi S, Ayati M (2010) Macroheterocycles 3:234–242. https://doi.org/10.6060/mhc2010.4.234
20. Tropsha A, Wang S (2007) In: Bourne H, Horuk R, Kuhnke J, Michel H (eds) GPCRs: from deorphanization to lead structure identification. Ernst Schering Foundation symposium proceedings, vol 2006/2. Springer, Berlin, Heidelberg, pp 49–74. https://doi.org/10.1007/2789_2006_003
21. Ahmadi S, Ganji S (2016) Curr Drug Discov Technol 13:232–253. https://doi.org/10.2174/1570163813666160725114241
22. Lotfi S, Ahmadi S, Kumar P (2021) J Mol Liq 338:116465. https://doi.org/10.1016/j.molliq.2021.116465
23. Habibpour E, Ahmadi S (2017) Curr Comput-Aid Drug Des 13:143–159. https://doi.org/10.2174/1573409913666170124100810
24. Lotfi S, Ahmadi S, Kumar P (2021) RSC Adv 11:33849–33857. https://doi.org/10.1039/D1RA06861J
25. Willett P, Barnard JM, Downs GM (1998) J Che Inf Comp Sci 38:983–996. https://doi.org/10.1021/ci9800211
26. Suhachev D, Pivina T, Shlyapochnikov V, Petrov E, Palyulin V, Zefirov N (1993) Dokl RAN 328:50–57
27. Harary F (1971) Graph theory, 2nd printing. Addison-Wesley, Reading, MA
28. Roy K (2004) Mol Divers 8:321–323. https://doi.org/10.1023/b:modi.0000047519.35591.b7
29. Wiener H (1947) J Am Chem Soc 69:17–20. https://doi.org/10.1021/ja01193a005
30. Randić M (1993) Chem Phys Lett 211:478–483. https://doi.org/10.1016/0009-2614(93)87094-J

31. Nikolić S, Trinajstić N, Randić M (2001) Chem Phys Lett 333:319–321. https://doi.org/10.1016/S0009-2614(00)01367-1
32. Li X-h, Li Z-g, Hu M-l (2003) J Mol Graph Model 22:161–172. https://doi.org/10.1016/S1093-3263(03)00157-8
33. Randic M (1975) J Am Chem Soc 97:6609–6615. https://doi.org/10.1021/ja00856a001
34. Kier LB, Hall LH (1986) In: Molecular connectivity in structure-activity analysis. Research studies. Wiley, Letchworth, Hertfordshire, England, New York, p 262. https://doi.org/10.1002/jps.2600760325
35. Randić M (1991) J Mat Chem 7:155–168. https://doi.org/10.1007/BF01200821
36. Randić M (2001) J Mol Graph Model 20:19–35. https://doi.org/10.1016/S1093-3263(01)00098-5
37. Kier L (1986) Acta Pharm Jugosl 36:171–188. https://doi.org/10.1002/med.2610070404
38. Balaban AT (1982) Chem Phys Lett 89:399–404. https://doi.org/10.1016/0009-2614(82)80009-2
39. Gutman I, Das KC (2004) MATCH Commun Math Comput Chem 50:83–92. https://match.pmf.kg.ac.rs/electronic_versions/Match50/match50_83-92.pdf
40. Furtula B, Graovac A, Vukičević D (2010) J Mat Chem 48:370–380. https://doi.org/10.1007/s10910-010-9677-3
41. Hosoya H (1971) B Chem Soc Jpn 44:2332–2339. https://doi.org/10.1246/bcsj.44.2332
42. Randić M, Zupan J (2001) J Chem Inf Comp Sci 41:550–560. https://doi.org/10.1021/ci000095o
43. Moreau G, Broto P (1980) Nouv J Chim 4(6):359–360
44. Gutman I (1994) Graph Theory Notes NY 27(9):9–15
45. Melville JL, Hirst JD (2007) J Chem Inf Model 47:626–634. https://doi.org/10.1021/ci6004178
46. Spowage BM, Bruce CL, Hirst JD (2009) J Cheminform 1:1–13. https://doi.org/10.1186/1758-2946-1-22
47. Ghose AK, Crippen GM (1986) J Comput Chem 7:565–577. https://doi.org/10.1002/jcc.540070419
48. Arnott JA, Kumar R, Planey SL (2013) J Appl Biopharm Pharmacokinet 1:31–36. http://creativecommons.org/licenses/by-nc/3.0/
49. Winiwarter S, Ax F, Lennernäs H, Hallberg A, Pettersson C, Karlén A (2003) J Mol Graph Model 21:273–287. https://doi.org/10.1016/S1093-3263(02)00163-8
50. Cramer RD, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5959–5967. https://doi.org/10.1021/ja00226a005
51. Seel M, Turner DB, Willett P (1999) Quant Struct-Act Relat 18:245–252. https://doi.org/10.1002/(SICI)1521-3838
52. Doweyko AM (1988) J Med Chem 31:1396–1406. https://doi.org/10.1021/jm00402a025
53. Kuz'min VE, Artemenko AG, Kovdienko NA, Tetko IV, Livingstone DJ (2000) J Mol Model 6:517–526. https://doi.org/10.1007/s0089400060517
54. Todeschini R, Consonni V (2008) Handbook of molecular descriptors. Wiley, p 667. https://doi.org/10.1002/9783527613106
55. Soltzberg LJ, Wilkins CL (1977) J Am Chem Soc 99:439–443. https://doi.org/10.1021/ja00444a021
56. Wagener M, Sadowski J, Gasteiger J (1995) J Am Chem Soc 117:7769–7775. https://doi.org/10.1021/ja00134a023
57. Vedani A, Dobler M (2000) In: Jucker E (ed) Progress in drug research, vol 55. Birkhäuser, Basel, pp 105–135. https://doi.org/10.1007/978-3-0348-8385-6_4
58. Easton RE, Giesen DJ, Welch A, Cramer CJ, Truhlar DG (1996) Theor Chim Acta 93:281–301. https://doi.org/10.1007/BF01127507
59. Kostal J, Voutchkova-Kostal A, Anastas PT, Zimmerman JB (2015) Proc Natl Acad Sci 112:6289–6294. https://doi.org/10.1073/pnas.1314991111
60. Lynch BJ, Truhlar DG (2004) Theor Chem Acc 111:335–344. https://doi.org/10.1007/s00214-003-0518-3

61. Azimi A, Ahmadi S, Kumar A, Qomi M, Almasirad A (2022) Polycycl Aromat Comp 1–21. https://doi.org/10.1080/10406638.2022.2067194
62. Cocchi M, Menziani MC, De Benedetti PG, Cruciani G (1992) Chemometr Intell Lab 14:209–224. https://doi.org/10.1016/0169-7439(92)80105-D
63. Franke R (1984) Pharm Libr, vol 7. Elsevier, Amsterdam, p 412
64. Bodor N, Gabanyi Z, Wong CK (1989) J Am Chem Soc 111:3783–3786. https://doi.org/10.1021/ja00193a003
65. Buydens L, Massart DL, Geerlings P (1983) Anal Chem 55:738–744. https://doi.org/10.1021/ac00255a034
66. Klopman G, Iroff LD (1981) J Comput Chem 2:157–160. https://doi.org/10.1002/jcc.540020204
67. Zhou Z, Parr RG (1990) J Am Chem Soc 112:5720–5724. https://doi.org/10.1021/ja00171a007
68. Ośmiałowski K, Halkiewicz J, Radecki A, Kaliszan R (1985) J Chromatogr A 346:53–60. https://doi.org/10.1016/S0021-9673(00)90493-X
69. Fukui K (1970) In: Orientation and Stereoselection. Fortschritte der Chemischen Forschung, vol 15/1. Springer, Berlin, Heidelberg, pp 1–85. https://doi.org/10.1007/BFb0051113
70. Sklenar H, Jäger J (1979) Int J Quantum Chem 16:467–484. https://doi.org/10.1002/qua.560160306
71. Tuppurainen K, Lötjönen S, Laatikainen R, Vartiainen T, Maran U, Strandberg M, Tamm T (1991) Mutat Res Fundam Mol Mech 247:97–102. https://doi.org/10.1016/0027-5107(91)90037-O
72. Becker H (1978) J Prakt Chem 320:879–880. https://doi.org/10.1002/prac.19783200525
73. Lewis D, Ioannides C, Parke D (1994) Xenobiotica 24:401–408. https://doi.org/10.3109/00498259409043243
74. Pearson RG (1989) J Org Chem 54:1423–1430. https://doi.org/10.1021/jo00267a034
75. Prabhakar YS (1991) Drug Des Deliv 7:227–239
76. Kurtz HA, Stewart JJ, Dieter KM (1990) J Comput Chem 11:82–87. https://doi.org/10.1002/jcc.540110110
77. Cammarata A (1967) J Med Chem 10:525–527. https://doi.org/10.1021/jm00316a004
78. Leo A, Hansch C, Church C (1969) J Med Chem 12:766–771. https://doi.org/10.1021/jm00305a010
79. Hansch C, Coats E (1970) J Pharm Sci 59:731–743. https://doi.org/10.1002/jps.2600590602
80. Lewis DF (1987) J Comput Chem 8:1084–1089. https://doi.org/10.1002/jcc.540080803
81. Cartier A, Rivail J-L (1987) Chemometr Intell Lab 1:335–347. https://doi.org/10.1016/0169-7439(87)80039-4
82. Gaudio AC, Korolkovas A, Takahata Y (1994) J Pharm Sci 83:1110–1115. https://doi.org/10.1002/jps.2600830809
83. Grunenberg J, Herges R (1995) J Chem Inf Comp Sci 35:905–911. https://doi.org/10.1021/ci00027a018
84. Kikuchi O (1987) Quant Struct-Act Relat 6:179–184. https://doi.org/10.1002/qsar.19870060406
85. Grüber C, Buss V (1989) Chemosphere 19:1595–1609. https://doi.org/10.1016/0045-6535(89)90503-1
86. Saura-Calixto F, Garcia-Raso A, Raso M (1984) J Chromatogr Sci 22:22–26. https://doi.org/10.1093/chromsci/22.1.22
87. Shusterman A (1991) ChemTech 21(10):624–627
88. Brusick DJ, Vogel EW, Nivard MJ, Klopman G, Rosenkranz HS, Enslein K, Gombar VK, Blake BW, Debnath AK, Shusterman AJ, de Compadre RL (1994) Mutat Res 305:321–323
89. Trapani G, Carotti A, Franco M, Latrofa A, Genchi G, Liso G (1993) Eur J Med Chem 28:13–21. https://doi.org/10.1016/0223-5234(93)90074-O
90. Ebert C, Linda P, Alunni S, Clementi S, Cruciani G, Santini S (1990) Gazz Chim Ital 120:29
91. Skagerberg B, Bonelli D, Clementi S, Cruciani G, Ebert C (1989) Quant Struct-Act Relat 8:32–38. https://doi.org/10.1002/qsar.19890080105

92. Flynn GL (1980) J Pharm Sci 69:1109–1109. https://doi.org/10.1002/jps.2600690938
93. Taft RW, Topsom R (1987) Prog Phys Org Chem 16:1–83
94. Ewing DF (1979) Org Magn Reson 12:499–524. https://doi.org/10.1002/mrc.1270120902
95. Waterbeemd H Van de, Testa B (1987) In: Advances in drug research, vol 16. Academic, London, pp 87–227
96. Yang GZ, Lien EJ, Guo ZR (1986) Quant Struct-Act Relat 5:12–18. https://doi.org/10.1002/qsar.19860050104
97. Kim KH, Martin YC (1991). In: Silipo C, Vittoria A (eds) QSAR: rational approaches to the design of bioactive compounds. Elsevier, Amsterdam, pp 151–154
98. Sjöström M, Wold S (1985) J Mol Evol 22:272–277. https://doi.org/10.1007/BF02099756
99. Sjöström M, Wold S, Wieslander A, Rilfors L (1987) EMBO J 6:823–831. https://doi.org/10.1002/j.1460-2075.1987.tb04825.x
100. Carlson R, Lundstedt T, Albano C (1985) Acta Chem Scand B 39:79–91. https://doi.org/10.3891/acta.chem.scand.39b-0079
101. Carlson R, Lundstedt T, Nordahl Å, Prochazka M (1986) Acta Chem Scand B 40:522–533. https://doi.org/10.3891/acta.chem.scand.40b-0522
102. Lundstedt T, Carlson R, Shabana R (1987) Acta Chem Scand B 41:157–163. https://doi.org/10.3891/acta.chem.scand.41b-0157
103. Carlson R, Prochazka M, Lundstedt T (1988) Acta Chem Scand B Org Chem Biochem 42:145–156. https://doi.org/10.3891/acta.chem.scand.42b-0145
104. Hellberg S, Sjostrom M, Skagerberg B, Wikstrom C, Wold S (1987) Acta Pharm Jugsl 37:53–65. https://doi.org/10.1021/jm00390a003
105. Hansch C, Leo A, Unger SH, Kim KH, Nikaitani D, Lien EJ (1973) J Med Chem 16:1207–1216. https://doi.org/10.1021/jm00269a003
106. Verloop A, Hoogenstraaten W, Tipker J (1976) In: Ariens EJ (ed) Drug design. Academic, New York
107. Rybińska-Fryca A, Mikolajczyk A, Puzyn T (2020) Nanoscale 12:20669–20676. https://doi.org/10.1039/D0NR05220E
108. Richarz A-N, Avramopoulos A, Benfenati E, Gajewicz A, Golbamaki Bakhtyari N, Leonis G, Marchese Robinson RL, Papadopoulos MG, Cronin MTD, Puzyn T (2017) In: Tran L, Bañares M, Rallo R (eds) Modelling the toxicity of nanoparticles. Advances in experimental medicine and biology, vol 947. Springer, Cham, pp 303–324. https://doi.org/10.1007/978-3-319-47754-1_10
109. Puzyn T, Gajewicz A, Leszczynska D, Leszczynski J (2010) In: Puzyn T, Leszczynski J, Cronin M (eds) Recent advances in QSAR studies. Challenges and advances in computational chemistry and physics, vol 8. Springer, Dordrecht, pp 383–409. https://doi.org/10.1007/978-1-4020-9783-6_14
110. Hassellöv M, Readman JW, Ranville JF, Tiede K (2008) Ecotoxicology 17:344–361. https://doi.org/10.1007/s10646-008-0225-x
111. Ahmadi S, Lotfi S, Kumar P (2020) SAR QSAR Environ Res 31:935–950. https://doi.org/10.1080/1062936X.2020.1842495
112. Ghiasi T, Ahmadi S, Ahmadi E, Talei Bavil Olyai M, Khodadadi Z (2021) SAR QSAR Environ Res 32:495–520. https://doi.org/10.1080/1062936X.2021.1925344
113. Ahmadi S, Lotfi S, Afshari S, Kumar P, Ghasemi E (2021) SAR QSAR Environ Res 32:1013–1031. https://doi.org/10.1080/1062936X.2021.2003429
114. Ahmadi S, Lotfi S, Kumar P (2022) Toxicol Mech Method 32:302–312. https://doi.org/10.1080/15376516.2021.2000686
115. Ahmadi S, Moradi Z, Kumar A, Almasirad A (2022) J Recept Signal Transduct 42:361–372. https://doi.org/10.1080/10799893.2021.1957932
116. Lotfi S, Ahmadi S, Kumar P (2022) RSC Adv 12:24988–24997. https://doi.org/10.1039/D2RA03936B
117. Weininger D (1988) J Chem Inf Comp Sci 28:31–36. https://doi.org/10.1021/ci00057a005
118. Weininger D, Weininger A, Weininger JL (1989) J Chem Inf Comp Sci 29:97–101. https://doi.org/10.1021/ci00062a008

119. Pinheiro GA, Mucelini J, Soares MD, Prati RC, Da Silva JL, Quiles MG (2020) J Phys Chem A 124:9854–9866. https://doi.org/10.1021/acs.jpca.0c05969
120. Lotfi S, Ahmadi S, Zohrabi P (2020) Struct Chem 31:2257–2270. https://doi.org/10.1007/s11224-020-01568-y
121. Toropov A, Toropova A, Martyanov S, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2011) Chemometr Intell Lab 109:94–100. https://doi.org/10.1016/j.chemolab.2011.07.008
122. Ahmadi S, Mehrabi M, Rezaei S, Mardafkan N (2019) J Mol Struct 1191:165–174. https://doi.org/10.1016/j.molstruc.2019.04.103
123. Ahmadi S, Akbari A (2018) SAR QSAR Environ Res 29:895–909. https://doi.org/10.1080/1062936X.2018.1526821
124. Heidari A, Fatemi MH (2017) J Chin Chem Soc-Taip 64:289–295. https://doi.org/10.1002/jccs.201600761
125. Kumar P, Kumar A (2020) SAR QSAR Environ Res 31:697–715. https://doi.org/10.1080/1062936X.2020.1806105
126. Ahmadi S (2020) Chemosphere 242:125192. https://doi.org/10.1016/j.chemosphere.2019.125192
127. Ahmadi S, Toropova AP, Toropov AA (2020) Nanotoxicology 14:1118–1126. https://doi.org/10.1080/17435390.2020.1808252
128. Ahmadi S, Ketabi S, Qomi M (2022) New J Chem 46:8827–8837. https://doi.org/10.1039/D2NJ00596D
129. Ahmadi S, Aghabeygi S, Farahmandjou M, Azimi N (2021) Struct Chem 32:1893–1905. https://doi.org/10.1007/s11224-021-01748-4
130. Toropov AA, Toropova AP (2019) Sci Total Environ 681:102–109. https://doi.org/10.1016/j.scitotenv.2019.05.114
131. Toropov AA, Kjeldsen F, Toropova AP (2022) Chemosphere 135086. https://doi.org/10.1016/j.chemosphere.2022.135086