25
Quantum-Chemical Descriptors in QSAR/QSPR Modeling: Achievements, Perspectives and Trends

Anna V. Gubskaya

25.1
Introduction

Quantitative structure–activity or property relationships (QSAR/QSPR) modeling has been used in the fields of medicinal (drug-design, toxicology), industrial, agricultural and environmental chemistry for more than 40 years. During the last decade it has also been successfully applied in biochemistry, molecular biology and material science. The concept of QSAR is based on the postulate that the structure of a molecule represented by selected molecular characteristics (descriptors) can be correlated to its biologic activity. Once such a correlation is established for the compounds with known biologic activity, it becomes possible using specified computational protocol to predict biologic activities for new or untested chemicals of the same class. The growing popularity of modern QSAR can be attributed to its ability to select compounds with desirable biologic response from combinatorial libraries containing thousands of molecules *in silico*, that is, without synthesis as well as time and labor-consuming screening.

Several excellent recent reviews published on QSAR, in addition to offering a historical overview, describe the most prominent trends in this field [1–3]. One of these trends is associated with the rapidly increasing amount of molecular descriptors represented by quantum-chemical or by various classical parameters [e.g., constitutional, topological, connectivity, weighted holistic invariant molecular descriptors (WHIM), etc.] that were designed and tested as potential variables for QSAR modeling. To date, the work of Todeschini *et al.* [4] is the most comprehensive source of QSAR descriptors. Quantum-chemical parameters represent a special class of molecular properties. They can be obtained from sophisticated *ab initio* calculations or by means of relatively inexpensive semiempirical methods, but in either case such calculations require more time and effort than those for one, two or three-dimensional classical parameters, which can be computed from molecular structures within a few minutes. However, in contrast to most classical descriptors, quantum chemical parameters are capable of expressing all the electronic and

Quantum Biochemistry. Edited by Chérif F. Matta

ISBN: 978-3-527-32322-7

geometric properties of the molecules being analyzed as well as their interactions. Therefore, in some cases the interpretation of quantum-chemical descriptors can provide much deeper insights into the nature of biologic or physicochemical mechanisms under consideration than that of classical descriptors.

To the best of the author's knowledge, the latest comprehensive review on quantum-chemical descriptors and their applications in QSAR studies was published by Karelson *et al.* [5] in 1996. Since then, significant progress has been achieved in computational hardware, the development of quantum mechanical methodologies such as density functional theory (DFT) [6, 7], the concepts of molecular quantum similarity measure and quantum topological molecular similarity [8] as well as the design of corresponding algorithms. Advances in combinatorial chemistry have given a new start to the use of machine-learning methods for the selection of the most relevant to specific bioactivity descriptors and to the development of more sophisticated QSAR models. The increasing complexity of investigations in life science-related fields has facilitated the process of generating and testing new molecular (including quantum-chemical) descriptors. As most biologic processes take place in aqueous media, the advantages of descriptors calculated by means of quantum-chemical approaches that account for specific and non-specific solvation effects are of prime importance. The present chapter focuses on QSAR/QSPR studies in biologic sciences carried out during past decade or so. It will cover the recent applications of quantum descriptors and the new conceptual and methodological trends associated with their use. The capability of quantum descriptors in predicting biologic activities and biologically important properties will be demonstrated.

25.2
Quantum-Chemical Methods and Descriptors

25.2.1
Quantum-Chemical Methods

Calculations of electronic properties of a molecule that have potential value to QSAR studies can be performed by various quantum-mechanical methods. These methods, represented by two major groups, *ab initio* and semiempirical methods, have been further classified and their methodological details, corresponding approximations as well as advantages of utilization have been described by Karelson *et al.* [5]. The authors mention no applications of density functional theory [6, 7] in QSAR studies, while in about one half of 80 original research articles reviewed in the present work DFT formalism was used to obtain descriptors for highly predictive QSAR/QSPR models. Among semiempirical methods, Austin model 1 (AM1) [9], modified neglect of differential overlap (MNDO) [10] and parametric model 3 (PM3) [11], known as evolution of MNDO parameterization, were chosen in 16, 4 and 13% of cases, respectively. Approximately 20% of the studies reviewed here were devoted to the development and applications of quantum similarity approaches. In several cases *ab initio* methods [12], namely, Hartree–Fock (12%) and Møller–Plesset theory of

second order (MP2) (3%), were used to calculate quantum descriptors. Some of these studies reconfirmed the conclusion made by Karelson *et al*. [5] that electronic descriptors as well as optimized geometrical parameters obtained from AM1 and PM3 calculations are more satisfactory than those from *ab initio* calculations carried out with insufficiently large basis set [13].

The development of DFT accelerated the utilization of electronic structure theory in determining molecular properties for biologically significant molecules. The DFT method belongs to the group of *ab initio* methods that allow calculations of quantum-chemical descriptors at a reasonable cost and with higher accuracy than that of semiempirical methods. QSAR models generated using quantum descriptors obtained by DFT were found to be more predictive than the models incorporating descriptors calculated by AM1 [14, 15] or PM3 [15] methods. The fact that DFT accounts for dynamic correlation effects makes it an attractive alternative to the Hartree–Fock (HF) method as well as the much more CPU-demanding post-HF methods: MP theory, coupled-cluster theory and configuration interaction approach [5]. In the commonly applied Kohn–Sham DFT formalism the functional of electron density, in addition to classical (kinetic and electrostatic) energy terms, includes contribution from exchange-correlation energy. Becke's three-parameter hybrid exchange functional and the Lee–Yang–Parr correlation functional (B3LYP) [16–18] is probably the most popular hybrid density functional used in QSAR-related DFT calculations at the time of writing. Omitting electron correlations in HF theory lowers the accuracy of computations in comparison with those carried out using DFT: statistical parameters of QSAR models obtained with descriptors computed at B3LYP/LANL2DZ and at HF/LANL2DZ levels showed the obvious advantage of DFT-based descriptors [19]. Several QSPR studies in material science demonstrated successful utilization of thermochemistry data calculated at the B3LYP/6-31G(d) level in predicting physicochemical properties of polymers [20–22]. Since density functional does not account for dispersion energy [23], its applicability might be limited in cases where contribution from dispersion interactions is expected or known to be significant.

Among all quantum-mechanical methods used for calculation of QSAR descriptors the quantum methodologies dealing with the principle of molecular similarity represent efficient tools for solving various chemistry-related problems. The molecular quantum similarity measures (MQSMs) approach developed by Carbó-Dorca and co-authors [24–27] establishes a quantitative measure of resemblance between two molecules based on their first-order density functions (DFs), constructed in a specific internal energy state. It was also possible to include one of extended DFs, namely kinetic energy DF, into MQSM formalism and to use it to correlate the antimalarial activity of two series of compounds [28]. In the framework of MQSM approach [27] a quantitative molecular similarity measure between two molecules (or molecular fragments [29]), A and B, described by density functions $\varrho_A(\boldsymbol{r})$ and $\varrho_B(\mathbf{r})$ is expressed as a direct volume integral:

$$Z_{AB}(\Omega) = \iint \varrho_A(\mathbf{r}_1)\Omega(\mathbf{r}_1, \mathbf{r}_2)\varrho_B(\mathbf{r}_2)\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 \tag{25.1}$$

where $\Omega(\mathbf{r}_1, \mathbf{r}_2)$ is a positive definite two-electron operator that acts as a weighting factor. The choice of this operator determines the type of QSM; for example, the identification of Ω with the Dirac delta function or with the Coulomb repulsion term produces overlap-like or Coulomb-like similarity measures [29], respectively. A quantitative measure is transformed into an absolute magnitude by means of the Carbó index given as:

$$C_{AB} = Z_{AB}(Z_{AA}Z_{BB})^{-1/2}. \tag{25.2}$$

The closer to unity the Carbó index is the more similar are the two compounds subject to comparison. All quantitative measures computed for molecular pairs can be transformed into Carbó indexes and represented as a matrix form whose columns are used as MQSM molecular descriptors [24, 26, 28]. Computations of Z_{AB} integral can be very time-consuming and thus the authors proposed the so-called atomic shell and promolecular approximations [30, 31] to reduce the MQSM computational costs without a significant loss of accuracy.

MQSM originally introduced by Carbó-Dorca provided a fundamental framework for the quantum-mechanical description of molecular similarity and stimulated the development of various computational methods and theories focused on this problem. Mezey and co-authors [32–35] have made a substantial contribution to shape similarity analysis, specifically to the shape analysis of electron density, $\varrho(\mathbf{r})$. The authors introduced the concepts of density domain and molecular isodensity contour surface and proposed two families of shape analysis techniques: the shape group methods [32] and T-hull method with its extensions [33]. Mezey also reviewed Münch and Reiss's proof that a unique mapping exists between the electron density of a subsystem and that of the entire system [35, 36]. The principles and methods of the molecular shape analysis were reflected in the new the molecular electron density "lego" assembler (MEDLA) [34] method for rapid "*ab initio* quality" computation of shapes for large molecules. The MEDLA approach is based on a simple electron density fragment additivity principle. The fragment density matrices (along with the basis set information) are stored in the MEDLA data bank and are used as fuzzy building blocks for the construction of electron densities for target molecules of any size. In the QShAR (here "Sh" stands for "shape") study of toxicological risk assessment of polycyclic aromatic hydrocarbons (PAHs) to *L. gibba*, utilization of similarity measures calculated by the MEDLA method provided excellent correlation coefficients [7, 34]. It is somewhat surprising that this interesting method had not yet been tested in a wider variety of cases. A conceptually similar (in terms of additivity and transferability) but methodologically alternative approach for computing the properties of large molecules not amenable to direct computation was presented by Matta [37]. This author described a real space approach to reconstructing the electron density of a large molecule from electron density fragments extracted from molds [37], using the quantum theory of atoms in molecules (QTAIM) developed by Bader [38] and members of his school [39–41].

Bader's QTAIM [38, 42] represents another promising approach to molecular similarity. Popelier and coworker [43–45] have proposed and extensively described

applications of a method called quantum topological molecular similarity (QTMS) – reviewed by Popelier in Chapter 24. Herein is provided only a brief description of it, supported by the most recent references. The AIM theory takes advantage of the topology of the electron density distribution in a molecule. The QTMS method probes the electron density of a molecule at critical points, specifically the points in 3D space where the gradient of electron density vanishes ($\nabla\varrho = 0$). The most chemically informative among the four types of critical points is the bond critical point (BCP), which is located on the intersection of the interatomic surface between bonded nuclei and the bond path linking these two nuclei. The BCP represents a quantum-chemical "signature" of a bond and can be identified by evaluating the Hessian matrix of the electron density at the location of the BCP. This second derivative matrix is composed of three eigenvalues ($\lambda_1, \lambda_2, \lambda_3$) and three associated eigenvectors. The first two eigenvalues represent the local curvatures that are perpendicular to the bond path and they must be negative, while the last eigenvalue is positive and corresponds to the curvature along the bond path. The properties evaluated at BCP are used as descriptors, that is, as measures of quantum topological similarity [46–53].

To date the QTAIM provides an essential link between the rigorous theory of quantum mechanics and the processes usually described by organic chemistry, that is, those that involve atoms as parts of larger molecular fragments. One of the applications of AIM approach in QSAR, the transferable atom equivalent method (TAE/RECON) [54], utilizes atomic contributions to generate electron-density-derived descriptors that approximate regular descriptors available through *ab initio* calculations. Mazza *et al.* [54] have employed this approach to model protein retention in ion-exchange systems.

25.2.2
Quantum-Chemical Descriptors: Classification, Updates

The choice of descriptors is crucial for obtaining a highly predictive QSAR model, whether they are chosen to be classical, electronic, that is, derived from quantum-mechanical calculations, or experimentally measured quantities or selected representatives of different types. For instance, to predict anticancer activity of carbocyclic analogs of nucleosides, Yao *et al.* [55] considered electronic descriptors such as charges, molecular orbital characteristics and polarity measures to account for the drug–receptor interaction effects and the computed octanol–water partition coefficient, solvent accessible surface areas and molecular volumes to model the drug delivery and size effects. The benefit of calculated parameters is that in contrast to experimental properties they are reproducible. They can be calculated for a set of compounds in question by defined software applying the same theoretical or methodological approximations. The errors associated with the assumptions needed for facilitating quantum calculations are considered to be constant within a series of related compounds and for most cases the direction of possible errors are known [5]. The additional advantage of quantum-chemical descriptors is that they allow characterization of an entire molecule as well as its fragments and substituents.

Table 25.1 gives a summary and classification of quantum-chemical descriptors available for contemporary QSAR/QSPR studies. The classification scheme in Table 25.1 closely reflects, combines and extends the classifications of quantum-chemical descriptors proposed by Karelson *et al.* [5] and later by Todeschini *et al.* [4]. Readers are also referred to these reviews for details about the descriptors of major groups, in terms of usage and physicochemical significance: atomic charges, molecular orbital characteristics, energy and polarity measures [5] as well as DFT-based descriptors (e.g., softness and hardness indices) [4]. Quantum-chemical measures that have recently been introduced in the literature in the context of potential application in QSAR studies are described below.

The special group of descriptors defined as electronic indices was derived by means of electronic indices methodology (EIM) to correlate structures of polycyclic aromatic hydrocarbons and their carcinogenic activity [56, 57]. This approach has been applied to investigate 5α-reductase inhibitory properties of benzquinolizin-3-ones [13]. The EIM approach uses concepts of density of states (DOS, i.e., the number of electronic states per energy unit) and local density of states (LDOS, i.e., density of states calculated over a specific region or atom) to estimate the contributions of the specific regions of the molecules to physicochemical or biologic response (s) [56]. The EIM indices are represented by the values of relative HOMO and HOMO-1 contributions (η_H) to the LDOS over the ring that contains the highest bond order and by the critical value of their energy separation (ΔH). The authors showed that QSAR models for carcinogenic activity of PAHs based on EIM indices exhibit about 80% of predictive power [57] and that the EIM approach performs remarkably well in constructing rules and patterns of classification for benzo[*c*]quinolizin-3-ones according to their biologic activity [13].

Clare and co-authors [58–61] have introduced new descriptors, the frontier orbital phase angles (FOPAs), which together with flip regression and orbital nodal orientation calculation methods were specifically designed for QSAR modeling of drug-like compounds containing five- or six-membered aromatic rings. The authors suggested that the FOPA parameters (e.g., $S2\Theta H$, $C2\Theta H$, $S4\Theta H$, $C4\Theta H$ variables in Table 25.1) affect activity because they approximate the actual orientation of the nodes of the orbitals in the compounds and that the π-like orbitals of aromatic substances presumably interact with π-like orbitals of the receptor [58]. This approach was successfully used to establish drug–receptor correlations for phenylalkylamine hallucinogens [58] and to model inhibitory activities of some carbonic anhydrase inhibitors [59], flavonoid analogs [60] and phenylisopropylamines [61].

Quantum similarity descriptors are represented by Hodgkin–Richards indices [62], the reactivity based similarity indices [63] and by the most often used Carbó indices. Interestingly, both Carbó and Hodgkin–Richards indices were used in a molecular quantum similarity (MQS) study [64] to assess the difference in information that can be obtained from conceptual DFT descriptors, specifically the electron density, the shape function, the Fukui functions and the local softness. In contrast to the Hodgkin–Richards index, the Carbó index clearly revealed that within the set of congeneric steroids the density function and local softness contain different chemical information while the shape function and the Fukui function are

Table 25.1 Classification of quantum-chemical descriptors.

Symbol[a)]	Name/definition
Energy measures	
E_T (TE)	Total energy [4, 5, 20–22, 84, 107, 109, 112, 124, 125, 128, 136, 139, 146]
E_e (EE)	Electronic energy [68, 112, 125, 128]
CCR	Core–core repulsion energy [125, 128]
E_b	Binding energy [5]
ΔH_f°	Heat of formation [4, 5, 112, 125]
$\Delta\Delta H_f^\circ$	Relative heat of formation [5, 152]
ΔH_{prot} (ΔE)	Energy of protonation given as the difference between the total energy of the protonated and neutral species [5, 143]
IP	Ionization potential [4, 5, 104, 139]
EA	Electron affinity [4, 5, 104]
$\chi_{MU,PU,\ etc.}$	Atom or molecular electronegativity given in different scales (Mulliken, Pauling, etc.) [4, 82, 86, 98, 101, 114]
ESG_i	Sanderson group electronegativity calculated as the geometric mean for atoms comprising considered group [4]
χ_μ	Orbital electronegativity of the μ-th atomic orbital [4]
μ	Electronic chemical potential for a molecule of N_{el} electrons [4, 104]
Local quantum-chemical properties	
$\varrho(\mathbf{r})$	Electron density [4]
$\bar{I}(\mathbf{r})$	Average local ionization energy [4]
P, $P_{\mu\nu}$, $P_{\mu\mu}$	Charge density matrix and its elements [4]
B_{ij}	Bond index, a measure of the multiplicity of bonds between two atoms [4]
V_i	Valency index, the valency of the i-th atom as the sum of the valencies of its atomic orbitals [4]
F_i	Free valence index, a measure of the residual valency of the i-th atom in π-electron molecular orbitals [4]
F_i', F_i' %	General free valence index, a representation of the residual covalent binding capacity of the i-th atom [4]
$\nu(\mathbf{r})$	Composite nuclear potential for a given configuration of the nuclei of a molecule [4]
$S(\mathbf{r},s)$	Somoyai function, representing the difference between the electronic density $\varrho(\mathbf{r})$ and the composite nuclear potential $\nu(\mathbf{r})$ at a point **r** and providing the information about chemical bonding [4]
MEP	Molecular electrostatic potential – defines the interaction energy of a molecule with a unit positive charge at position **r** [4, 74]
EP_i (P_i)	Electrostatic potential on the specified i-th atom [82, 101, 116]
MNEP, LNEP (P_{min}, P_{max})	The most negative and the least negative electrostatic potentials [82, 101]

(*Continued*)

Table 25.1 (*Continued*)

Symbol[a)]	Name/definition
F_k^{α} (FF)	Fukui function – defines electrophilicity associated with a site k in a molecule, where α represents local electrophilic quantities describing nucleophilic (+), electrophilic (–) and radical (0) attacks [4, 48, 66]
ω_g (ω)	Group or generalized electrophilicity [4, 48, 82, 98, 101, 104, 114]
Molecular orbital characteristics	
E_{HOMO}, E_{LUMO} (HOMO, LUMO)	Energies of the highest occupied and the lowest unoccupied molecular orbitals [4, 5, 15, 21, 55, 60, 75, 82, 84, 86, 98, 108, 109, 113–115, 124, 125, 128, 136, 137, 142, 146, 147]
$E_{LUMO} - E_{HOMO}$ (GAP, Δ_{LH})	Difference between the HOMO and LUMO energies [4, 5, 55, 84, 98, 114, 124, 128, 147]
$E_{LUMO} + E_{HOMO}$	Sum of HOMO and LUMO energies [124, 128]
E_{SOMO} (SOMO)	Energy of the singly occupied molecular orbital [110]
$E_{LUMO} - E_{SOMO}$	Difference between the SOMO and LUMO energies [110]
$f_{H/L}$ (r_{LH}))	HOMO/LUMO energy fraction, a stability index given by the ratio between HOMO and LUMO energies [4, 55]
HOP, LUP (HOPO, LUPO)	Energies of the highest occupied and the lowest unoccupied π orbitals [60, 61]
Φ_H, Φ_L	Angle between node in highest occupied π orbital and specific functional group [59]
$S2\Theta H, C2\Theta H, S4\Theta H, C4\Theta H$ $S2\Theta L, C2\Theta L, S4\Theta L, C4\Theta L$	Variables that account for orientation of nodes in π-like orbitals defined by sin and cos of the nodal angle in the highest occupied and lowest unoccupied molecular orbitals, respectively [60, 61]
η	Absolute hardness index [66, 82, 86, 98, 101, 104, 110, 114]
Δη	Activation hardness index represents the difference between absolute hardness of reactant and transition states [4]
S	Total softness index [66, 82, 86, 98, 101, 104]
Orbital electron densities	
f_i^E, f_i^N (f_i^-, f_i^+)	Electrophilic and nucleophilic frontier electron density of atom i at HOMO and LUMO, respectively [4, 5, 115]
F_i^E, F_i^N (F_i^-, F_i^+)	Indices of electrophilic and nucleophilic frontier electron density [4, 5, 113]
Superdelocalizability measures	
ES_i, NS_i (S_i^-, S_i^+)	Electrophilic and nucleophilic superdelocalizabilities measure, respectively, the availability of electrons in the i-th atom and the availability of space on the i-th atom for additional electron density [4, 5, 80, 142]
ES_T, NS_T (S_T^-, S_T^+)	Total electrophilic and nucleophilic superdelocalizability indices [4, 5, 80, 142]
Polarity measures	
π_{ii}, π_{ij}	Self-atom and atom–atom polarizabilities [5]

(*Continued*)

Table 25.1 (*Continued*)

Symbol[a)]	Name/definition
$\alpha, \bar{\alpha}, \alpha$	Molecular polarizability [5, 55, 74, 105, 110, 114, 116, 125, 128, 136], average polarizability [5, 20–22, 108, 113, 139, 147], polarizability tensor [5]
β	First-order hyperpolarizability [122, 128]
γ	Second-order hyperpolarizability [68, 128]
μ (DM_T)	Total molecular dipole moment [5, 20–22, 55, 74, 75, 78, 82, 84, 86, 101, 108, 110, 114, 124, 125, 128, 136, 139, 146, 147]
D_x, D_y, D_z (DM_x, DM_y, DM_z)	Components of dipole moment along inertia axes [5, 82]
Θ (Q_{ii}, MQM)	Molecular quadrupole moment [20, 22, 98, 106, 107]
Φ	Average hexadecapole moment of a molecule [106]
Δ (SPP)	Submolecular polarity parameter, defined as the largest difference in electron charges between two atoms [5, 78]
APT	Atomic polar tensor [147]
Charges	
Q_A	Net atomic Mulliken charge at specific atom [5, 59, 86, 98, 105, 113, 115, 128, 136, 137, 140, 146, 147]
Q_A^+, Q_B^- (Q_{min}, Q_{max} MPC, MNC)	The most positive and the most negative Mulliken atomic charges [5, 20–22, 55, 82, 84, 86, 101, 108, 109, 116, 124, 125, 128, 147]
Q_T, Q_A ($\sum Q_A$, SAC)	Sum of absolute values of charges of all atoms in a molecule or functional group [5, 82, 86, 101]
Q_T^2, Q_A^2 ($\sum Q_A^2$, SSC)	Sum of squares of charges of all atoms in a molecule or functional group [5, 82, 86, 101]
Q_m	Average of the absolute values of the charges on all atoms [5, 86, 101, 114]
Q_i	Electronic charge on the *i*-th atom [80, 142]
QTMS indices	
ϱ	Electron density [46, 47, 49, 51, 52]
$\nabla^2\varrho$	Laplacian of electron density [46, 47, 49, 51, 52]
$\lambda_1, \lambda_2, \lambda_3$	Three Hessian eigenvalues [46, 49, 51, 52]
ε	Ellipticity of a bond at the bond critical point (BCP) – provides a measure of an extent to which charge is accumulated in a given plane [46, 47, 49, 51, 52]
$G(r)$	Lagrangian kinetic energy density [46, 49, 51, 52]
$K(r)$	Hamiltonian kinetic energy density [46, 47, 49, 51, 52]
Thermal properties[b)]	
$E_{thermal}$	Thermal energy [20–22, 105, 107, 109]
E_{int}	Internal energy [20]
C_v	Heat capacity at constant volume [20–22]
S	Entropy [20–22]

a) Possible variations of symbols adopted in the contemporary literature are given in parenthesis.
b) Thermal properties calculated at $T = 298.15$ K, $P = 1.00$ atm.

redundant; this was in agreement with other studies [64]. Carbó-Dorca and co-authors have also proposed several new molecular descriptors based on quantum similarity measures. These are (i) molecular quantum self-similarity measures (MQS-SMs) that were used to generate statistically significant QSAR models for steroids binding to corticosteroid-binding human globulin [29] as well as QSPR models for series of organic compounds [31] and (ii) an electron–electron repulsion energy descriptor, V_{ee}, tested on a widespread set of molecules as a complement to steric and electronic parameters in the description of molecular properties and bioresponses [65].

The QTMS approach takes advantage of the BCP space concept, providing topological similarity descriptors that are discrete distance-like measures defined in three or higher dimensional BCP space [7, 43, 44]. Some of the QTMS descriptors listed in Table 25.1 represent components of a so-called chemical vector that describes a bond in 3D BCP space (e.g., the electron density, ϱ_b, the Laplacian of the electron density, $\nabla^2\varrho$, and the ellipticity, ε); others characterize a bond by evaluating Lagrangian and Hamiltonian kinetic energy densities, $G(\mathbf{r})$ and $K(\mathbf{r})$, respectively [46–50]. In terms of chemical interpretation, it was shown that sometimes QTMS descriptors provide measures of σ (ϱ_b and λ_3) and π ($\lambda_1 + \lambda_2$, ε) character of a bond or a simple measure of a covalent character of a bond ($\nabla^2\varrho$). The ellipticity and Laplacian of electron density can also provide information about structural stability and local concentration of electronic charge, respectively [46, 51, 52]. QTMS descriptors have been used to build a wide variety of QSAR/QSPR models in medicinal and ecological chemistry [46], details of which are given in the publications of Popelier and co-authors (see also Chapter 24).

It is necessary to mention that QSAR studies in which quantum-chemical descriptors are obtained from calculations that account for the solvent effect are fairly rare [61, 66], due to the obvious reason of reducing computational time and costs, especially if it is assumed that the presence of solvent does not change significantly the geometrical and electronic characteristics of the molecule [67]. It has been noticed, however, that in certain cases this assumption is not valid [5]. In computational modeling of biologic macromolecules associated with rational drug-design, accounting for solvent can be crucial. Khandogin and York [66] have presented a set of descriptors for the characterization of macromolecules in solution that were obtained with modest computational cost using linear-scaling semiempirical methods combined with a conductor-like screening model (COSMO). The authors demonstrated the stability and convergence of derived descriptors and their applications to study several nucleocapsid proteins [66].

From Table 25.1 one can see that the total energy of the molecule, HOMO and LUMO energies, the HOMO–LUMO energy gap, the total molecular dipole moment, the molecular polarizability and Mulliken atomic charges can be ranked as the most frequently used electronic properties in the life science-related QSAR. The references cited in Table 25.1 refer readers to publications in which certain quantum-chemical descriptors were considered and then their "predictive power" evaluated by statistical methods. It is not always obvious at the beginning of a particular QSAR study what descriptors (quantum or classical) and in what amount

or combination must be selected. Two major approaches to the problem are worth mentioning. Some researchers choose the statistical or chemometric approach: they prefer starting from the entire pool of available descriptors and then perform computer-aided selection of significant descriptors before including them as variables into QSAR models [68–79]. Interestingly, when classical and quantum-chemical descriptors are combined in such an automated procedure, the chances of classical descriptors being selected for the final model are much higher than those for quantum-chemical descriptors. Every so often this fact makes it difficult or almost impossible to achieve a meaningful physicochemical interpretation of a predictive model. An alternative approach includes knowledge or experience-based initial selection, preferably of quantum-chemical (i.e., more interpretable) descriptors and then the possible addition of certain classical parameters to increase the accuracy and predictive ability of the final model, if needed [15, 55, 80–86]. However, the latter approach, which is better known from the historical perspective of QSAR, is sometimes considered as biased and it also may lead to an excessive amount of trial and error in the process of building predictive models when complex biologic phenomena are involved. The next section introduces the most commonly used and promising algorithms for the selection of descriptors, together with contemporary statistical methods.

25.3
Computational Approaches for Establishing Quantitative Structure–Activity Relationships

25.3.1
Selection of Descriptors

Contemporary QSAR has thousands of parameters available from experiment and *in silico* calculations that could potentially serve as independent variables (descriptors) in statistical analysis. It is already known, however, that utilization of an excessive number of descriptors leads to over-fitting of QSAR models and/or increases the risk of chance correlations. Despite the existence of "five golden rules" for building successful and meaningful QSAR models, formulated in 1973 by Hansch and Unger [87], the increasing complexity of biologic mechanisms on the one hand creates the need for considering a large variety of variables and, on the other, makes a knowledge-based approach to the identification of the most significant descriptors for a particular case extremely difficult. In his more recent review Kubinyi mentions the disadvantages of elimination procedures using forward, backward and stepwise regressions [1]. The conceptual description of mathematical approaches to the selection of descriptors that have shown to be efficient in generating optimal or near-optimal predictive models is presented below. A more general discussion of this topic, as well as of the process known as "feature selection" that allows identification and elimination of redundant or ineffective descriptors, can be found in a review of Nikolova and Jaworska [8] and the references therein.

Principal component analysis (PCA) performs reduction of data by generating linear combinations of original variables, that is, descriptors [57, 88–91]. The PCA method identifies correlated variables, groups them into linear combinations and generates an entire new set of uncorrelated (orthogonal) variables called principal components (PC). The direction of the first principal component (PC_1) is chosen to maximize the variance in the data; the next, PC_2, is orthogonal to PC_1 and directed to maximize as much variance left as possible and so on. The process of data transformation is given by:

$$\mathbf{X} = \mathbf{TP}^{\mathrm{T}} \tag{25.3}$$

where **X** represents the initial data matrix, **T** is a score matrix that defines the position of data points in a new coordinate system and **P** is a loadings matrix. The loadings indicate how much each original descriptor contributes to the corresponding PC. Scores and loadings allow the data points to be mapped into the new vector space defined by PCs [89, 91].

The decision tree method was employed by Smith *et al.* [92, 93] and later by Gubskaya *et al.* [79] to select the most significant descriptors for predicting the adsorption of fibrinogen onto polymer surfaces. Decision trees are usually used for description, classification and generalization of data [94]. The decision trees involved in the selection and analysis of descriptors classify data points by starting at the top of the tree (root node) and moving down the tree by creating a hierarchy of descriptor values on an "if-then-else" basis at each branch point until the terminal leaf (node) is reached. In these top-down constructions the data are recursively divided into subsets based upon the best classifying descriptors at each level (Figure 25.1). The C5 decision tree algorithm [95] employed by the authors evaluates the significance of each descriptor with respect to the set of experimental fibrinogen adsorption data using the concept of information gain introduced by Shannon [96, 97]. The conventional C5 algorithm was augmented by a Monte Carlo procedure to account for the experimental uncertainty [79]. All descriptors with the highest information gain were

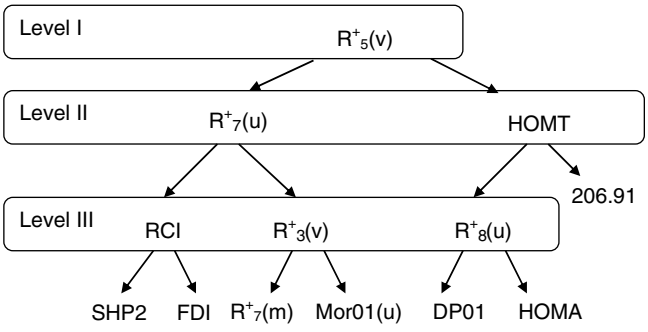

Figure 25.1 Schematic representation of C5 decision tree. The three most significant descriptors (level I and level II) were used to build an ANN model for predicting fibrinogen adsorption to polymeric surfaces [79].

extracted from thousands of Monte Carlo pseudo-experiments and summarized into a histogram. The three descriptors with the highest counts in this histogram were used as input variables to build QSAR models. It was shown that decision tree algorithms appear to be valuable tools for identifying the most relevant descriptors; however, so far they have only been used in the cases when classical descriptors are concerned. It would be beneficial from a methodological viewpoint to test the application of these algorithms on the QSAR models based exclusively on quantum-chemical properties whose biologic relevance to the particular bio-response was defined in advance by the knowledge-based approach.

A genetic algorithm (GA), a meta-heuristic method for the optimization of a function, was recommended by Kubinyi [1] as a method of choice for variable selection in QSAR [85, 98, 99]. The concept of the genetic algorithm is similar to that proposed in Darwin's theory of biologic evolution due to natural selection. In genetic algorithm terminology, an initial group ("population") of random organisms (sometimes called "chromosomes") evolves according to a fitness function that determines their survival. The algorithm searches for the "fittest" organisms through a selection, mutation and crossover genetic operation. "Genes" represent the properties of organisms; in the case of a feature selection these are descriptors. In other words, the method generates a set of potential solutions to a problem and then this set is iteratively modified and tested until an optimal solution is found. Genetic algorithms were successfully employed to select the most relevant descriptors from the large pools of variables containing various classes of classical and quantum-chemical descriptors [78, 100] as well as for the identification of the most significant characteristics from the relatively small sets obtained mainly by quantum simulations [76, 82, 98, 99, 101].

25.3.2
Linear Regression Techniques

Since 1964, the year the contemporary QSAR approach was born due to the contributions of C. Hansch, T. Fujita, S.M. Free and J.W. Wilson [102, 103], most QSAR models have been built using a multivariate regression technique. Regression analysis establishes a correlation between a dependent variable representing the biologic activity and multiple independent variables, that is, the descriptors (predictors). This correlation is most often expressed in the form of a multiple linear regression (MLR) equation as:

$$\text{activity} = \sum_i x_i \alpha_i + b \tag{25.4}$$

or, as in the case of a forward stepwise multiple regression (SMLR) technique [81, 98], as:

$$\text{activity} = \sum_i x_i (\alpha_i + \Delta\alpha_i) \tag{25.5}$$

where x_i denotes molecular descriptors, α_i and b are coefficients to be optimized, and $\Delta\alpha_i$ are coefficient errors. The best models are selected based on the correlation

coefficient (R or/and R^2), standard deviation (S) and the value F that represents the level of statistical significance of the model.

Clare and co-authors [59–61] have proposed a procedure called flip regression whereby MLR was applied to build QSAR models for benzene derivatives with each molecule independently in both possible orientations. For phenethylamine, for example, flipping includes rotating or reflecting the benzene ring in such a way that the charge and substituent values for two *o*- and two *m*-positions are interchanged. Then for N molecules 2^N regressions are generated and the one giving the best fit is selected [58].

The statement that "multiple regression is the most widely used mathematical technique in QSAR analysis" [2] is valid not only for all varieties of QSAR studies but also for those focused on the utilization of quantum-chemical descriptors [14, 20–22, 82, 104–115] and their combinations with other types of molecular characteristics [68, 70, 71, 76–78, 98–100, 116–123].

The partial least squares (PLS) regression method is the next multivariate linear regression technique commonly applied in QSAR modeling [51, 52, 67, 88, 124, 125]. It is often combined with principal component analysis or genetic algorithm to select the most appropriate input variables [52, 67, 85]. PLS establishes relationships between highly correlated input and output variables represented by arrays of data. It performs a reduction of the dimensionality of the raw data using both input and output data (i.e., **X** and **Y** matrices, respectively). Decomposition of **X** and **Y** is carried out simultaneously according to:

$$\mathbf{X} = \mathbf{TP} + \mathbf{D} \tag{25.6}$$

$$\mathbf{Y} = \mathbf{UQ} + \mathbf{F} \tag{25.7}$$

where **T** and **U** are the **X**- and **Y**-block score matrices, **P** and **Q** are the **X** and **Y** loadings, and **D** and **F** are residuals. A distinctive feature of this method is that it builds a regression model according to equation:

$$\mathbf{U} = \mathbf{BT} \tag{25.8}$$

with **U** and **T** representing scores (or projections) of dependent and independent variables, respectively. The *N*-way or multilinear partial least-squares method (*N*-PLS) is an extension of bilinear PLS designed for use in 3D-QSAR. *N*-PLS simultaneously decomposes and processes three-dimensional arrays of data such as GRID descriptors calculated by the comparative molecular field analysis (CoMFA) method or data generated by the quantum topological molecular similarity approach. Esteki *et al.* [51] have used both bilinear and *N*-PLS for QTMS indices-based QSAR modeling and prediction of acidity constant (pK_a) for some phenol derivatives.

25.3.3
Machine-Learning Algorithms

Since the late 1980s artificial intelligence methods have become an essential tool in QSAR analysis due to an increasing demand for accuracy and to the rapidly growing

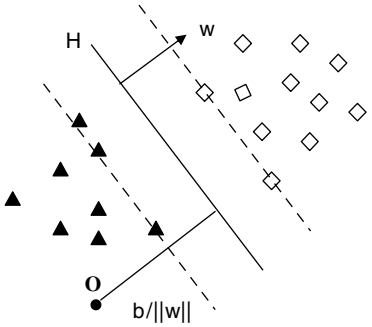

Figure 25.2 Schematic of SVM. The hyperplane *H* is shown by solid line and situated to maximize the margin, $d = 2/||\mathbf{w}||$ depicted by two dashed lines. Support vectors are located on the margin. Here **w** is the normal vector of hyperplane and $b/||\mathbf{w}||$ is its perpendicular distance from the origin.

number of SAR cases that exhibit highly nonlinear relationships. This section introduces the most promising SAR analysis nonlinear predictive methodologies and, when possible, their performance will be compared. In all cases mentioned here quantum-chemical descriptors were employed throughout or/and were found among the most significant variables responsible for accurate and meaningful correlations.

Not long ago a machine-learning technique called support vector machines (SVMs) became a part of the data analysis toolbox used for solving classification and regression problems in computational chemistry [126], specifically in drug-design [74, 83, 127], QSAR [73, 128], chemometrics and chemical engineering. The concept of SVM, originally introduced by Vapnik [129] and Lerner [130], is one of the major developments in statistical learning theory. Its principles can be briefly summarized as follows. The algorithm is designed to find an optimal hyperplane, *H*, between data points (i.e., all descriptors), **x**, separating two distinct classes labeled as $y = 1$ and $y = -1$. The hyperplane, $\mathbf{w}\cdot\mathbf{x} + b$, is defined by its normal vector, **w**, and the perpendicular distance from the origin $b/||\mathbf{w}||$ (Figure 25.2). The classification problem involves the optimization of Lagrangian multipliers α_i to generate a decision function $f(\mathbf{x})$ given by:

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{l} y_i \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b\right) \tag{25.9}$$

where $0 \leq \alpha_i \leq C$ and $\sum \alpha_i y_i = 0$ are constraints to be satisfied, C is a regularity parameter, $\boldsymbol{x_i}$ are support vectors (i.e., subset l of descriptors) and $K(\boldsymbol{x}, \boldsymbol{x}_i)$ is a Kernel function. A kernel function is defined in descriptor spaces of high dimensionality $\Phi(\boldsymbol{x})$ and $\Phi(\boldsymbol{x}_i)$ as $K(\boldsymbol{x}, \boldsymbol{x}_i) = \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x}_i)$ and it can be computed without explicit use of $\Phi(\boldsymbol{x})$. To solve the regression problem, Equation 25.9 has to be rewritten as:

$$f(\boldsymbol{x}) = \sum_{i=1}^{l} y_i \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b, \tag{25.10}$$

where constrains are similar to those given for Equation (25.9). Several types of kernel functions are known [126]: linear, polynomial, Gaussian, exponential radial basis function (RBF) kernel and so on. The kernel function and the regularity parameter C are the only features that specify the SVM algorithm for a given data set. The Gaussian kernel is the most commonly used for support vector classification and regression. In addition to the conventional SVM algorithm, a few of its modifications (such as least squares (LS) SVM [131] and v-SVM [132]) were recently employed in medicinal chemistry to build successful QSAR models [83, 127].

Zheng *et al.* [128] reported predictions of receptor relative binding affinities for polybrominated diphenyl ethers using SVM and radial basis function neural network (RBFNN). After correlating relative binding affinities with seven of the most significant quantum-chemical descriptors, the authors concluded that SVM models generalize better than RBFNN models [128]. A similar conclusion was drawn by Liu *at al.* [73]; their analysis of classification models built with SVM and the k-nearest neighbor (k_{NN}) method for a novel series of cyclooxygenase selective inhibitors revealed that the performance of the k_{NN} algorithm is much less satisfactory than that of SVM. Comparison of nonlinear and linear QSAR models for a set of pyrazine-pyridine biheteroaryls, inhibitors of vascular endothelial growth factor receptor-2, showed that nonlinear models obtained using LS-SVMs perform better in terms of generalization and predictive ability than the multiple linear regression model [127]. Interestingly, the SVM and cluster-genetic algorithm-partial least squares discriminant analysis [133] classification models generated to predict chemical metabolism by human UDP-glucuronosyltransferase isoforms [83] demonstrated very similar and somewhat unsatisfactory performance with two-dimensional (2D) chemical descriptors employed. The authors applied an electronegativity equalization method (EEM) [134, 135] developed for the fast calculations of DFT-based molecular and atomic properties to compute quantum-chemical descriptors. Combination of 2D and EEM-derived descriptors using the so-called consensus approach made it possible to achieve significant improvement (up to 84%) of overall substrate and non-substrate predictability [83].

Artificial neural network (ANNs) are machine-learning algorithms that like SVM can be used to handle classification and regression problems. Classifying ANN estimates the so-called membership functions and doing so translates continuously changing output value into discrete nominal categories [84]. Such a procedure increases the robustness of the algorithm, simplifies and speeds up the training process and provides some measures of similarity between investigated objects (i.e., molecules). Regression ANN handles a prediction problem by estimating the value of a continuous output variable that is calculated for the known values of input parameters. It requires higher accuracy and therefore increases the time of the training process and the number of learning examples [84]. Additionally, regression ANN models must demonstrate extrapolation capabilities that, in contrast to radial basis function networks, are associated with the multilayer perceptron-type of NN architectures. Multilayer perceptrons are feed-forward, multilayer networks designed to provide adaptable frameworks for nonlinear function estimation. They consist of neurons or nodes arranged in layers: an input layer, one or more hidden

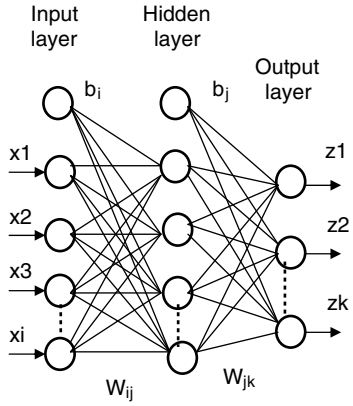

Figure 25.3 Structure of ANN represented by a two-layer perceptron with multi-output node, where the input variables and output values are denoted as x and z, respectively; b_i and b_j stay for biases; w_{ij} and w_{jk} are the connection weights between input/output, respectively, and a hidden nodes.

layers and an output layer. Multiple connections between neurons of adjacent layers are unidirectional (i.e., from input to output) and reflected in the connection weights (Figure 25.3). First, input neurons distribute the initial input variables (x_i) among hidden layer neurons without additional computations. Then the hidden layer variables (y_j) are calculated as:

$$y_j = f\left(\sum_i x_i w_{ij} + b_i\right) \tag{25.11}$$

where w_{ij} are the connection weights between input and hidden nodes, b_i is the bias of input layer, and $f(\varsigma) = (1 + e^{-k\varsigma})^{-1}$ is a sigmoid transfer function that varies between 0 and 1 with coefficient k to be user-specified. Finally, the output values (z_k) are computed as:

$$z_k = \sum_j y_j w_{jk} + b_j \tag{25.12}$$

where w_{jk} and b_j are the weights and the bias for the output layer, respectively. The value E represents the target error defined as:

$$E = \frac{1}{2}\sum_{k=0}^{s-1}(z_k - \hat{z}_k)^2 \tag{25.13}$$

where z_k and $\hat{z}_k$ are the predicted and experimentally determined output values, respectively.

Subsequent minimization of E modifies the connection weights to achieve the best fit.

In addition to conventional feed-forward (or as they are sometimes called, back-propagation) ANN algorithms [69, 82, 84, 91, 105, 136–139] and their variations, such as projection pursuit regression models [140], several alternative designs, namely, counter-propagation ANN architecture [141], Bayesian-regularized [101] and fuzzy ARTMAP [139] neural networks, were explored to handle complex quantitative relationships in biology-related systems. Counter-propagation ANNs are based on an unsupervised learning strategy that does not require a priori knowledge about the process to be modeled. Although this type of algorithm was not extensively used in QSAR, Novic *et al.* [141] used it to study the correlation of inhibitory activity of 105 flavonoid inhibitors of the enzyme $p56^{lck}$ protein tyrosine kinase and proved that its results compare favorably with those obtained by classical back-propagation methods. The Bayesian regularization (BR) approach, instead of looking for global minima, finds locally the most probable network parameters, estimates their effectiveness and produces predictors that are reliable and well-matched to the data. Multiple linear regression model and BRANN combined with feature selection by genetic algorithm were recently utilized to model the biologic activity of luteinizing hormone-releasing hormone antagonists [101]. The better results achieved by the GA-BRANN suggest that nonlinear analysis is more suitable to model complex SAR of these non-peptide antagonists [101]. The fuzzy ARTMAP-based model was developed to construct log K_{ow} QSPR for 442 compounds and it demonstrated clear superiority with respect to back propagation ANN and multiple linear regression QSPR models, due to its ability to handle noisy data with fuzzy logic and to avoid the plasticity–stability dilemma of standard back propagation architectures [139].

There is a common belief that because SVM is based on structural risk minimization its predictions are better than those of other algorithms [126]. It was also suggested that SVMs perform better than ANNs because they provide a unique solution, whereas ANNs can become trapped in local minima and the optimum number of hidden neurons of ANN sometimes requires time-consuming computations. It is known, however, that ANNs can provide constantly good suboptimal solutions and that this is not always the case for SVMs. Thus, only an empirical comparison of results obtained by various machine-learning algorithms can demonstrate, for a particular SAR problem, the superiority of one method over the other [126].

25.4
Quantum-Chemical Descriptors in QSAR/QSPR Models

25.4.1
Biochemistry and Molecular Biology

Just over a decade ago, Karelson *et al.* [5] emphasized that quantum-chemical descriptors have a long history of use in QSAR studies in biochemistry and also indicated that a trend of such descriptors as HOMO–LUMO energies, frontier orbital

electron densities and superdelocalizabilities correlates well with various biologic activities. In the present section several examples of the most successful QSAR models will be shown and discussed without providing specific details of correlation equations, for which readers are referred to the original publications.

The way in which the contemporary QSAR approach pursues its twofold goal of understanding the chemical mechanisms associated with biochemical effects and providing practical recommendations for optimal molecular design can be demonstrated in investigations of inhibitory activity. The potency of the intermediate conductance Ca^{2+}-activated K^{+} channel (IKCa1) blockade by triarylmethane (TRAM) derivatives has been studied by Fernández and coworker [86]. The results showed that *ab initio* derived electronic properties in combination with topological (WHIM) descriptors are important parameters influencing the binding of TRAMs with IKCa1: global quantum-chemical properties (i.e., molecular dipole moments, HOMO and LUMO energies) describe the electronic environment of the molecules and local parameters (i.e., local Mulliken charges) help to locate molecular regions responsible for given bioactivity [86]. Analogous behavior in terms of selected descriptors tendency was observed by Safarpour *et al.* [82], who conducted QSAR analysis on the Ca^{2+} channel antagonist activity of some newly synthesized 1,4-dihydropyridine derivatives. From the six descriptors selected for the multiple linear regression model, five were related to the electronic (i.e., electrophilicity, electronegativity and dipole moment) or physicochemical (surface area and molar refractivity) properties of the whole molecules and one (electrostatic potential) described electronic properties of individual atoms [82]. Inhibition activity of flavonoids was investigated based on their ability to inhibit replication of the human immunodeficiency virus (HIV) [81] and to maintain the balance between neuronal excitation and inhibition in the central nervous system (CNS) by binding to the γ-aminobutyric acid type A [GABA(A)] receptor [77]. In the former case it was concluded that HIV-inhibitory activity is mainly the outcome of electronic interactions between atomic charges within flavonoids and possible receptor-like structures in the HIV or the lymphocyte itself. In the latter case it was shown that the binding affinities of selected flavonoids to GABA(A) receptor are highly dependent on conformational changes involved in the interactions [77]; in other words, no electronic properties were found to be significant after exploring the entire pool of 1176 classical and electronic variables. This example, however, appears to be an exception: the results of many other researchers clearly indicate a trend of preferable utilization of electronic properties in QSAR studies of inhibition activity [59–61, 85, 112–114].

Special types of inhibitory activity, such as the inhibition of parasite or bacterial growth, have been studied by several research groups [28, 71, 110, 142]. Katritzky and co-authors [71] reported QSAR models obtained using the B(est)MLR method and descriptors calculated with CODESSA PRO software [143] for two diverse sets of potentially active compounds against the D6 and NF54 strains of malaria. The mechanism of antimalarial activity showed the significance of charge-related interactions as well as the shape and branching of the molecules [71]. Utilization of kinetic energy DF similarity measures as descriptors by Gironés *et al.* [28] also led

to satisfactory correlations for all antimalarial activities in all studied molecular sets. Molina and co-authors [142] have developed a new QSAR strategy that includes a linear piecewise regression and discriminant analysis and examines the possibility of combining different types of molecular descriptors. Their best models for inhibition of respiration in *Escherichia coli* by 2-furylethylenes included only quantum-chemical descriptors such as local charges and electrophilic and nucleophilic superdelocalizabilities [142].

QSAR/QSPR approaches have been used over the years to develop models capable of estimating or predicting the toxic potency of various compounds ranging from small molecules to biosystems. The most recent achievements in this area include QSAR modeling of toxicity of various aromatic compounds [48, 49, 68, 100, 121, 124, 125] and halocarbons [15] to develop efficient and inexpensive methods for the estimation of their effects on human and environmental health. The quantum-chemical descriptors that quantify the electrophilic nature of the molecules were identified as premier quantities in the prediction of toxicity [15, 48, 100, 121, 124]. Again, in some cases quantum-chemical and classical descriptors were combined to fortify models obtained solely with classical descriptors [15, 68, 121].

The chemicals, which are known as potential carcinogens, have been in the focus of the scientific community for at least 70 years. Among these, polycyclic aromatic hydrocarbons take a special place: some of them are ranked as the strongest carcinogens, but some are identified as inactive. Several research groups have investigated the relationship between carcinogenic activity and the chemical structure of PAHs using quantum-mechanical calculations [56, 57, 124, 144, 145]. The latest studies include the work of Barone *et al.* [56], who proposed a new methodology, EIM, that utilizes electronic descriptors (Section 25.2.2), as well as the work of Vendrame *et al.* [57], who performed a comparative study of EIM and PCA-ANN approaches and confirmed that the key descriptors in EIM are indeed the relevant descriptors for classifying the carcinogenic activity of PAHs. In addition, Lu *et al.* [124] have reported a QSPR model for water solubility of PAHs obtained using electronic descriptors calculated at the B3LYP/6-31G(d) level and PLS statistics. Their results demonstrate the superiority of electronic parameters in comparison with known physicochemical properties and/or topological indices.

25.4.2
Medicinal Chemistry and Drug Design

In modern QSAR the distinction between biochemistry and molecular biology at one end and medicinal chemistry and drug-design at the other is somewhat artificial: in all these fields, biologic processes and phenomena are the matters of primary concern. Thus the present section will focus on examples of practical interest to medicine and pharmacology.

The blood–brain barrier (BBB) is represented by a complex cellular system that maintains the homeostasis of the CNS by separating the brain from systemic blood circulation. The ability of a drug to penetrate the BBB is of utmost importance

in the design of neurological drugs: CNS-active drugs require high penetration, while it is more desirable for drugs intended for peripheral activity to exhibit minimal penetration to prevent CNS-related side effects [118]. To build a QSAR model for the brain/plasma partition coefficient, log (C_{brain}/C_{blood}) or log BB, Hutter employed variables from the AM1 optimized geometry of 90 compounds [119]. The electrostatic potential and, related to it, the set of variables that accounts for the polarity of molecular surfaces were identified as the most significant descriptors of his 12-term MLR-based model [119]. Van Damme *et al*. [118] have developed and presented a new *in silico* model to predict log BB, for a set of 82 structurally diverse compounds using a combination of classical and quantum-chemical descriptors. The final eight-parameter model among the others included several Mulliken charge-related descriptors and the dipole moment, confirming the known fact that non-polar molecules cross the barrier more easily than polar molecules [118]. Among a multitude of CNS-related drugs, valproic acid (VPA) is an established antiepileptic drug, which is generally well tolerated but has two serious potential side effects: hepatotoxicity and teratogenicity [98]. In an attempt to find a superior compound that would retain the anticonvulsant activity of the basic structure of VPA but would not cause the adverse side effects, Hashemianzadeh *et al*. [98] performed a QSAR study of 25 potent VPA derivatives utilizing DFT calculations and QTAIM to obtain quantum-chemical descriptors. Their statistically significant MLR model (with a correlation coefficient of 0.937) suggested that polarizability and electrostatic potential at certain carbon atoms of the drugs are strongly correlated with antiepileptic activity of these types of VPA derivatives [98].

Electronic descriptors were the major descriptors employed in QSAR studies of the anesthetic action of some polyhalogenated ethers [116] and the antioxidant properties of phenolic compounds [104]. Mehdipour *et al*. [116] reported an MLR equation (with correlation coefficient of 0.97) that clearly indicated the significant effects of coulombic (i.e., electrostatic potential and most positive charge descriptors), steric and polar interactions (molecular polarizability) as well as lipophilicity (log *P*) on the anesthetic activity of the polyhalogenated ethers. Reis *et al*. [104] employed and compared the predictive quality of quantum-chemical descriptors potentially relevant to antioxidant activity that were calculated at AM1, PM3, HF and B3LYP levels of theory for 41 phenolic compounds. The best regression equations included E_{HOMO}, vertical ionization potential, IPv, and charge on oxygen atom, Q_O. These descriptors, obtained at both HF and DFT/B3LYP levels, revealed that low values of $IPv_{HOMO(DFT)}$ (from Koopman's theorem) combined with negative charges on O7 lead to an increase in the antioxidant activity [104]. The antioxidative nature of hydroxyphenylureas has been investigated by Deeb *et al*. [120]; their eight parametric model consisted of five topological and three quantum-chemical descriptors: Q_m, Q_{max} and μ.

As mentioned in Section 25.2.2, QSAR studies that combine both classical and quantum-chemical descriptors are becoming more common. Alvarez-Ginatre and co-authors [76, 99] reported several predictive QSAR models of anabolic and androgenic activities for selected steroid analogs. In this study the combination of electronic and physicochemical descriptors helped to identify molecular shape,

hydrophobicity and electronic properties as three major factors responsible for these types of steroid activity. A similar approach to the initial selection of QSAR descriptors was employed for the discovery of less toxic, more selective and more effective agents to treat [55, 69, 123, 127] and prevent [73] cancer.

25.4.3 Material and Biomaterial Science

The use of QSAR/QSPR in the development of biomaterials and, in particular, materials for tissue-engineering applications is relatively new. The recent success of DFT in the accurate determination of electronic properties of biologically significant molecules has initiated QSPR studies in material science that focus specifically on the design of polymeric (bio)materials. DFT provides an invaluable tool for calculating quantum-chemical descriptors that demonstrate high potential in generating predictive QSPR models without the addition of classical descriptors for various classes of polymers represented by their repeat units. Yu, Gao, Liu and co-authors [20–22, 105–109, 136, 146, 147] have published a series of articles devoted to the prediction of the most important physicochemical properties of polymers, namely, refractive index [21], dielectric constant [109], glass transition temperature [105, 106, 109], melting point [136], cohesive energy [108], molar stiffness function (i.e., conformational property) [22], thermal decomposition [107] and reactivity parameters in free-radical polymerization [147]. It is somewhat surprising that no attempts to build QSAR models for prediction of biologic response (e.g., polymer adsorption [79, 92, 93] onto or cellular proliferation [89, 148] to the polymer surfaces) using quantum-chemical descriptors have been reported by the time of writing this chapter.

The conformational properties of polymer chains play a key role in synthesis (i.e., polymerization in solution), processing (i.e., solvent casting on thin films) and bioresponse onto the surfaces of biocompatible and biodegradable polymers. The molar stiffness function, K, is directly related to the intrinsic viscosity of polymer solutions and can be estimated using QSPR models. Yu *et al.* [22] used three quantum-chemical descriptors (quadrupole moment, thermal and total energies) to predict the molar stiffness function K for 47 vinyl polymers. A physically meaningful QSPR model with correlation coefficient 0.958 and mean error 7.1% was generated using stepwise MLR analysis. Yu and co-authors [147] also reported an accurate QSPR model (root mean square errors of 0.37 and 0.19, respectively) using back propagation ANN for predicting the reactivity parameters Q and e in radical copolymerization of vinyl monomers. The authors showed that the Mulliken charges and frontier molecular orbital energies are the descriptors most correlated with reactivity parameters [147].

An interesting comparison can be made between two independently generated QSAR models to predict the refractive index for the same representative set of polymers previously investigated by Bicerano [149]. Holder *et al.* [117] used CODESSA program [143] to calculate a total of 600 (classical and quantum-chemical) descriptors for each of the 60 polymers represented by dimer models. Quantum

Table 25.2 Comparison of QSAR models for T_g of polymeric materials.

Significant descriptors	Methods	Correlation coefficient	Class of polymers	References
PMA,[a] M,[a] U,	DFT		Polyamides	Gao *et al.* [146]
E_{LUMO}, Q_O	MLR, ANN	0.889, 0.898		
Θ, Φ	DFT		Polyvinyls, polyethylenes and polymethacrylates	Yu *et al.* [106]
	MLR	0.952		
L-1.356[a]	DFT		Polymethacrylates	Liu *et al.* [105]
$E_{thermal}$, α, Q_C	MLR, ANN	0.960, 0.991		
$E_{thermal}$, E_{HF}	DFT		Polyalkanes and polyacrylamides	Liu *et al.* [109]
	MLR	0.921		

a) Classical descriptors.

descriptors were obtained by the AM1 method. The final QSPR model with a correlation coefficient of 0.953 was obtained by MLR and featured two electronic descriptors: HOMO–LUMO gap and a polarizability index. Yu *et al.* [21] used refractive index data for 95 polymers from the same dataset and calculated only ten quantum-chemical descriptors for given monomers using DFT at the B3LYP/6-31G (d) level. In this case four descriptors (LUMO energy, molecular polarizability, heat capacity at constant volume and the most positive charge on hydrogen atom) were selected to build optimal QSPR models by means of stepwise MLR (correlation coefficient 0.926). These independently obtained and comparable results clearly identify the main electronic parameters that affect the values of refractive index for vinyl polymers.

The glass transition temperature, T_g, is one of the most important characteristics of amorphous polymers. Table 25.2 summarizes the results of several QSPR studies devoted to the prediction of T_g for various classes of polymers. There have been attempts in the past decade to evaluate the usefulness of different classes of descriptors in establishing correlations for certain types of activities/properties by comparing their statistical performance separately or in combination using the same or different data sets [150]. Clearly, from Table 25.2 (as well as from the previous example), it is impossible to provide any reliable generalizations in this regard and this fact can be used as the main argument against broad utilization of knowledge-based initial selection of descriptors.

25.5 Summary and Conclusions

Since QSAR/QSPR has been established as a methodology that allows one to estimate the properties of chemicals at a much lower cost than that of actual laboratory screening, it has been widely applied in all chemistry-related fields of the life

sciences. The necessity to produce robust and reliable QSAR models, the results of which would be of immense practical value, has led to the development of the "Guidance document on the validation of (Q)SAR models" that was recently introduced by a group of experts from the Organization for Economic Co-operation and Development [151]. The principles formulated in this document clearly reflect the original goal of QSAR to be a reliable predictive tool and to provide, when possible, mechanistic interpretation(s) of the model. Both the predictive ability of the model and the scientific insights into the mechanisms involved in a particular kind of biologic activity depend on the descriptors selected in the modeling process.

The use of quantum-chemical descriptors has an obvious advantage over other calculated or experimentally measured properties because they are reproducible (in the framework of the chosen approximation), they allow meaningful interpretation of QSAR models in terms of the mechanism of actions, metabolic or toxicological routes and, thus, can offer clear guidance for molecule optimization or design. In some cases it becomes necessary to consider both classical and quantum-chemical descriptors as potential candidates for successful QSAR modeling that sometimes dramatically increases the number of input variables. Among all available tools for automatic selection of descriptors, genetic algorithms have shown to be the most promising.

Rapid developments in combinatorial synthesis have facilitated the production of large amounts of chemical compounds whose activity cannot be easily estimated even by modern high-throughput screening techniques. The utilization of machine-learning methodologies such as artificial neural nets and support vector machines provides contemporary QSAR reliable ways of handling nonlinear statistics and noticeably increases accuracy and predictive power not only of large industrial-scale models but also of models generated for local scientific or testing purposes.

One of the major reasons for the acceleration of the use of quantum-chemical descriptors in QSAR modeling has been the development of DFT. Both relatively low costs and reasonable accuracy have led to the successful utilization of DFT for the calculation of a broad range of properties for larger molecules. Many researchers have confirmed the superior performance of DFT in comparison with semiempirical calculations. The concept of quantum similarity has stimulated the development of various theories and algorithms that have given rise to a new generation of descriptors known as quantum similarity measures. These descriptors are rapidly becoming competitive with conventional electronic parameters in solving QSAR problems in life science-related fields. Similar to the previous decade, the recent applications of quantum-chemical descriptors include prediction of inhibitory activity, chronic and acute toxicity, ligand–receptor binding affinity, antimicrobial activity, carcinogenesis and mutagenesis. Successful utilization of DFT-derived electronic parameters in the prediction of various properties of polymers has created the basis for *in silico* design of new biomaterials. Clearly, quantum-chemical descriptors have significant applicability potential in traditional areas of QSAR and their applications in novel and rapidly growing fields of biomedical science has yet to be explored.

Abbreviations

AIM	atoms in molecules (theory)
AM1	Austin model 1
ANN	artificial neural networks
BCP	bond critical point
BRANN	Bayesian regularization ANN
CNS	central nervous system
DF	density function
DFT	density functional theory
EIM	electronic indices methodology
GA	genetic algorithm
HF	Hartree–Fock
HOMO	highest occupied molecular orbital
LUMO	lowest unoccupied molecular orbital
MLR	multiple linear regression
MNDO	modified neglect of differential overlap
MP	Møller–Plesset
MQSM	molecular quantum similarity measure
PAH	polycyclic aromatic hydrocarbons
PCA	principal component analysis
PLS	partial least-squares
PM3	parametric model 3
QSAR/QSPR	quantitative structure activity/property relationship
QTMS	quantum topological molecular similarity
RBFNN	radial basis function neural network
SVM	support vector machines
WHIM	weighted holistic invariant molecular (descriptors)

Acknowledgments

The author is thankful to Professor C. F. Matta for the invitation to contribute a chapter to this book and his useful comments on its content. The constructive criticism and professional remarks of Drs Y. V. Lisnyak and V. Kholodovych are also deeply appreciated.

References

1 Kubinyi, H. (2002) *Quantum Struct.-Act. Relat.*, **21**, 348–356.
2 Selassie, C.D. (2003) *Burger's Medicinal Chemistry and Drug Discovery*, 6th edn, vol. 1 (ed. D.J. Abraham), John Wiley & Sons, Inc., New York.
3 Gramatica, P., Sumathy, K.V.C., and Suraj, S. (eds) (2008) A Strand Life Sciences Web Resource. http://www.qsarworld.com/qsar_archives.php, (September 2008).
4 Todeschini, R., Mannhold, R., Kubinyi, H., Consonni, V., and Timmerman, H.

(2000) *Handbook of Molecular Descriptors*, John Wiley & Sons, Inc., New York.
5 Karelson, M., Lobanov, V.S., and Katritzky, A.R. (1996) *Chem. Rev.*, **96**, 1027–1043.
6 Parr, R.G. and Yang, W. (1989) *Density Functional Theory of Atoms and Molecules*, Oxford University Press, Oxford.
7 Koch, W. and Holthausen, M.C. (2001) *A Chemist's Guide to Density Functional Theory*, 2nd edn, Wiley-VCH Verlag GmbH, Weinheim.
8 Nikolova, N. and Jaworska, J. (2003) *QSAR Comb. Sci.*, **22**, 1006–1026.
9 Dewar, M.J.S., Zoebisch, E.G., Healy, E.F., and Stewart, J.J.P. (1985) *J. Am. Chem. Soc.*, **107**, 3902–3909.
10 Dewar, M.J.S. and Thiel, W. (1977) *J. Am. Chem. Soc.*, **99**, 4899–4907.
11 Stewart, J.J.P. (1989) *J. Comput. Chem.*, **10**, 209–220.
12 Levine, I.N. (1991) *Quantum Chemistry*, Prentice Hall, Englewood Cliffs, New Jersey.
13 Braga, S.F. and Galvão, D.S. (2004) *J. Chem. Inf. Comput. Sci.*, **44**, 1987–1997.
14 Trohalaki, S., Gifford, E., and Pachter, R. (2000) *J. Comput. Chem.*, **24**, 421–427.
15 Basak, S.C., Balasubramanian, K., Gute, B.D., Mills, D., Gorczynska, A., and Roszak, S. (2003) *J. Chem. Inf. Comput. Sci.*, **43**, 1103–1109.
16 Lee, C., Yang, W., and Par, R.G. (1988) *Phys. Rev. B*, **37**, 785–789.
17 Becke, A.D. (1993) *J. Chem. Phys.*, **98**, 1372–1377.
18 Becke, A.D. (1993) *J. Chem. Phys.*, **98**, 5648–5652.
19 Wang, Z.-Y., Zhai, Z.-C., and Wang, L.-S. (2005) *QSAR Comb. Sci.*, **24**, 211–217.
20 Yu, X., Wang, X., Gao, J., Li, X., and Wang, H. (2005) *Polymer*, **46**, 9443–9451.
21 Yu, X., Yi, B., and Wang, X. (2007) *J. Comput. Chem.*, **28**, 2336–2341.
22 Yu, X., Yi, B., Xie, Z., Wang, X., and Liu, F. (2007) *Chemom. Intell. Lab. Syst.*, **87**, 247–251.
23 Hobza, P. and Syponer, J. (1999) *Chem. Rev.*, **99**, 3247–3276.
24 Gironés, X. and Carbó-Dorca, R. (2006) *QSAR Comb. Sci.*, **25**, 579–589.
25 Carbó-Dorca, R. (2007) *SAR QSAR Environ. Res.*, **18**, 265–284.
26 Gallegos, A., Robert, D., Gironés, X., and Carbó-Dorca, R. (2001) *J. Comput. Aided Mol. Des.*, **15**, 67–80.
27 Carbó-Dorca, R., Robert, D., Amat, L., Gironés, X., and Besalú, E. (2000) *Molecular Quantum Similarity in QSAR and Drug Design*, Springer Verlag, Berlin.
28 Gironés, X., Gallegos, A., and Carbó-Dorca, R. (2000) *J. Chem. Inf. Comput. Sci.*, **40**, 1400–1407.
29 Amat, L., Besalú, E., and Carbó-Dorca, R. (2001) *J. Chem. Inf. Comput. Sci.*, **41**, 978–991.
30 Lobato, M., Amat, L., Besalú, E., and Carbó-Dorca, R. (1997) *Quant. Struct.-Act. Relat.*, **16**, 465–472.
31 Ponec, R., Amat, L., and Carbó-Dorca, R. (1999) *J. Comput. Aided Mol. Des.*, **13**, 259–270.
32 Mezey, P.G. (1993) *Shape in Chemistry: An Introduction to Molecular Shape and Topology*, VCH Publishers, New York.
33 Mezey, P.G. (1996) *J. Chem. Inf. Comput. Sci.*, **36**, 1076–1081.
34 Mezey, P.G., Zimpel, Z., Warburton, P., Walker, P.D., Irvine, D.G., Dixon, D.G., and Greenberg, B. (1996) *J. Chem. Inf. Comput. Sci.*, **36**, 602–611.
35 Mezey, P.G. (1999) *Mol. Phys.*, **96**, 169–178.
36 Riess, I. and Münch, W. (1981) *Theor. Chim. Acta*, **58**, 295–300.
37 Matta, C.F. (2001) *J. Phys. Chem.*, **105**, 11088–11101.
38 Bader, R.F.W. (1990) *Atoms in Molecules: A Quantum Theory*, Oxford University Press, Oxford, UK.
39 Matta, C.F. and Bader, R.F.W. (2000) *Proteins Struct., Funct., Genet.*, **40**, 310–329.
40 Matta, C.F. and Bader, R.F.W. (2002) *Proteins Struct., Funct., Genet.*, **48**, 519–538.
41 Matta, C.F. and Bader, R.F.W. (2003) *Proteins Struct., Funct., Genet.*, **52**, 360–399.
42 Matta, C.F. and Boyd, R.J. (eds) (2007) *The Quantum Theory of Atoms in Molecules: From Solid state to DNA and Drug Design*, Wiley-VCH Verlag GmbH, Weinheim.
43 Popelier, P.L.A. (1999) *J. Phys. Chem.*, **103**, 2883–2890.
44 O'Brien, S.E. and Popelier, P.L.A. (1999) *Can. J. Chem.*, **77**, 28–36.
45 O'Brien, S.E. and Popelier, P.L.A. (2001) *J. Chem. Inf. Comput. Sci.*, **41**, 764–775.
46 Popelier, P.L.A., Smith, P.J., and Chaudry, U.A. (2004) *J. Comput. Aided Mol. Des.*, **18**, 709–718.

47 Loader, R.J., Singh, N., O'Malley, P.J., and Popelier, P.L.A. (2006) *Bioorg. Med. Chem. Lett.*, **16**, 1249–1254.

48 Roy, D.R., Parthasarathi, R., Subramanian, V., and Chattaraj, P.K. (2006) *QSAR Comb. Sci.*, **25**, 114–122.

49 Roy, K. and Popelier, P.L.A. (2008) *Bioorg. Med. Chem. Lett.*, **18**, 2604–2609.

50 Roy, K. and Popelier, P.L.A. (2008) *QSAR Comb. Sci.*, **27**, 1006–1012.

51 Esteki, M., Hemmateenejad, B., Khayamian, T., and Mohajeri, A. (2007) *Chem. Biol. Drug Des.*, **70**, 413–423.

52 Mohajeri, A., Hemmateenejad, B., Mehdipour, A., and Miri, R. (2008) *J. Mol. Graphics Modell.*, **26**, 1057–1065.

53 Jezierska, A., Panek, J., Ryng, S., Glowiak, T., and Koll, A. (2003) *J. Mol. Model.*, **9**, 159–163.

54 Mazza, C.B., Sukumar, N., Breneman, C.M., and Cramer, S.M. (2001) *Anal. Chem.*, **73**, 5457–5461.

55 Yao, S.-W., Lopes, V.H.C., Fernández, F., and García-Mera, X. (2003) *Bioorg. Med. Chem.*, **11**, 4999–5006.

56 Barone, P.M.V.B., Camilo, A.J., and Galvão, D.S. (1996) *Phys. Rev. Lett.*, **77**, 1186–1189.

57 Vendrame, R., Braga, R.S., Takahata, Y., and Galvão, D.S. (1999) *J. Chem. Inf. Comput. Sci.*, **39**, 1094–1104.

58 Clare, B.W. (2002) *J. Comput. Aided Mol. Des.*, **16**, 611–633.

59 Clare, B.W. and Supuran, C.T. (2005) *Bioorg. Med. Chem.*, **13**, 2197–2211.

60 Deeb, O. and Clare, B.W. (2007) *Chem. Biol. Drug Des.*, **70**, 437–449.

61 Deeb, O. and Clare, B.W. (2008) *Chem. Biol. Drug Des.*, **71**, 352–362.

62 Bowen-Jenkins, P.E. and Richards, W.G. (1986) *Int. J. Quantum Chem.*, **30**, 763–768.

63 Boon, G., Langenaeker, W., De Proft, F., De Winter, H., Tollenaere, J.P., and Geerlings, P. (2001) *J. Phys. Chem. A*, **105**, 8805–8814.

64 Bultinck, P. and Carbó-Dorca, R. (2005) *J. Chem. Sci.*, **117**, 425–435.

65 Gironés, X., Amat, L., Robert, D., and Carbó-Dorca, R. (2000) *J. Comput. Aided Mol. Des.*, **14**, 477–485.

66 Khandogin, J. and York, D.M. (2004) *Proteins Struct., Funct., Bioinf.*, **56**, 724–732.

67 Stenberg, P., Norinder, U., Luthman, K., and Artursson, P. (2001) *J. Med. Chem.*, **44**, 1927–1937.

68 Maran, U., Karelson, M., and Katritzky, A.R. (1999) *QSAR Comb. Sci.*, **19**, 3–10.

69 Katritzky, A.R., Dobchev, D.A., Fara, D.C., and Karelson, M. (2005) *Bioorg. Med. Chem.*, **13**, 6598–6608.

70 Katritzky, A.R., Dobchev, D.A., Hür, E., Fara, D.C., and Karelson, M. (2005) *Bioorg. Med. Chem.*, **13**, 1623–1632.

71 Katritzky, A.R., Kulshyn, O.V., Stoyanova-Slavova, I., Dobchev, D.A., Kuanar, M., Fara, D.C., and Karelson, M. (2006) *Bioorg. Med. Chem.*, **14**, 2333–2357.

72 Katritzky, A.R., Pacureanu, L.M., Dobchev, D.A., Fara, D.C., Duchowicz, P.R., and Karelson, M. (2006) *Bioorg. Med. Chem.*, **14**, 4987–5002.

73 Liu, H.X., Zhang, R.S., Yao, X.J., Liu, M.C., Hu, Z.D., and Fan, B.T. (2004) *J. Comput. Aided Mol. Des.*, **18**, 389–399.

74 Kriegl, J.M., Arnhold, T., Beck, B., and Fox, T. (2005) *QSAR Comb. Sci.*, **24**, 491–502.

75 Gafourian, T., Safari, A., Adibkia, K., Parviz, F., and Nokhodchi, A. (2007) *J. Pharm. Sci.*, **96**, 3334–3351.

76 Alvarez-Ginatre, Y.A., Crespo-Otero, R., Marrero-Ponce, Y., Noheda-Marin, P., de la Vega, J.M.G., Montero-Cabrera, L.A., García, J.A.R., Caldera-Luzardo, J.A., and Alvarado, Y.J. (2008) *Bioorg. Med. Chem.*, **16**, 6448–6459.

77 Duchowicz, P.R., Vitale, M.G., Castro, E.A., Autino, J.C., Romanelli, G.P., and Bennardi, D.O. (2008) *Eur. J. Med. Chem.*, **43**, 1593–1602.

78 Mercader, A.G., Duchowicz, P.R., Fernández, F.M., Castro, E.A., Bennardi, D.O., Autino, J.C., and Romanelli, G.P. (2008) *Bioorg. Med. Chem.*, **16**, 7446–7470.

79 Gubskaya, A.V., Kholodovich, V., Knight, D., Kohn, J., and Welsh, W.J. (2007) *Polymer*, **48**, 5788–5801.

80 Estrada, E., Perdomo-López, I., and Torres-Labandeira, J.J. (2001) *J. Chem. Inf. Comput. Sci.*, **41**, 1561–1568.

81 Olivero-Verbel, J. and Pacheco-Londoño, L. (2002) *J. Chem. Inf. Comput. Sci.*, **42**, 1241–1246.

82 Safarpour, M.A., Hemmateenejad, B., and Jamali, M. (2003) *QSAR Comb. Sci.*, **22**, 997–1005.

83 Sorich, M.J., McKinnon, R.A., Miners, J.O., Winkler, D.A., and Smith,

P.A. (2004) *J. Med. Chem.*, **47**, 5311–5317.
84 Szaleniec, M., Witko, M., Tadeusiewicz, R., and Goclon, J. (2006) *J. Comput. Aided Mol. Des.*, **20**, 145–157.
85 Dai, Y., Zhang, X., Wang, H., and Lu, Z. (2008) *J. Mol. Model.*, **14**, 807–812.
86 Fernández, M. and Caballero, J. (2008) *QSAR Comb. Sci.*, **27**, 866–875.
87 Unger, S.H. and Hansch, C. (1973) *J. Med. Chem.*, **16**, 745–749.
88 Mager, P.P. (1984) *Multidimensional Pharmacochemistry: Design of Safer Drugs*, Academic Press, Inc., London.
89 Kholodovych, V., Smith, J.R., Knight, D., Abramson, S., Kohn, J., and Welsh, W.J. (2004) *Polymer*, **45**, 7367–7379.
90 Gini, G. and Lorenzini, M. (1999) *J. Chem. Inf. Comput. Sci.*, **39**, 1076–1080.
91 Molfetta, F.A.d., Angelotti, W.F.D., Romero, R.A.F., Montanari, C.A., and Silva, A.B.F.d. (2008) *J. Mol. Model.*, **14**, 975–985.
92 Smith, J.R., Knight, D., Kohn, J., Rasheed, K., Weber, N., Kholodovych, V., and Welsh, W.J. (2004) *J. Chem. Inf. Comput. Sci.*, **44**, 1088–1097.
93 Smith, J.R., Kholodovych, V., Knight, D., Kohn, J., and Welsh, W.J. (2005) *Polymer*, **46**, 4296–4306.
94 Murthy, S.K. (1998) *Data Min. Knowl. Discovery*, **2**, 345–389.
95 P. L. R. Research C5.0, v.5.0 (2002) 5.0 ed., St Ives NSW 2075, Australia, 2002.
96 Shannon, C.E. (1948) *Bell System Tech. J.*, **27**, 379–423.
97 Shannon, C.E. (1948) *Bell System Tech. J.*, **27**, 623–656.
98 Hashemianzadeh, M., Safarpour, M.A., Gholamjani-Moghaddam, K., and Mehdipour, A.R. (2008) *QSAR Comb. Sci.*, **27**, 469–474.
99 Alvarez-Ginatre, Y.M., Crespo, R., Montero-Cabrera, L.A., Ruiz-Garcia, J.A., Ponce, Y.M., Santana, R., Pardillo-Fontdevila, E., and Alonso-Becerra, E. (2005) *QSAR Comb. Sci.*, **24**, 218–225.
100 Isayev, O., Rasulev, B., Gorb, L., and Leszczynski, J. (2006) *Mol. Diversity*, **10**, 233–245.
101 Fernández, M. and Caballero, J. (2007) *J. Mol. Model.*, **13**, 465–476.
102 Hansch, C. and Fujita, T. (1964) *J. Am. Chem. Soc.*, **86**, 1616–1626.
103 Free, S.M. Jr and Wilson, J.W. (1964) *J. Med. Chem.*, **7**, 395–399.
104 Reis, M., Lobato, B., Lameira, J., Santos, A.S., and Alves, C.N. (2007) *Eur. J. Med. Chem.*, **42**, 440–446.
105 Liu, W., Yi, P., and Tang, Z. (2006) *QSAR Comb. Sci.*, **25**, 936–943.
106 Yu, X., Yi, B., Wang, X., and Xie, Z. (2007) *Chem. Phys.*, **332**, 115–118.
107 Yu, X., Xie, Z., Yi, B., Wang, X., and Liu, F. (2007) *Eur. Polym. J.*, **43**, 818–823.
108 Yu, X., Wang, X., Li, X., Gao, J., and Wang, H. (2006) *J. Polym. Sci., Part B: Polym. Phys.*, **44**, 409–415.
109 Liu, A., Wang, X., Wang, L., Wang, H., and Wang, H. (2007) *Eur. Polym. J.*, **43**, 989–995.
110 Chaviara, A.T., Kioseoglou, E.E., Pantazaki, A.A., Tsipis, A.C., Karipidis, P.A., Kyriakidis, D.A., and Bolos, C.A. (2008) *J. Inorg. Biochem.*, **102**, 1749–1764.
111 Song, Y., Zhou, J., Zi, S., Xie, J., and Ye, Y. (2005) *Bioorg. Med. Chem.*, **13**, 3169–3173.
112 Pasha, F.A., Neaz, M.M., Cho, S.J., and Kang, S.B. (2007) *Chem. Biol. Drug Des.*, **70**, 520–529.
113 Wan, J., Zhang, L., Yang, G., and Zhan, C.-G. (2004) *J. Chem. Inf. Comput. Sci.*, **44**, 2099–2105.
114 Eroglu, E. and Türkmen, H. (2007) *J. Mol. Graphics Modell.*, **26**, 701–708.
115 Zhang, L., Wan, J., and Yang, G. (2004) *Bioorg. Med. Chem.*, **12**, 6183–6191.
116 Mehdipour, A.R., Hemmateenejad, B., and Miri, R. (2007) *Chem. Biol. Drug Des.*, **69**, 362–368.
117 Holder, A.J., Ye, L., Eick, J.D., and Chappelow, C.C. (2006) *QSAR Comb. Sci.*, **25**, 905–911.
118 Van Damme, S., Langenaeker, W., and Bultinck, P. (2008) *J. Mol. Graphics Modell.*, **26**, 1223–1236.
119 Hutter, M.C. (2003) *J. Comput. Aided Mol. Des.*, **17**, 415–433.
120 Deeb, O., Youssef, K.M., and Hemmateenejad, B. (2008) *QSAR Comb. Sci.*, **26**, 417–424.
121 Toropov, A.A., Rasulev, B.F., and Leszczynski, J. (2008) *Bioorg. Med. Chem.*, **16**, 5999–6008.
122 Katritzky, A.R., Pacureanu, L., Dobchev, D., and Karelson, M. (2007) *J. Mol. Model.*, **13**, 951–963.

123 Matysiak, J. (2008) *QSAR Comb. Sci.*, **27**, 607–617.

124 Lu, G.-N., Dang, Z., Tao, X.-Q., Yang, C., and Yi, X.-Y. (2008) *QSAR Comb. Sci.*, **27**, 618–626.

125 Niu, J., Long, X., and Shi, S. (2007) *J. Mol. Model.*, **13**, 163–169.

126 Ivanciuc, O. (2007) *Reviews in Computational Chemistry*, vol. 23 (eds K.B. Lipkowitz and T.R. Cundari), Wiley-VCH Verlag GmbH, Weinheim.

127 Li, J., Qin, J., Liu, H., Yao, X., Liu, M., and Hu, Z. (2008) *QSAR Comb. Sci.*, **27**, 157–164.

128 Zheng, G., Xiao, M., and Lu, X. (2007) *QSAR Comb. Sci.*, **26**, 536–541.

129 Vapnik, V. (1995) *The Nature of Statistical Learning Theory*, Springer, New York.

130 Vapnik, V. and Lerner, A. (1963) *Automat. Remote Contr.*, **24**, 774–780.

131 Suykens, J.A.K. and Vandewalle, J. (1999) *Neural Process. Lett.*, **9**, 293–300.

132 Scholkopf, B., Smola, A.J., Williamson, R.C., and Bartlett, P.L. (2000) *Neural Comput.*, **12**, 1207–1245.

133 Sorich, M.J., Miners, J.O., McKinnon, R.A., and Smith, P.A. (2004) *Mol. Pharmacol.*, **65**, 301–308.

134 Bultinck, P., Langenaeker, W., Lahorte, P., De Proft, F., Geerlings, P., Waroquier, M., and Tollenaere, J.P. (2002) *J. Phys. Chem. B*, **106**, 7887–7894.

135 Bultinck, P. and Carbó-Dorca, R. (2002) *Chem. Phys. Lett.*, **364**, 357–362.

136 Gao, J., Wang, X., Yu, X., Li, X., and Wang, H. (2006) *J. Mol. Model.*, **12**, 521–527.

137 Tang, Y., Chen, K.-X., Jiang, H.-L., and Ji, R.-Y. (1998) *Eur. J. Med. Chem.*, **33**, 647–658.

138 Hu, L.-H., Chen, G.-H., and Chau, R.M.-W. (2006) *J. Mol. Graphics Modell.*, **24**, 244–253.

139 Yaffe, D., Cohen, Y., Espinosa, G., Arenas, A., and Giralt, F. (2002) *J. Chem. Inf. Comput. Sci.*, **42**, 162–183.

140 Nguyen-Cong, V. and Rode, B.M. (1996) *Eur. J. Med. Chem.*, **31**, 479–484.

141 Novic, M., Nikolovska-Coleska, Z., and Solmajer, T. (1997) *J. Chem. Inf. Comput. Sci.*, **37**, 990–998.

142 Molina, E., Estrada, E., Nodarse, D., Torres, L.A., González, H., and Uriarte, E. (2008) *Int. J. Quantum Chem.*, **108**, 1856–1871.

143 Katritzky, A.R., Karelson, M., and Petrukhin, R. (2001–2005) CODESSA PRO, University of Florida, Florida.

144 Coulson, C.A. (1953) *Adv. Cancer Res.*, **1**, 1–56.

145 Pullman, A. and Pullman, B. (1955) *Adv. Cancer Res.*, **3**, 117–169.

146 Gao, J., Wang, X., Li, X., Yu, X., and Wang, H. (2006) *J. Mol. Model.*, **12**, 513–520.

147 Yu, X., Liu, W., Liu, F., and Wang, X. (2008) *J. Mol. Model.*, **14**, 1065–1070.

148 Kholodovich, V., Gubskaya, A.V., Bohrer, M., Harris, N., Knight, D., Kohn, J., and Welsh, W.J. (2008) *Polymer*, **49**, 2435–2439.

149 Bicerano, J. (2002) *Prediction of Polymer Properties*, Marcel Dekker, New York.

150 Katritzky, A.R. and Gordeeva, E.V. (1993) *J. Chem. Inf. Comput. Sci.*, **33**, 835–857.

151 Environment Directorate OECD (2007) *OECD Environment Health and Safety Publications*, Environment Directorate OECD, Paris.

152 Ferrari, A.M., Sgobba, M., Gamberini, M.C., and Rastelli, G. (2007) *Eur. J. Med. Chem.*, **42**, 1028–1031.