QSAR/QSPR Modeling in the Design of Drug Candidates with Balanced Pharmacodynamic and Pharmacokinetic Properties

George Lambrinidis, Fotios Tsopelas, Costas Giaginis and Anna Tsantili-Kakoulidou

Abstract Drug discovery and development is a slow complicated multi-objective and expensive enterprise. Drug candidates are a compromise output of competing pharmacodynamics and pharmacokinetic processes. To facilitate this task and avoid failures in clinical phases, computational techniques and in silico modeling using the endpoints offered by high technology, are extremely valuable. In this chapter, some historical aspects and a background overview for constructing Quantitative Structure-Activity Relationships (QSAR) and Quantitative Structure-Property Relationships (QSPR) are provided. The different goals for the establishment of QSAR/QSPR models are defined. Representative examples and success stories of in silico modeling along the different drug discovery processes are presented. Examples include models for optimizing efficient binding to receptor, using both ligand- and structure-based approaches, for in vitro permeability predictions, predictions for human intestinal absorption and blood brain barrier penetration, as well as for plasma protein binding and drug metabolism. The value of global and local models as well as their interpretability and the criteria for their evaluation and proper use are discussed throughout this chapter.

G. Lambrinidis · A. Tsantili-Kakoulidou (✉)
Faculty of Pharmacy, Department of Pharmaceutical Chemistry, National and Kapodistrian
University of Athens, Panepistimiopolis, 157 71 Athens, Zografou, Greece
e-mail: tsantili@pharm.uoa.gr

G. Lambrinidis
e-mail: lambrinidis@pharm.uoa.gr

F. Tsopelas
Laboratory of Inorganic and Analytical Chemistry, School of Chemical Engineering,
National Technical University of Athens, Athens, Greece
e-mail: ftsop@central.ntua.gr

C. Giaginis
Department of Food Science and Nutrition, School of Environment,
University of the Aegean, Limnos, Greece
e-mail: cgiaginis@aegean.gr

K. Roy (ed.), *Advances in QSAR Modeling*, Challenges and Advances
in Computational Chemistry and Physics 24, DOI 10.1007/978-3-319-56850-8_9

Keywords Multi-objective drug discovery • Local models • Global models • Model interpretability • Drug-likeness • Affinity predictions • ADME properties predictions

1 Introduction

The advancement of a new chemical entity (NCEs) to become a drug candidate is a slow, complex, expensive and multi task process. Along this long road, identification of the disease and the isolation and validation of the molecular target(s) are the first crucial steps. Next, the right drug candidates to interact with the validated target are designed, synthesized and tested for their preclinical and clinical efficacy and safety (Satyanarayanajois 2011; Speck-Planche and Cordeiro 2015). Despite the great advances in science and technology, this process can take around 15 years with a cost of hundreds of millions of dollars (Paul et al. 2010). In fact, much of this cost comes from failures, which account for 75% of the total drug discovery and development expenses. On the other hand such failures if appropriately consolidated, contribute to the body of knowledge on biological complexity.

To prevent late-stage project interruptions, research is shifted to reduce the uncertainties and obtain a proof of concept (POC) for a molecule as a potential medicine in earlier phases of development. Thus, investigation of the fate of a molecule in the organism, considering appropriate pharmacokinetics as well as safety and adverse reactions profiles should advance in parallel with affinity for the target receptor(s) (Gaviraghi et al. 2001; Swift and Amaro 2013). The fate of drug molecules within the organism is principally controlled by ADME properties which stand for absorption, distribution, metabolism and elimination. (Rogge and Taft 2010; Testa et al. 2005b). Poor absorption and thereupon poor bioavailability have been in the past one of the main reasons for the failure of drug candidates. According to more recent statistics, the most important issues to be confronted are drug efficacy and drug safety, associated mainly with plasma protein binding, metabolism and off target activity (Kola and Landis 2004).

Computer-aided approaches and chemoinformatics, applied during the different stages of the pipeline, permit an effective handling of such failures and uncertainties, facilitate candidate selection and speed up their long journey to the market. Reliable models obtained by Quantitative Structure-Activity Relationships (QSAR) and Quantitative Structure-Property Relationships (QSPR) offer decision support upon rationalizing the drug discovery procedure in line with the Quick Win, Fast Fail concept, allowing a pre-selection of compounds with more chances to succeed in later phases (Owens et al. 2015). In this context, a new scientific area has emerged, defined as pharmacoinformatics, which enables the management of all available information from binding to kinetics and toxicity for safer drug candidates (Goldmann et al. 2014).

In fact, successful drug candidates usually represent a compromise between the numerous, sometimes competing objectives so that the advantages for patients outweigh potential drawbacks and risks. However, in order to benefit from QSAR/QSPR models, the appropriate criteria for their evaluation and thereupon their proper use and/or interpretation are essential. Such criteria as well as the ultimate goal of the models may differ according to the timeline and the particular process modeled.

The present chapter provides an outline of the philosophy, the state of the art and the strategies for QSAR/QSPR generation. Distinction between QSAR and QSPR is primarily associated with the traditional drug design steps, concerning lead optimization for efficient receptor binding and predictions of pharmacokinetic/toxicity properties, respectively. After an overview of the common features for in silico modeling, QSAR models for pharmacodynamics properties, e.g., binding to target receptor(s) or off-target proteins and QSPR models for pharmacokinetic process (ADME properties) are discussed in separate sections. According to the underlying mechanism QSPR models concern both models for passive phenomena and for bonding to proteins. In all cases, two critical interdependent issues are addressed throughout the chapter: (i) the value of global models built on large and chemically diverse datasets and that of local models, built specifically for a series or project, and (ii) the importance or not of model interpretability (Cox et al. 2013; Fujita and Winkler 2016).

2 Historical Aspects and Background

Early QSAR studies were based on the assumption that biological activity can be quantitatively expressed as a function of chemical structure (Brown and Fraser 1868). They involved the establishment of model equations in order to understand and if possible to predict biological activity on the basis of structural parameters, as expressed by equation of type (1).

$$\text{Biological activity} = a_0 + a_1P_1 + a_2P_2 + \cdots + a_nP_n \tag{1}$$

where $P_1 \ldots P_n$ are physicochemical/molecular properties characterizing the compound structures and a_o $a_1 \ldots a_n$ the constants derived by multiple linear regression analysis (Hansch et al. 1995b; Hansch and Fujita 1964; Martin 1978).

Although biological activity was not always considered at the molecular level, it was recognized as an essential prerequisite that the analyzing compounds should act at the same receptor and with the same mechanism of action. Within a congeneric series it was assumed that all other factors influencing the manifestation of drug action should have similar impact. In regard to the description of chemical structure, the well-known Hansch analysis recognized three major categories of physicochemical parameters, namely lipophilicity, electronic properties and steric (geometric) properties (Eq. 2).

$$\mathrm{logBR} = -\mathrm{alogP}^2 + \mathrm{blogP} + \rho\sigma + \delta \mathrm{E}\varsigma + \mathrm{c} \tag{2}$$

where logBR is a general expression for biological activity in its logarithmic form to be linearly related to free energy, logP is the logarithm of octanol-water partition coefficient, the widely accepted measure of lipophilicity, σ Hammett's electronic substituent constant and $\mathrm{E}\varsigma$ Taft's steric substituent constant (Hansch 1969; Hansch and Fujita 1964).

Evidently, early QSAR models could be developed only for congeneric compounds, having a common skeleton and different substituents. In those models, lipophilicity was considered as the physicochemical property of primary importance, since it was understood to influence both pharmacokinetics and pharmacodynamics (Kubinyi 1979; Leo et al. 1971; Pliška et al. 1996; Van de Waterbeemd and Testa 1987). A parabolic relationship between lipophilicity and membrane passage was assumed; thus the quadratic term in Eq. 2 reflects transport to the active site, considering all other pharmacokinetic issues equal within a congeneric series (Hansch and Clayton 1973). Since, the parabolic relationship between potency and logP did not fit all datasets, Kubinyi proposed a bilinear relationship, which allows for different slopes at low and high logP values (Kubinyi and Kehrhahn 1978). At the same time calculation methods for logP were developed, based on the additivity principle. The hydrophobic substituent constant π and soon later the hydrophobic fragmental constant *f* or their Σπ and Σ*f*, accounting for all substituents/fragments on the parent structure, could replace logP of the whole molecule, in line with the other substituent constants in Hansch analysis (Hansch and Leo 1979; Rekker and Mannhold 1992).

In fact, Hansch analysis, firstly applied in agrochemistry, drug design, toxicology, industrial and environmental chemistry (Dunn 1988; Hansch et al. 1995a, 1963; Muir et al. 1967), marked a breakthrough in the way of thinking in medicinal chemistry and the start of the new discipline of QSAR (Ganellin 2004), with the mission to exploit the increasing amount of information in the aim to facilitate drug discovery.

Since those early days, QSAR has undergone a tremendous evolution in regard to all aspects, the target end points, the structural representation, the implemented statistical tools, as well as its own standpoints (Cherkasov et al. 2014; Cramer 2012; Puzyn et al. 2010; Tsantili-Kakoulidou and Agrafiotis 2011). In view of biological complexity QSAR has adapted to the multi-task concept, taking advantage of technological achievements and moving from the perception of single-objective drug design to the multi-objective drug discovery and development (Fujita and Winkler 2016; Jorgensen 2004; Speck-Planche and Cordeiro 2015). The multiple tasks addressed by QSAR/QSPR and the tools implemented to construct the models are illustrated in Fig. 1.

Thus, QSAR/QSPR models are generated to address two goals, each of which has its own value: One goal is to establish models which provide an insight of the properties or chemical features that correlate with a biological assay and thereupon an understanding of the mechanism of action. Such models are valuable support for

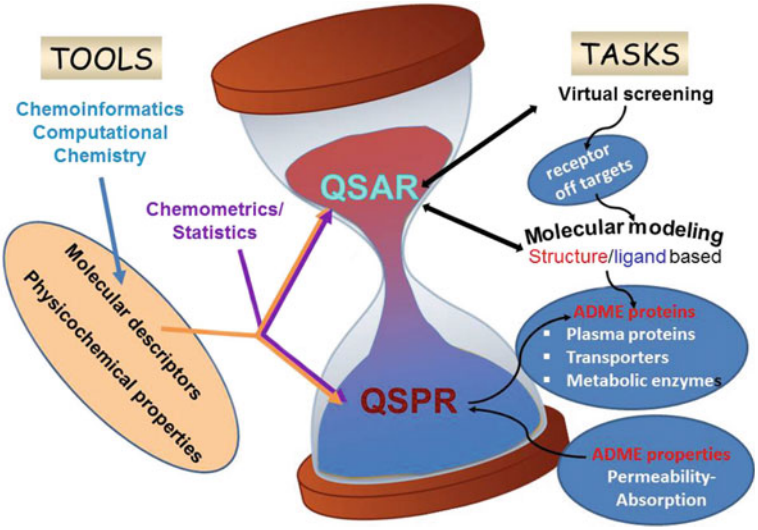

Fig. 1 Tasks addressed by QSAR/QSPR and tools implemented in model construction

the design of novel compounds with affinity to a target protein. The second goal is to create models, which provide accurate prediction of large chemically diverse datasets and address a variety of biological endpoints, as well as different pharmacokinetic processes. Such models allow ranking of compounds prior to synthesis or set priorities among drug candidates for proceeding to further development (Birchall et al. 2008a, b; Nicolotti et al. 2002).

3 Experimental Data and Endpoints in QSAR/QSPR

The multi-objective QSAR starts with data analysis for hit identification, followed by hit-to-lead optimization (lead discovery) and lead optimization (Jorgensen 2009). For hit identification, virtual screening has gained a crucial role, as a consequence also of the continuous emergence of novel biological targets (Schneider 2010; Vasudevan and Churchill 2009). QSAR end-points are usually measured at the molecular or cellular level. The advent of robotized biological testing in the 1990s (Ashour et al. 1987; Houston and Banks 1997; Löfås and Johnsson 1990; Navratilova et al. 2007) has led to the creation of large databases, freely accessible in the public domain, which incorporate millions of compounds with associated bioactivities. PubChem (https://pubchem.ncbi.nlm.nih.gov) and ChemSpider (http://www.chemspider.com), the two major collections of chemical structures on the web, currently include over 30 million compounds each. ZINC (http://zinc.docking.org), a database frequently used for virtual screening applications, incorporates a total of approximately 21 million compounds (Irwin 2008; Moura Barbosa and Del Rio 2012; Wang et al. 2012). In such databases the results of many screens are presented in the form of scores for many compounds on a given assay, while they also contain information on the structures of compounds and the target of particular assays. More detailed data about binding assays can also be found in Binding Database (www.bindingdb.org) which is a public web-accessible database of measured binding affinities containing more than 1 million binding data for nearly 500,000 small molecules and thousands of proteins (Gilson et al. 2016).

However there is a warning on the use of the databases, since they may include inconsistencies concerning both chemical and biological data, while the chemical structures may be inaccurate or presented in a non-consistent way. Therefore curation of the data sets is recognized as a critical step for the establishment of good quality models (Akhondi et al. 2012; Cherkasov et al. 2014).

More to the point, there are databases with sets of inactive compounds (decoys) for several biological targets together with a small set of known active compounds (Mysinger et al. 2012) or even software to produce decoy datasets based on similarity with known active compounds (Cereto-Massagué et al. 2012). Decoy data sets are useful for validation of the QSAR/QSPR models.

When searching in structural databases for experimental binding affinities, one could find different biological data. They may be expressed as continuous response such as IC_{50}, EC_{50}, K_i, K_d, % inhibition, etc., or as categorical response, e.g., active/inactive. Continuous response values are preferably used in their negative logarithms, so as to be in linear correlation with free energy. In line with this concept, ChEMBL database introduced the pChEMBL activity value, defined as $-\log(IC_{50}$, XC_{50}, EC_{50}, AC_{50}, K_i, K_d or Potency) in M units (Papadatos et al. 2015). This value allows a number of roughly comparable measures of half-maximal response concentration/potency/affinity to be compared on a negative logarithmic scale (https://www.ebi.ac.uk/chembl/faq#faq67). This approach has also been implemented in software for large scale off-target pharmacology and predictive safety of small molecule such as CTLink (http://www.chemotargets.com).

Besides the compound databases, there is also a wealth of deposited gene expression data available for downloading and/or online interrogation For example, the NCBI gene expression omnibus (GEO) (Barrett et al. 2007) hosts over half a million single array chip expression profiles and the EBI hosts the Array Express database (Parkinson et al. 2010) with a similar largely overlapping number of arrays. Gene expression-based screening (GE-HTS) represents a strategy for identifying modulators of biological processes with little a priori information about their underlying mechanisms. It is mainly used in cancer research, where it detects compounds, which may revert undesired oncogenic states to nonmalignant or drug-sensitive states (Evans and Guy 2004; Williams 2012). It is evident that for the screening procedure, good prediction models are necessary, complying with the second goal as described in Sect. 2. In such case model interpretability is not a priority. In contrast, the transition from hit identification to lead discovery and optimization requires models which should provide an understanding of the molecular factors involved and a sound physicochemical interpretation, while in-house affinity measurements of the novel compounds are used as endpoints.

The range of affinity values is a crucial issue for model construction. Generally it should be significantly greater than the experimental error among the biological data. Considering that such errors can often exceed half a log unit (Gedeck et al. 2006) it is recommended an endpoint value range of at least 1.0 log unit to obtain a reasonable QSAR model (Cherkasov et al. 2014).

Lead optimization in regard to other pharmaceutical properties, while maintaining affinity, is a next important step. This is a multi-objective process involving many experimental parameters (assays) related to physicochemical properties, ADME properties, plasma and tissue protein binding, target selectivity, off-target activities and toxicity. These properties influence considerably the efficacy and safety of drug candidates and are potential causes for attrition. Rapid in vitro measurements have been and are being developed for permeability and for plasma protein binding assessment and toxicity protocols have been established (Artursson et al. 2001; Kansy et al. 1998; Kariv et al. 2001; Rich and Myszka 2000). On the other hand, there are many efforts for in silico prediction of many of these endpoints by constructing appropriate QSARs or QSPRs (A Cabrera-Perez et al. 2012; Dearden 2007; Lambrinidis et al. 2015; Swift and Amaro 2013). Certain global models for toxicity predictions are approved by OECD and provide support to regulatory authorities (Larregieu and Benet 2013). More to the point, predictions on secondary targets may be useful for the safety profile as well as for drug repurposing (Hodos et al. 2016; Sheridan et al. 2015).The implementation of QSAR/QSPR in the complex drug discovery process is demonstrated in Fig. 2.

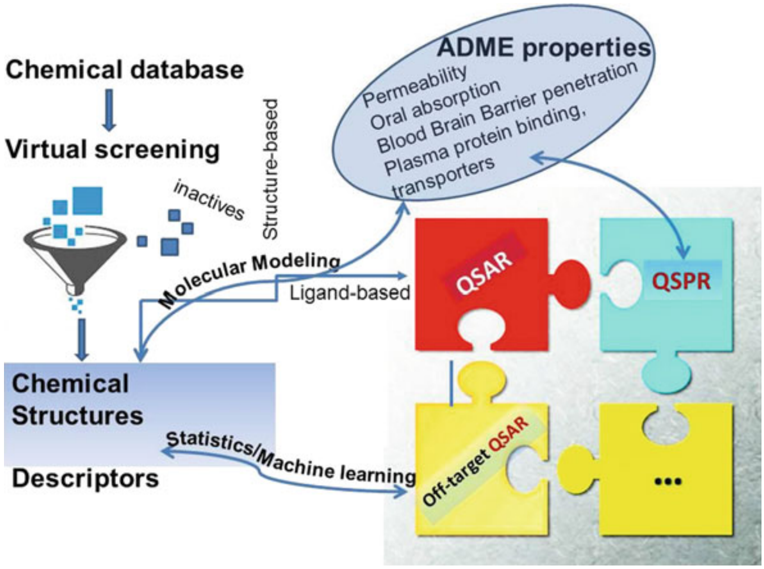

Fig. 2 Implementation of QSAR/QSPR in the drug discovery process

The splitting of QSAR models to encompass various areas of biological complexity has challenged the development of workflows, which integrate QSAR/QSPR models of selected endpoints, including affinities for different target proteins/off-targets and pharmacokinetic data (Cartmell et al. 2005). Consensus predictions using all acceptable models may contribute to further decisions in selecting future experimental screening sets. In inductive knowledge transfer approaches, treating multi-task modeling, the individual QSAR models are not considered separately but they are viewed as nodes in a network of inter-related models (Cherkasov et al. 2014; Qiu et al. 2016). Evidently, the quality of such integrated models largely depends on the quality of the available experimental data compiled in relevant databases, which should be carefully curated, as well as on the range of endpoint values, as already commented (Cherkasov et al. 2014; Gedeck et al. 2006). Interpretability of such models as a prerequisite depends on the purpose and the timeline that they are used along the drug discovery process. In regard to toxicity, for QSAR models to be accepted for regulatory purposes, interpretability is often a crucial issue. According to OECD "*To facilitate the consideration of a QSAR model for regulatory purposes, it should be associated with … a mechanistic interpretation, if possible*" (www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf).

4 Tools Implemented in Model Construction

4.1 Molecular Structure Representation-Descriptors

Molecular structures are represented by descriptors which mediate their relation with activity. Thus, molecular descriptors are at the core of QSAR modeling.

In line with the definition of Todeschini and Consonni (2009), molecular representation has moved forward from substituent constants to variables suitable to portray diverse molecules, belonging to different chemical classes. A variety of software calculates a large number of different physicochemical/molecular properties and theoretical descriptors, starting from SMILES, 2D-chemical graphs to 3D-x, y, z-coordinates or based on mathematical algorithms or statistics. Some of the most popular software are DRAGON, which calculates more than 4000 descriptors (http://www.talete.mi.it/products/dragon_description.htm), ADAPT (Stuper and Jurs 1976) (http://research.chem.psu.edu/pcjgroup/adapt.html), OASIS (Mekenyan and Bonchev 1986), CODESSA (Katritzky et al. 1994), MOE-Chemical Computing Group (https://www.chemcomp.com/) and MolConnZ (http://www.edusoft-lc.com/molconn/).

According to molecular structure representation, descriptors may reflect various levels of dimensionality, ranging from 0D to 4D and xD. 0D are based on molecular formula and are independent from molecular connectivity and conformations. 1D descriptors, reflect the substructure representation of a molecule, 2D descriptors are based on the two-dimensional structural formula (2D), while 3D descriptors are conformation dependent. 3D descriptors are based on thermodynamically favored conformation and necessitate geometry optimization. 4D descriptors reflect interactions with some probe within a grid, while higher dimension (xD) are receptor dependent descriptors. They represent each ligand molecule as an ensemble of conformations, orientations, tautomeric forms and protonation states (Ekins et al. 1999; Hopfinger et al. 1997; Vedani et al. 2000, 2005). Using enhanced molecular dynamic simulations, the overall conformational change of the receptor upon ligand binding can be simulated, producing more vital structural descriptors (Sohn et al. 2013). Such approaches can be considered as a promising link between structure and ligand based strategies (Polanski 2009; Caporuscio and Tafi 2011). An atlas of the available descriptors, the theory used for their calculation and their information content, has been compiled by Todeschini and Consonni (2009). In Table 1, a classification of representative descriptors is presented.

Among the physicochemical descriptors, logP keeps its central role in drug-protein and drug-membrane interactions, as well as in permeability models. Nowadays, there are many algorithms for logP or logD calculation, implemented in relevant software. They are based on the additivity principle and have been developed upon analysis of a large amount of experimental data (Mannhold and Dross 1996). More to the point, calculation of logD necessitates knowledge on pK_a, while charge is a crucial determinant also in drug action (Csizmadia et al. 1997). Actually most of the logP, pK_a and solubility prediction algorithms are QSPR models per se. Some global logP models are implemented in software workflows, which allow the user to utilize his/her own compound library as input in order to refine predictions (Tetko et al. 2001). A comprehensive description and classification of the logP/logD calculation systems and software is provided by Mannhold et al. (Mannhold et al. 2009). Among them, ClogP is often considered as a reference calculation system, while it has been included in most rules for druglikeness (see Sect. 5). Some software for logP/logD prediction are free available on the web.

Despite the large arsenal of available software, the correct selection for logP/logD prediction is not always easy, since often the outcome of the different algorithms shows considerable variations. Although this is not an issue for models intended to screen large compound libraries, it becomes crucial for local models established for lead optimization or for predictions within congeneric compounds (Chrysanthakopoulos et al. 2009; de Melo et al. 2009). In such cases it is important that the compounds analyzed fall within the applicability domain of the training set, used to construct the prediction algorithm (see Sect. 4.3) (Tetko et al. 2009).

Next to lipophilicity, other molecular properties such as molecular volume and surface area, polarizability, molar refractivity, polarity descriptors, dipole moments, hydrogen bond acidity/basicity, as well quantum chemical descriptors, including energy parameters like E_{HOMO} and E_{LUMO}, maximum and minimum electrostatic potentials, partial charges etc., are most commonly used by medicinal chemists. Such descriptors considered as “well-founded”; actually fall within the frame of the three categories: lipophilicity, electronic and geometric descriptors, reflecting the recognition forces and steric requirements of binding to receptor active site.

Table 1 Classification of representative descriptors according to the theory used

Information content/ Theory used	Representative descriptors	Dimensionality	Representative publications
Physical Chemistry	logP solubility ionization constants polarizability molar refractivity …	2D 2D 2D 2D 2D	Hansch et al. 1995b
Molecular properties	molecular weight molecular volume molecular surface hydrogen bond descriptors polarity flexibility-rotatable bonds angles distances …	0D 3D 3D 2D 2D 2D 3D 3D	Helguera et al. 2008
Linear Solvation Energy	hydrogen bond acidity hydrogen bond basicity dipolarity/polarizability excess molar refractivity Mac Goyan Volume	2D 2D 2D 2D 2D	Abraham et al. 2002
Quantum Chemistry	partial charges energy parameters electrostatic potentials dipole moment electron density descriptors HOMO-LUMO energy gap …	3D 3D 3D 3D 3D 3D	De Benedetti and Fanelli 2010

(continued)

Table 1 (continued)

Information content/ Theory used	Representative descriptors	Dimensionality	Representative publications
Molecular Interaction Field, MIF*	GRID descriptors CoMFA CoMSIA	4D 4D 4D	Goodford 1985; Cramer et al. 1988; Klebe et al. 1994
Molecular Interaction Fields (MIF) projected at different energy levels	Volsurf**	2.5D	Cruciani et al. 2000
Representation of molecules as ensemble of conformations, orientations, tautomeric forms, protonation states or any other experimental condition		xD	Vedani et al. 2005
Constitutional	atom counts bond counts number of heavy atoms number of specific atoms …	0D 0D 0D 0D	
Sub structural search	functional groups number of rings fragment counts fingerprints similarity diversity …	2D 2D 2D/3D 2D/3D 2D/3D	Willett 2004
Graph Theory	molecular connectivity electrotopological state indices (E-state) shape indices Wiener indices Balaban indices …	2D 2D 2D 2D 2D	Kier and Hall 1999; Balaban 1997

(continued)

Table 1 (continued)

Information content/ Theory used	Representative descriptors	Dimensionality	Representative publications
Information theory	Entropy Shannon indices		Godden and Bajorath 2000
Autocorrelation descriptors 3D-Autocorrelation descriptors	Topological maximum cross correlation (TMACC) descriptors Maximum Auto-Cross-Correlation MACC descriptors	2D 3D	Spowage et al. 2009; Pastor et al. 2000
Structural keys	MACCS (MDL)	2D	MACCS 2011
QuaSAR descriptors	MOE descriptors (Molecular Operating Environment)	2D/3D	MOE-Chemical Computing Group
Statistical indices	WHIM descriptors GETAWAY descriptors	3D 3D	Gramatica 2006; Consonni et al. 2002

*Interaction Energy Fields value at each point of a Grid for each probe

**The information present in 3D maps, compressed into a few 2D numerical descriptors

Thus they provide insight into the mechanism of action. More to the point, easily calculated physicochemical and molecular properties have created the basis for the development of the drug-like concept (see Sect. 4.1.1).

On the other hand, theoretical descriptors may be considered to reflect a direct detailed representation of molecular structure. However they are not easily interpretable and they do not provide a straightforward perception of the mechanism of action. Their use in QSAR/QSPR models is often faced with some skepticism and their contribution to model quality and validity performance compared to classical descriptors has been questioned in the case of lead optimization (Vallianatou et al. 2013). However, it is true that in some cases the most predictive model may not be the most interpretable (Birchall et al. 2008a, b; Nicolotti et al. 2002). The value of models with high prediction accuracy but low interpretability has already been discussed in Sect. 3.

To obtain information about molecular structure from QSAR/QSPR models with low interpretability, a procedure called *reversible decoding* or *inverse* QSAR is being developed. Topological and molecular signature descriptors are considered to be more suitable for inverse QSAR/QSPR (Faulon et al. 2005; Gozalbes et al. 2002).

Moreover, sub-structural descriptors and molecular fingerprints are important to establish similarity/diversity approaches, which gain increasing interest within the scientific community (Willett 2004). Such approaches are widely used for virtual screening and design of chemical libraries, which aid in the primary identification of promising hits.

Recently, chemical similarity between molecules is being extended to evaluate clinical effects, if combined with information derived from computing similarity based upon lexical analysis of patient package inserts. It is expected, that drugs with highly structurally similarity (both by 2D and 3D comparison) are much more likely to have significant overlap of their clinical effects, compared to drugs that are structurally different (low 2D similarity but high 3D similarity Yera et al. 2014). However in the search of new candidates chemical similarity does not always lead to biological similarity. Structure-Activity landscape may present the so called activity cliffs. Such discontinuities cannot be predicted by statistically derived QSAR models (Guha 2011).

In the case of toxicity predictions the incorporation of biodescriptors (short-term assays) as independent variables is suggested. Such descriptors are derived by in vitro quantitative high through put screening (qHTS) and in combination with chemical descriptors lead to hybrid models, which may exhibit higher accuracy (Sedykh et al. 2011).

Gene expression signatures of a desired biological state, derived from gene expression data are used to screen a compound library to identify compounds that induce this target signature and corresponding phenotype, while they may also be used as descriptors (Hieronymus et al. 2006; Stegmaier et al. 2004).

4.1.1 Drug-Like Filtering

The use of combinatorial methods during the last 30 years has produced a vast number of compounds, which tend to be more lipophilic, less soluble and with higher molecular weight than conventional drug entities (Hertzberg and Pope 2000). Such properties are often associated with unfavorable absorption, poor or inconsistent bioavailability, as well as with lack of selectivity and increased toxicity (Oprea 2000). To face this situation the concept of druglikeness was launched, defining boundaries on the chemical space and functioning as filter to guarantee a physicochemical profile enabling further development (Leeson and Springthorpe 2007; Yusof et al. 2013). Druglikeness provides useful guidelines for early stage drug discovery, following simple rules of thumb, which suggest cut-off values or ranges for certain properties. According to the rule of 5 (RoF), molecular weight (MW) should not exceed 500 Da, calculated lipophilicity (clogP) should not exceed 5, hydrogen bond donor sites (HBD) should not be more than 5, and hydrogen bond acceptor (HBA) sites not more than 10. Upon pairwise violation of these limits, bioavailability problems may occur in the case of orally administered drugs (Lipinski et al. 1997). RoF was further extended including cutoff values or ranges for additional properties, the most common being: Polar Surface Area (PSA) < 140, number of rotatable bonds (ROTB) < 10, Molar Refractivity (MR) in the range of 40–130, number of aromatic rings (AROM) < 3, total number of atoms in the range of 20–70 (Ursu et al. 2011; Veber et al. 2002). Lipophilicity is related also to safety endpoints. Increased relative risk (6:1) for an adverse event may be anticipated for compounds possessing high lipophilicity (ClogP > 3) and low topological polar surface area (TPSA < 75 A) (Hughes et al. 2008). It is also reported that for ClogP > 3 there is a dramatic higher risk for hERG channel inhibition, an endpoint associated with cardiotoxicity (Wager et al. 2011). More strict cutoff values are proposed for compounds intended to act in the Central Nervous System (CNS-likeness Pajouhesh and Lenz 2005). A quantitative estimate of drug-likeness (QED) has been proposed by Bickerton et al. (Bickerton et al. 2012) which relates the similarity of a compound's properties to those of oral drugs based on eight commonly used molecular properties: MW, log P, HBDs, HBAs, PSA, ROTBs, AROMs and count of alerts for undesirable substructures.

For lead compounds the rule of 3 is suggested according to which MW < 300, logP < 3, HD < 3, and HA < 6 (Congreve et al. 2003). The rule of 3 is applicable mainly for fragment-based lead generation.

The rules of thumb are very simple and understandable, however they do not take into account inaccuracies in the prediction of logP and more important they do not consider the receptor demands. For instance, receptors of the PPAR family possess a very large hydrophobic cavity in their active center, requiring lipophilic ligands with high molecular weight, which in many cases violate twice the rule of 5 (Giaginis et al. 2008, 2007). Target specific lipophilicity profiles obtained through calculation of the logP and logD of ligand series for different receptors have recently investigated, showing also other targets where the compound libraries had mean logP ≥ 5, i.e., outside of traditional RoF space with respect to lipophilicity.

Such knowledge in the early stages of drug development is very useful for the formulation strategy in later stage (Bergström et al. 2016).

The advantages of smaller and less lipophilic compounds as safer and more selective drug candidates were further recognized in terms of receptor binding. According to metrics such as ligand efficiency (LE) and ligand lipophilicity efficiency (LLE) affinity is normalized against molecular size, expressed as heavy atoms, or lipophilicity respectively (Abad-Zapatero 2007; Hopkins et al. 2014). Ligand efficiency dependent lipophilicity (LELP) takes both lipophilicity and molecular size into consideration by dividing logP (clogP) by LE (Tarcsay et al. 2012). In terms of thermodynamics, according to the above metrics drug—receptor binding should be optimized in regard rather to the enthalpic component through specific interactions. Such metrics may be used to prioritize drug candidates with quasi equal potency (Hann 2011; Leeson and Springthorpe 2007).

An update on recent applications of efficiency metrics and strategies to control drug-like properties and to replace problematic elements for improving drug design, is recently published by Meanwell (2016).

4.2 Modeling Techniques

Statistical tools mediate the relationship between structural descriptors and the response variable(s) leading either to regression or to classification models. Model building methods are incorporated in different software packages (Bruce et al. 2007). Multiple linear regression (MLR) analysis is a simple and still widely used technique, which however can handle a limited number of variables. Thus, as a first step, variable selection methods are applied to reduce the large number of calculated descriptors to a set which is information rich but as small as possible. Redundant descriptors and descriptors which show low variance or/and collinearity are removed. For further descriptor reduction, stepwise regression approaches are commonly used, with the drawback however that they are local search processes and may converge to local optima (Paterlini and Minerva 2010).

A promising alternative for variable selection is the use of genetic algorithms (GA). GAs explore the descriptor space simultaneously by a population of candidate solutions which compete and recombine, mimicking the process of natural selection (Mitchell 1998).

Reduction of the descriptors space is inherent in multivariate data analysis (MDVA) a popular statistical technique, which permits the simultaneous (not one at a time) treatment of large number of descriptors, while tolerating inter-relation between them (Eriksson et al. 2001; Wold et al. 2001). It is a projection method from a space with high dimensionality to a space with few dimensions (latent variables), characterized as principal components. Principal component analysis (PCA) is a powerful unsupervised classification method. Projection to latent structures defined also as partial least squares (PLS) is the regression extension of PCA. PLS can handle more than one response variables, under the precondition that they are to some degree inter-related. This is very important for multi-target drug design, for toxicity models or for the establishment of activity profiles of antimicrobial or anticancer agents (Vallianatou et al. 2013; Koukoulitsa et al. 2009). PLS analysis generates coefficients for the original variables (descriptors), which permit a straight-forward interpretation of the model.

MLR and PLS are linear methods and any non-linearity should be incorporated through data transformation before the analysis. On the other hand, machine learning (ML) methods are gaining increasingly important roles in the construction of classification and/or prediction models in several steps of the drug discovery process (Tao et al. 2015). They are effective dimension reduction methods, while allowing for non-linearity to be included in the models and the incorporation of variable interactions. Thus they can reflect biological complexity leading to models with high accuracy. Their drawback is their black box character, e.g., the inability for their rationalization and interpretation in chemical terms. Most popular ML techniques are artificial neural networks (ANN) and associative neural networks (ASNN), inspired by the function and structure of neural network correlations in brain, the k-nearest neighbor technique (k-NN), support vector machines (SVM), regression trees (RT) or random forest (RF) (Byvatov et al. 2003; Sakiyama 2009). The latter are also very useful in the creation of gene expression signatures (Lima et al. 2016). An overview of the machine learning methods, used mainly as prediction tools for ADME properties is given in a recent review by Tao et al. (Tao et al. 2015). Table 2 includes commonly used statistical tools, which are referred in the representative QSAR and QSPR examples, discussed in Sect. 5.

Models are evaluated by statistical data, the most commonly being correlation coefficient (R or r) and determination coefficient (R^2 or r^2), standard error of estimate(s), given also as root mean square error of estimate (RMSE). The adjusted determination coefficient (R^2_{adj}) for degrees of freedom allows for comparison between QSARs with different numbers of descriptors and can indicate if a given QSAR model is overfit incorporating too many descriptors. The Fisher test *F* provides an indication of a chance correlation, while the Student test *t* is used to evaluate the significance of descriptors in MLR. In multivariate data analysis, the variable importance to projection (VIP) criterion is used instead. In ANN, the contribution of molecular descriptors is based on the ratio between the performance of neural network before and after the elimination of each descriptor (sensibility analysis).

Visualization of the results, fitting the line on the graph of observed versus predicted values, enables to check for outliers or trends in the data, while it provides an overview of the predictive power of the model. In fact a good model should show an 1:1 correlation between observed and predicted values. Detected outliers should be submitted to further investigation—they may unravel interesting information. Further statistical data are related to model internal or external validation (Sect. 4.3).

For classification models, % sensitivity defined as the ratio of percentage of true positives in respect to the sum of true positives + false negatives, % specificity, defined as the ratio of percentage of true negatives in respect to the sum of true

Table 2 Statistical tools, commonly used in QSAR/QSPR prediction or classification models

	Linear	Non-linear	Prediction	Classification
Multiple Linear Regression Analysis (MLR)	x		x	
Partial Least Square/Projection latent Structures, PLS	x		x	
Principal Component Regression, PCR	x		x	
Principal Component Analysis, PCA	x			x(unsupervised)
PLS-Discriminant Analysis, PLS-DA	x			x (supervised)
Linear Discriminant Analysis (LDA)	x			x (supervised)
Artificial neural networks, ANN Bayesian NN Associative NN		x	x	x (unsupervised/ supervised)
Support Vector Machine, SVM		x	x	x (supervised)
k-Nearest Neighbors		non-parametric	x	x (supervised)
K-means		x	x	x(unsupervised)
Decision trees and Random forests		x	x	x(unsupervised)
Classification and Regression Tree (CART)		x	x	x (supervised)
Ensemble methods-Bagging-Boosting trees		x	x	x (supervised)

negatives + false positives and %CCR (correct classification rate or balanced accuracy) equal to (sensitivity + specificity)/2 are common statistical data to evaluate the merit of the models. It should be noted that acceptance criteria depend on the quality of experimental data, as well as on the ultimate goal of the QSAR/QSPR performed.

4.3 Model Validation

Whatever modeling technique is used, validation of QSAR models has received considerable attention in the last decades (Guha and Jurs 2005; Tropsha et al. 2003; Veerasamy et al. 2011). Validation requirements are becoming increasingly strict so as to assure robust models, which can lead to reliable predictions and to proof of concepts. According to the European center for the validation of alternative methods (ECVAM) four tools, the methods accepted for estimating the prediction accuracy include (i) cross-validation, (ii) bootstrapping, (iii) randomization of the response data, and (iv) external validation (Worth et al. 2004).

Cross-validation as an internal model validation method is usually performed by the ‘leave-one out’ (LOO) or ‘leave many out’ (LMO) procedure to determine PRESS and cross-validated correlation coefficient q^2, which are metrics reflecting the internal predictive ability of the model. In contrast to r^2 which increases with the number of variables included in the model with a tendency to approximate the value of 1, Q^2 follows a quadratic relationship reaching a maximum corresponding to optimal number of variables.

To check that the obtained model is not a result of chance factors, randomization of the Y response is recommended (Rücker et al. 2007). All models obtained with the randomized training set should be inferior, with r^2 and q^2 values around 0 or with negative values respectively for a set with 0% similarity with the original set (Gasteiger et al. 2003; Klopman and Kalos 1985).

A prerequisite for model validation is external validation, either by dividing the data set into training and test sets and rebuilding the models or/and using a blind test set. The errors produced in the predictions should be comparable to those achieved for the training set. Recently, Roy et al. have proposed a modified correlation coefficient r_m^2 as a novel metric for external validation, which represents the actual difference between the observed and predicted response data without consideration of training set mean and taking into account the r^2 with intercept and r_0^2, without intercept. Change of the axes denoting observed and predicted y modified correlation coefficient r'^2_m may be different from r_m^2 A threshold for the difference delta $r_m^2 = \mathrm{abs}(r_m^2 - r'^2_m)$ less than 0.2 and an average $r_m^2 = (r_m^2 + r'^2_m)/2$ higher than 0.5 indicate robustness of the model (Roy et al. 2009; Roy et al. 2012).

Model applicability domain (AD), defined as the region of chemical space where predictions can be made without extrapolation is an important issue that should be taken into consideration for the proper use of QSAR/QSPR. There are different methods for the assessment of applicability domain, for particular types of QSAR models (Jaworska et al. 2005; Netzeva et al. 2005; Sahigara et al. 2012). Distance/leverage based methods are usually applied. In regard to QSAR models for regulatory purposes, OECD clearly states that the AD should be described “*in terms of the most relevant parameters, i.e., usually those that are descriptors of the model*” (Jaworska et al. 2003).

The performance of the models over time, in particular in the case of global QSPR models, has been addressed by continuous updating of the original models, so as to extend the applicability domain allowing predictions for new compounds of different chemotypes (Rodgers et al. 2011).

5 QSAR/QSPR Applications in the Drug Discovery Process

QSAR/QSPR models can be established for all processes across the drug discovery pipeline. Initial virtual screening may be followed by modeling of the affinity of ligand series to the receptor or to other off-target proteins. In parallel, models for permeability and other pharmacokinetic properties like plasma protein binding, affinity to uptake or efflux transporters and metabolic stability may be established to evaluate safety and efficacy of the candidates.

5.1 Modeling Pharmacodynamics

Pharmacodynamic models focus on predictions of receptor affinity. It should be noted however that binding to proteins is governed by the same recognition forces, regardless if they are target receptors, plasma and tissue proteins, metabolizing enzymes or off-target proteins. They reflect interactions between the small molecules and the amino acid residues within the active site of the protein.

Computational techniques to detect and/or and optimize efficient binding involve both ligand- and structure-based methods and are applied to optimize receptor binding as well as to predict ADME properties involving proteins, like plasma protein binding, binding to metabolizing enzymes or transporters (Fig. 2).

5.1.1 Ligand-Based Drug Design (LBDD)

Ligand-based Quantitative Structure-Activity Relationships (QSAR), established by the procedures, already discussed in Sects. 3–5, do not require or ignore knowledge on the structure of the target protein. In most cases, they are two dimensional models, although they may embrace three dimensional information by incorporating 3-D descriptors. Such models take advantage of the large number of available descriptors and the progress in the statistical techniques as well as of the associated philosophy (see Sect. 4). They can be further classified as global or local models.

Global models are useful for virtual screening, off target screening or for plasma/tissue protein binding (Helgee et al. 2010; Sheridan 2014). For global models, the goal is to encompass a large applicability domain, while interpretability may not be an issue, at least in the early stages. More important may be the continuous updating of the models to incorporate new chemotypes, so as to expand their applicability domain (Rodgers et al. 2011). In fact, the goal of such global models is not the search for new chemical entities, but to prioritize existing or virtual compounds. In contrast, for lead optimization on receptor binding, local models are more helpful. They are built under the precondition that all analyzed molecules interact with the same type of receptor in the same manner. Evidently, in these cases interpretability defines a determinant factor since the primary goal is to understand the receptor requirements and search for novel compounds with the desired physicochemical/molecular properties. Yet, the inverse-QSAR methodology (see Sect. 4), although based on descriptors which do not confer interpretability, may still allow to construct viable molecules (Wong and Burkowski 2009).

The three dimensional structure of the molecules can serve to create 3-D QSAR models, which provide a direct link to potency. 3D-QSAR has emerged as an extension to the classical 2D-QSAR, using robust chemometric techniques, such as PLS. In 3D-QSAR the precondition for identical binding sites in the same relative geometry for all molecules should be strictly obeyed. After geometry optimization, molecules are superimposed and carefully, aligned in a rational and consistent way to create a hypermolecule. A sufficiently large box is positioned around this hypermolecule and a grid distance is defined. Different atomic probes, e.g., a carbon atom, a positively or negatively charged atom, a hydrogen bond donor or acceptor, or a lipophilic probe, are used to calculate field values in each grid point, i.e., the energy values which the probe would experience in the corresponding position of the regular 3D lattice. Using these fields as input descriptors in PLS analysis, principal components, defined by different proportions of the fields, are generated.

The most popular 3D-QSAR methodology is Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA). CoMFA, developed by Cramer in 1988, is based upon the calculated energies of steric and electrostatic fields (Cramer et al. 1988). CoMSIA, instead of interaction fields, calculates similarity indices using a distance-depended Gaussian functional form. Five types of similarity indices are calculated, steric, electrostatic, hydrophobic, and hydrogen-bond donor and acceptor (Klebe 1998). An important advantage of CoMFA and CoMSIA is the graphical representation of the results. 3-D contour maps in CoMFA display the different contributions of the potentials to the activity, while in CoMSIA they highlight the areas within the region occupied by the ligands, that 'favor' or 'dislike' the presence of a structural feature with a given physicochemical property. In this sense the CoMSIA representation is more easily interpretable than CoMFA contour maps.

The difficulties of both methods are associated with the structure alignment, which may affect the results, while it limits their application to strictly similar compounds. The use of a single conformation for a given ligand represents a limitation of 3D-QSAR since the bioactive conformation may not be necessarily the thermodynamically optimal one. Moreover, orientation in the binding site may be ambiguous, especially in the absence of structural information on the biological receptor. To face such problems, higher dimension QSAR methodologies (xD-QSAR) have been developed. Additional dimensions offer the possibility to represent each ligand molecule as an ensemble of conformations, orientations, tautomeric forms and protonation states (Ekins et al. 1999; Hopfinger et al. 1997; Vedani et al. 2005, 2000).

A general drawback of ligand-based QSAR models is the underlying assumption that chemical similarity correlates with biological similarity, considering a rather smooth structure-activity landscape. The presence of activity outliers however shows that this is not always the case and structure-activity landscape may present activity cliffs (Guha 2011). In such cases, outliers deserve special attention and should be investigated separately. Outliers representing activity cliffs can be identified by structure-based methods, like docking or pharmacophore approaches. In this aspect combination of both ligand- and structure-based approaches may provide insight on the behavior of such outliers (Vallianatou et al. 2013).

5.1.2 Structure-Based Drug Design (SBDD)

Structure-based methods rely on detailed knowledge of target protein structures and target protein-ligand complex providing a more straightforward understanding of the mechanistic aspects in drug-receptor interactions. X-ray crystallography as well as NMR have contributed immensely in this field (Anderson 2003).

In the PDB database (http://rcsb.org), more than 120,000 biological macromolecular structures are deposited, covering more than 40,000 organisms and 38,000 distinct protein sequences. However, in order to use those data, a proper and detailed preparation of the protein must be performed (Anderson 2003; Sastry et al. 2013). The preparation process includes hydrogen addition, protonation or deprotonation based on pK_a prediction of acid or basic side chains, and side chain optimization to achieve the optimum number of hydrogen bond interactions. Once the structure of the protein is well studied and analyzed, all essential parts for interactions between the co-crystalized ligand and the protein are gathered to design new optimized molecules. In this aspect, the key issue for a successful structure-based design is the identification of the target and the appropriate binding site. In Fig. 3 a representative crystal structure of a protein-ligand complex and the interaction points is illustrated. In Fig. 3a, PPARα receptor is represented by ribbons in complex with aleglitazar, represented in space-filling way (CPK representation). Figure 3b shows the ligand interaction diagram of aleglitazar inside the binding pocket.

Additionally, the crystal structure of a protein-target can be used for virtual screening procedure. Virtual screening procedures are based on the structure of a protein while a large database is screened and all molecules are ranked based on empirical docking scoring function for binding affinity (Hillisch et al. 2015). Top ranked molecules are than tested in vitro to validate the model, and the new lead compounds are optimized using computer-aided combinatorial techniques (CombiGlide, version 4.1, Schrödinger, LLC, New York, NY, 2016). Thus, using fragment based algorithms, new virtual chemical libraries are designed based on the core skeleton of the hit compound previous, and top ranked "theoretical" molecules are passed to medicinal chemists for synthesis and further in vitro testing.

However prediction of binding constants based on the correlation with docking scores is not always feasible, especially in the case of structurally diverse compounds. ΔG values calculated by molecular docking may have an acceptable calculation error of 2 kcal/mol corresponding to 2 log units of dissociation constants K_d (Enyedy and Egan 2008; Keserü 2001). Moreover, they may show little differentiation, since they are the outcome of enthalpy–entropy compensation (Brandt et al. 2011). Therefore docking calculations alone are not sufficient, if the principal query is to predict binding constants.

In the past years, many success stories have been achieved using structure-based drug design (SBDD). Some representative examples are reported below:

Amprenavir (Agenerae) and nelfinavir (Viracept) (Kaldor et al. 1997) were the first drugs reaching the market designed against HIV protease using SBDD methodology. Zanamivir (Relenza) was designed against neuraminidase (Varghese 1999),

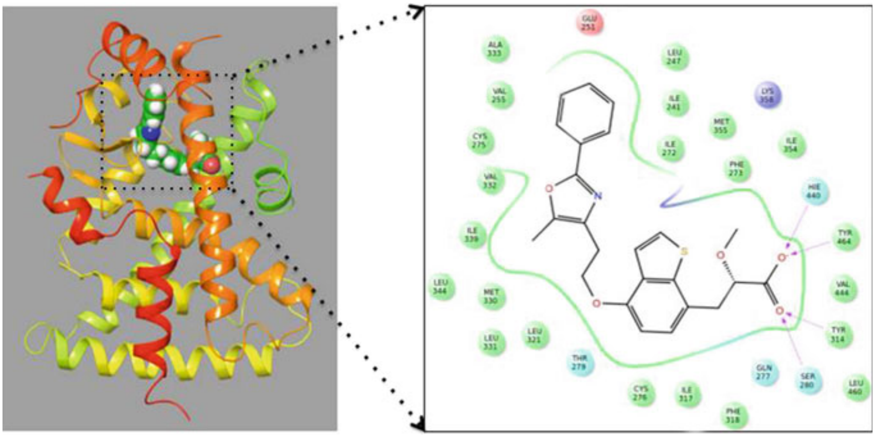

Fig. 3 **a** Ribbon representation of PPARα in complex with aleglitazar (CPK representation), **b** Ligand interaction diagram of Aleglitazar inside the binding pocket. Hydrophobic residues are colored *green*, hydrophilic residues are colored *cyan*, positive charged residues are colored *blue* and negative charged residues are colored *red*. Hydrogen bonds are depicted with *dashed lines*

Tomudex against thymidylate synthase (Rutenber and Stroud 1996) and imitinabmesylate (Glivec) against Abl tyrosine kinase (Schindler et al. 2000). Moreover, SBDD has contributed to address more complicated targets, like nucleic acids as well as protein-protein interactions. Thus, inhibitors have been developed for HIV-1 RNA target TAR (Lind et al. 2002, Filikov et al. 2000), the IL2/IL2Rα receptor interaction (Tilley et al. 1997), the VEGF/VEGF receptor (Wiesmann et al. 1998) and Bcl2 (Enyedy et al. 2001).

5.2 *Modeling Pharmacokinetics*

Pharmacokinetic processes are controlled both by passive phenomena and binding to proteins, the latter concerning plasma and tissue proteins, metabolizing enzymes and transporters. Passive phenomena include passive diffusion through various biological barriers, hemolysis or cell retention. They are governed primarily by lipophilicity, while molecular weight and hydrogen bonding may contribute as additional factors (Avdeef 2012; van de Waterbeemd and Smith 2001). There are also border cases between passive diffusion and binding such as phospholipidosis or drug membrane interactions (Hanumegowda et al. 2010). Volume of distribution is also the outcome of membrane permeability and tissue binding (Hollósy et al. 2006). Among the biological barriers, the gastrointestinal tract and the blood brain barrier are of highest interest and relevant QSPR models are discussed in the following sections.

5.2.1 Modeling Permeability

Several in vitro techniques have been developed for rapid estimation of membrane permeability in vitro. Artificial membranes used in parallel artificial membrane permeability assay (PAMPA) (Kansy et al. 1998) or in immobilized artificial membrane (IAM) chromatography (Tsopelas et al. 2016a, b) provide easy measurements. However, cell-based protocols such as Caco2 or MDCK cell lines are more widely accepted as measures of effective permeability, which is considered as a reliable index mainly for intestinal human absorption (Thiel-Demby et al. 2008; Usansky and Sinko 2005; Volpe 2008; Yee 1997). The Caco-2 model is recommended by the US FDA for the classification of compounds according to the bio-classification system (BCS) (Larregieu and Benet 2013). Several QSPR models to predict Caco-2 or MDCK permeability have been published, which however include a limited number of compounds (Castillo-Garit et al. 2008; Irvine et al. 1999; van De Waterbeemd et al. 1996). It has been shown however from local models, that high Caco-2 permeability rate should correspond to the high human intestinal permeability rate (or extent of absorption), independent of the laboratories of origin and regardless of whether carrier-mediated transport is occurring (Larregieu and Benet 2014).

Due to the considerable inter- and intra-laboratory variability of Caco-2 effective permeability, classification models may be a better option, while meeting the requirements for BCS. Two representative studies performed on large datasets are reported below. Sherrer et al. applied random forest (RF) to the largest dataset ever reported (15791 compounds) to establish a moderate model with a $R^2 = 0.52$, RMSE = 0.20 using 8 descriptors (Sherer et al. 2012). A later model derived by ruled-based decision trees using 1289 compounds achieved determination of 3 permeability classes (High-H, Medium-M, Low-L). The best rule, based on the combination of *PSA*-*MW*-log*D* (3P Rule), was able to identify the H, M and L classes with accuracy of 72.2, 72.9 and 70.6%, respectively, while a consensus system based on three voting binary classification trees predicted 78.4/76.1/79.1% of H/M/L compounds on the training and 78.6/71.1/77.6% on the test set (Pham-The et al. 2013).

Recently, a QSPR study to predict Caco-2 cell permeability was performed on a large data set of 1272 compounds, which were filtered and curated (Wang et al. 2016). Four different methods including multiple linear regression (MLR), partial least squares (PLS), support vector machine (SVM) regression, and boosting trees were employed to build prediction models with 30 molecular descriptors. The nonlinear model derived by Boosting performed better with $R^2 = 0.97$, RMSE = 0.12, $Q^2 = 0.83$, RMSECV = 0.31 for the training set and $R^2 = 0.81$, RMSE = 0.31 for the test set.

5.2.2 Predicting Human Intestinal Absorption/Oral Bioavailability

Considerable efforts are oriented to establish QSPR models for human intestinal absorption and oral bioavailability. Relevant software packages are available either for direct predictions or for predictions of ADME properties like lipophilicity, solubility, ionization, which would allow a rough evaluation of the potential of drugs to be orally absorbed. The rules of thumb, discussed in Sect. 4.1, are very helpful in this case.

Human intestinal absorption (HIA) is usually measured as the percentage of the dose that reaches the portal vein after passing the intestinal wall (%HIA). On the other hand, oral bioavailability (%F) describes the passage of a substance from the site of absorption into the systemic circulation after first pass hepatic metabolism. Intestinal metabolism, acidic stability and the effect of transporters contribute to the outcome. Absorption in gastrointestinal tract is governed by permeability through cell membranes (transcellular absorption) or through the intercellular space between cells of the gastrointestinal mucosa (paracellular transport). The effect of lipophilicity on absorption has been previously described by linear, bilinear, sigmoidal or parabolic models (Kubinyi et al. 1993; Kubinyi and Kehrhahn 1978). However, for the establishment of global QSPR models, which would permit predictions for different chemotypes of novel compounds, additional physicochemical parameters or molecular descriptors, should be implemented. Molecular weight, polarity or hydrogen bonding parameters as well as the charge state are most commonly used, being also consistent to describe Caco-2 permeability as discussed above (Kumar et al. 2011; Tsopelas et al. 2016a, b; Veber et al. 2002).

The main problems to be addressed for the establishment of robust global HIA models concern the significant variability of the datasets from one source to another and the distribution of endpoints, since they include commercially available drugs and are often heavily biased towards compounds with high intestinal absorption values (Hou et al. 2007). This fact will influence the predictive capacity of the in silico models and better predictions will be obtained for compounds with high intestinal absorption values, compared to the rest of the dataset. A scientific and technical report of the European Commission Joint Research Centre and the Institute for Health and Consumer Protection compiles literature models for HIA published till 2010, along with databases with ADME endpoints (Mostrag-Szlichtyng and Worth 2010). In this chapter, representative examples and latest investigations are discussed.

One of the first attempts to predict %HIA was published by Wessel et al. who applied a genetic algorithm with a neural network (GA-NN) technique to develop a non-linear model for set of 86 drugs. They identified six most significant variables, namely: the cube root of gravitational index, related to the size of molecule, the normalized 2D projection of the molecule on the YZ plane (SHDW-6, related to the shape, the number of single bonds (NSB), related to flexibility, as well as the charge on hydrogen bond donor atoms (CHDH-1), the surface area multiplied by the charge of hydrogen bond acceptor atoms (SCAA-s) and the surface area of hydrogen bond acceptor atoms (SAAA-2), related to hydrogen-bonding properties.

The predicted %HIA values achieved good statistics with root mean square errors (RMSE) of 9.4%HIA units for the training set, 19.7%HIA units for the cross-validation (CV) set, and 16.0%HIA units for the external prediction set (Wessel et al. 1998).

The general solvation equation developed by Abraham's group (Abraham et al. 2002) was used by Zhao et to model the human intestinal absorption data of 169 drugs (Zhao et al. 2001). The model Eq. (3) derived by stepwise MLR was based on Abraham's linear solvation energy (LSE) descriptors, namely: excess molar refraction (E), solute polarity/polarizability (S), the McGowan characteristic volume (V), solute overall hydrogen bond acidity (A) and basicity (B).

$$\begin{aligned} &\%\mathrm{HIA} = 92 + 2.94\mathrm{E} + 4.10\mathrm{S} + 10.6\mathrm{V} - 21.7\mathrm{A} - 21.1\mathrm{B} \\ &\mathrm{R}^2 = 0.74,\ \ \mathrm{s} = 14 \end{aligned} \tag{3}$$

According to Eq. (3) the volume and the hydrogen bond descriptors were found to be the most important.

Klopman et al. compiled a large dataset of 467 drug molecules for human intestinal absorption. The data were split into a training set of 417 and external prediction set of 50 molecules. Structural fragments promoting or preventing HIA were identified using the CASE program (http://www.multicase.com/) and their occurrence was subsequently used in a multiparameter linear equation (4) to predict human intestinal absorption (Klopman et al. 2002).

$$\%\mathrm{HIA} = \mathrm{c}_0 + \mathrm{c}_\mathrm{i}\mathrm{G}_\mathrm{i}, \tag{4}$$

where c_0 is a constant, ci are the regression coefficients and Gi is the presence (1) or absence (0) of a certain structural fragment. The final QSAR model included 37 descriptors: 36 statistically significant structural descriptors identified by CASE analysis and one important physicochemical parameter—the number of hydrogen bond donors (H donors). The model was able to predict the %HIA with an $r^2 = 0.79$ and a standard deviation s = 12.32% for the compounds of the training set. The standard deviation for the external test set (50 drugs) was 12.34%. The merit of the model is that it indicates certain substructures with negative impact in %HIA, such as quaternary nitrogens, SO_2 groups connected to an aromatic ring and others with positive impact on HIA. A drawback of the model is that the training set was biased towards high absorption values (Klopman et al. 2002).

Using Zhao's data set, Sun proposed a PLS-DA classification approach for human intestinal absorption modeling, using atom type descriptors. Drugs were classified as classified them as "good" (absorption > 80%) "medium" (80% < absorption > 20%) or "poor" (absorption < 20%), according to their %HIA. A five component PLS-DA model separated very well all 169 compounds with $r^2 = 0.921$ and $q^2 = 0.787$. Since in the case of virtual screening, only poorly absorbed compounds would need to be identified and removed the authors proposed also a three-component PLS-DA with $r^2 = 0.939$ and $q^2 = 0.861$ to separate the compounds with less than 20% absorption (Sun 2004).

Recently, a dataset of 578 compounds, split into a training set of 403 compounds a validation set of 87 and an external prediction set of 87, was analyzed, using ensemble learning (EL) techniques, (gradient boosted tree, GBT and bagged decision tree, BDT) to derive both qualitative (classification) and quantitative models. Topological polar surface area proved to be the most important descriptor with negative contribution, followed by lipophilicity expressed as XlogP. Classification accuracy > 99% was reported, while the QSAR models yielded correlation coefficients $R^2 > 0.91$ between the measured and predicted HIA values (Basant et al. 2016).

Prediction models are available also for the more complex process of oral bioavailability (Andrews et al. 2000; Hou et al. 2007; Kim et al. 2014; Kumar et al. 2011; Martin 2005; Moda et al. 2007; Tian et al. 2011). Till the year 2010 they are compiled in the scientific and technical report of the Joint Research Center of the European Union. In the same report relevant software for prediction of oral bioavailability are provided (Mostrag-Szlichtyng and Worth 2010).

Recently, in silico approaches focus more on physiologically based pharmacokinetics (PBPK), which go beyond human intestinal absorption and oral bioavailability, providing realistic descriptions of absorption, distribution, metabolism, and excretion processes (Bois and Brochot 2016; Jamei 2016). PBPK modeling has gained a significant impact on regulatory science and decisions (Huang et al. 2013) and best practice for its use to address regulatory questions, has been reported (Zhao et al. 2012).

5.2.3 Predicting Blood Brain Barrier Penetration

In drug discovery for CNS active drugs, it is important to determine whether a candidate molecule is capable of penetrating the blood brain barrier (BBB). For drugs targeted at the CNS, the BBB penetration is a necessity, whereas for drugs acting in peripheral tissues, the BBB penetration may lead to undesirable adverse effects (Di et al. 2009; Ecker and Noe 2004). The log BB, defined as the logarithm of the ratio of the concentration of a drug in the brain and in the blood, measured at equilibrium, is an index of BBB permeability. The optimal threshold for classification as a CNS acting drug is typically specified between 0 and −1 (Clark 2003). Log BB values, although widely used, do not take into account plasma and tissue binding, and therefore, do not reflect the free amount of the drug in the brain. Permeability surface area product (PS, quantified as logPS) representing the uptake clearance across the BB is used as a direct measure of permeability and theoretically is not confounded by the plasma and brain tissue binding.

Several models have been published trying to predict blood-brain barrier permeability from various physicochemical properties of molecules, including, among others, molecular size, lipophilicity or number of groups that can establish potential hydrogen bonds (Clark 1999; Kaliszan and Markuszewski 1996;

Konovalov et al. 2007; Luco 1999; Vastag and Keseru 2009). Rules of thumbs are also suggested, as discussed in Sect. 4.1. Till the year 2010, literature models are compiled in the scientific and technical report of the European Commission Joint Research Centre and the Institute for Health and Consumer Protection (Mostrag-Szlichtyng and Worth 2010). Some representative models and recent publications are discussed in this chapter.

Already in 1980, Levin had related log Pc (which is close analog of log PS) to a simple linear function of logP and molecular weight. The overall effect was represented as log $(P * MW^{-1/2})$ = logP−½logMW, whereby increasing log P was supposed to reflect a steady increasing log PS effect, whereas increasing MW had an opposite effect (Levin 1980). In 1999, Clark analyzed a set of 55 diverse organic compounds and generated a multiple linear regression model based on in silico calculated polar surface area (PSA) and logP values with negative and positive contribution respectively (Clark 1999).

The linear solvation energy relationship approach (LSER), also used to model human intestinal absorption, has been applied to blood/brain permeability prediction (Platts et al. 2001). For a dataset of 148 diverse compounds using MLR, they obtained a transparent QSAR incorporating 5 Abraham descriptors and an indicator variable (equal 1 for carboxylic acids and 0 for other compounds) has been reported. The model shows good statistics ($R^2 = 0.74$, $s = 0.34$, $RCV^2 = 0.71$). According to the model, the increasing size of molecules strongly enhances brain uptake, while increasing polarity/polarizability, hydrogen-bond acidity, basicity and the presence of carboxylic acid groups have a detrimental effect. Platt's model has been implemented in the commercially available ADME Boxes software (previously Pharma Algorithms; now ACD Labs, http://www.acdlabs.com/), providing a very fast estimation of logBB. Later, the data set was extended to include 328 compounds with in vivo and in vitro logBB values. A correlation coefficient $r^2 = 0.75$ and a standard deviation $s = 0.3$ was achieved by incorporating an additional indicator for in vitro data (Abraham et al. 2006).

For a data set of 88 diverse compounds using a variable selection and modeling method, a QSAR with three or four descriptors out of 324 descriptors has been reported for logBB prediction. In both models, calculated lipophilicity (AlogP98) was combined either with the atomic type E-state index (SsssN) and Van der Waal's surface ($r = 0.842$, $q = 0.823$, and $s = 0.416$) or with kappa shape index of order 1, atomic type E-state index (SsssN), atomic level based AI topological descriptor (AIssssC) ($r = 0.864$, $q = 0.847$, and SE = 0.392). The success rate of the reported models in test sets was 82% in the case of BBB + compounds. A similar success rate was observed with BBB-compounds (Narayanan and Gunturi 2005).

The VolSurf technique, which is based on molecular interaction fields, has also been used for blood/brain partitioning modeling (Crivori et al. 2000). The model was built on the basis of 230 diverse compounds and more than 70 VolSurf descriptors. Its prediction accuracy (assessed against an external test set) is 90% for BBB permeable molecules and 60% for non-permeable ones. The computational procedure is fully automated and fast and it provides a valuable tool for the virtual screening of large datasets of diverse molecules (Cruciani et al. 2000). The shortcoming of this approach however is its low interpretability.

Linear discriminant analysis (LDA) based on physicochemical descriptors calculated in silico has been used to establish two distinct classification models (Vilar et al. 2010). The data set consisted of the 307 compounds used by Abraham et al. (Abraham et al. 2006) for which in vivo logBB values were available. Considering that molecules with log BB > 0.3 cross the BBB readily while molecules with log BB < -1 are poorly distributed to the brain, these values were selected thresholds for classifying the compounds into two categories. For the threshold 0.3, a two component model was obtained with lipophilicity and topological polar surface area (TPSA), the latter with a negative coefficient. For the threshold-1, the total number of acidic and basic atoms was additionally incorporated, also with a negative sign. The models were validated with external data sets using the area under receiver operating characteristic (ROC) curves as evaluation criterion. In ROC the fraction of true positives (sensitivity) is plotted against the fraction of false positives (1-specificity). An area under the ROC curve of 0.95 for model 1 and 0.97 for model 2 is reported, demonstrating the high predictive power of the models, considering that for a perfect classifier the area under the curve is 1 and for a random classifier it is 0.5 (Vilar et al. 2010).

Based on logPS values in rats, Suenderhauf et al. developed predictive computational models (decision tree induction) for a dataset of 153 compounds. The established models exhibited a corrected classification rate of 90%. The models confirmed the involvement of lipophilicity, molecular size and charge in BBB permeation (Suenderhauf et al. 2012).

5.2.4 Modeling Plasma Protein Binding

A special case of binding of small molecules to macromolecules is plasma protein binding. Plasma protein binding (PPB) is the reversible association of a drug with the proteins of the plasma and is mainly due to hydrophobic and electrostatic interactions. Since only the fraction of unbound (fu) drug is able to pass across cell membranes, PPB strongly influences volume of distribution, half-life and efficacy of drugs. Extended plasma protein binding may be associated with drug safety issues, low clearance, low brain penetration, as well as drug–drug interactions (Ito et al. 1998; Rowley et al. 1997). In fact, plasma protein binding belongs to the ADME properties, representing mainly the "D" of the acronym.

Among the plasma proteins, human serum albumin (HSA) has a central role and the affinity of drugs to this protein is considered to dominate PPB and the thereupon related pharmacokinetic issues. Two primary active sites on HSA have been recognized for drug binding, the Sudlow's sites 1 (warfarin site) and 2 (benzodiazepine site), α_1-acid glycoprotein (AGP) is the second essential plasma protein with two main variants and a complicated physiological role (Lambrinidis et al. 2015).

Modeling of total plasma protein binding or/and of HSA binding has been the objective of many researchers and offers a representative case where combined structure- and ligand-based methods act synergistically. Structure based methods are very helpful to initially classify the compounds according to the preferred binding site or protein, prior to proceeding to ligand-based methods. Since PPB is practically involved in any class of therapeutics, the ultimate goal is to construct global HSA or PPB models, where structural diversity plays an important role. Representative successful efforts are described below. Often more than one model are suggested by the same research group, where interpretability may compete with accuracy in predictions.

A multiple computer-automated structure evaluation method (M-CASE) was used by Saiakhov et al. (Saiakhov et al. 2000) to analyze 154 structurally diverse compounds for total plasma protein binding. M-CASE starts by searching for ‘baseline correlation’ via an internal baseline activity identification algorithm subroutine (BAIA), using the octanol–water partition coefficient which is the most important parameter. For compounds showing residual binding when predicted by the baseline correlation, the algorithm continues to identify responsible structural characteristics, called biophores. Several local QSAR models built for subsets with common biophores are included in the final global model. The binding site(s) of each biophore, including the warfarin, benzodiazepine and digitoxin sites, as well as AGP and lipoproteins, are also characterized. Lipophilicity as the prevalent parameter showed different contribution in each local QSAR, indicating different lipophilicity requirements for each binding site. A crucial structural fragment present in the molecules was found to be part of a phenyl ring. The model, after classifying the compounds according to their biophores, was able to predict correctly the percentage bound to plasma for 80% of the compounds with an average error of 14%.

A large data set of 1008 compounds, partitioned into a training set of 808 compounds and an external validation test set of 200 compounds was used by Votano et al. for model construction of human serum protein binding (Votano et al. 2006). A robust ANN model based of topological descriptors in combination with logP was established with $r^2 = 0.90$, MAE = 7.6 and $r^2 = 0.70$, MAE = 14.1 respectively. MAE stands for Mean Absolute Error.

Votano’s data set was used by Ghafourian et al. (Ghafourian and Amin 2013) to construct linear regression and nonlinear models using classification and regression trees (CART), boosted trees and random forest. Interpretable linear regression and simple regression trees models were able to identify the important contribution of hydrophobicity, van der Waals surface area and aromaticity for high PPB. On the other hand, the more complicated ensemble method of boosted regression trees produced the most accurate PPB predictions.

Combination of chemometrics with molecular modeling confirmed the preponderant contribution of hydrophobic regions of drug molecules and the specific roles of polar groups, which anchor drugs to HSA 1 and 2 binding sites (Estrada et al. 2006). Identification of the binding site before performing QSAR analysis can evidently lead to better models. For 889 chemically diverse compounds with binding affinity for domain III-A, a group contribution model was developed based on 74 chemical fragments. ($R^2 = 0.94$, $Q^2 = 0.90$) (Hajduk et al. 2003).

The authors further suggested a combination of QSAR models for full-length albumin and for domain-IIIA to allow for discrimination between compounds that bind to the latter site and those that bind elsewhere on the protein. An important issue is that the fragments used in the model are mapped by most of the topological descriptors included in Votano's model, indicating that they can be considered quite universal. Thus, they provide a convenient look-up table for quantitatively estimation of the effect of a particular group to albumin binding.

A free web prediction platform was constructed by Zsila et al. who combined support vector machine (SVM) classification model with molecular docking calculations. The classification model was based on 45 descriptors, with logP being the most important. The platform (http://albumin.althotas.com) enables the users (i) to predict if albumin binds the query ligand, (ii) to determine the probable ligand binding site (site 1 or site 2) according to the classification model (iii) to select using the Tanimoto similarity the albumin X-ray structure which is complexed with the most similar ligand and (iv) to calculate complex geometry using molecular docking calculations (Zsila 2013).

The continuous update of the HSA models in order to maintain their performance over time is essential for the drug discovery and development settings, extending their applicability domain and robustness. In this sense, Rodgers et al. proposed a procedure for monthly updating human plasma protein binding models over a period of 21 months (Rodgers et al. 2007), which was extended to three years, using partial least squares (PLS), random forest (RF) and Bayesian neural networks (BNN). The authors started with a large data set, the size of which was doubled by the end of the study (Rodgers et al. 2011). Consensus predictions of HSA binding constants using the final models, generated by all three techniques showed, RMSE = 0.55. These results justified the need for the automatic regular updating of QSAR models (autoQSAR) in the case of ADME properties.

An analogous approach for modeling HSA binding, as well as other ADME properties, over time is implemented in a software architecture, the so called "Discovery Bus" which allows exhaustive exploration of descriptor and model space, automates model validation and their continuous updating providing an automated QSPR through competitive workflow (Cartmell et al. 2005).

Recently, ensemble machine learning-based QSPR models have been established for a four-category classification and PPB affinity prediction, using a dataset of 930 compounds. The structural diversity of the compounds was tested by the Tanimoto similarity index. In the test set, the classification QSPR models proved superior with an accuracy > 93%, while the regression QSPR models yielded $r^2 > 0.920$ between the measured and predicted PPB affinities, with the root mean squared error < 9.77. Lipophilicity, expressed as XLogP, was the most important descriptor (Basant et al. 2016).

For further PPB models and for the state of the art in predicting binding to a1-acid glycoprotein, the second important plasma protein, the reader is referred to a recent comprehensive review by Lambrinidis et al. (2015).

5.2.5 Prediction Models for Metabolism

Metabolism, the M in 'ADME', is one of the main factors influencing the fate and toxicity of a chemical. Metabolism or (biotransformation) includes a large set of chemical reactions, which generally convert drugs or other xenobiotics into more polar and more easily excreted, i.e., less toxic forms. However, in some cases, metabolism may lead to toxic metabolites or/and intermediates. Thus, metabolites with physicochemical and pharmacological properties that differ substantially from those of the parent drug have important implications for both drug safety and efficacy (Testa et al. 2004; Testa 2009).

The utility of conventional QSARs predicting the metabolic fate of chemicals is rather limited. Most of the models are established to predict the phase I metabolism, mainly addressing cytochrome P450 (CYP450) isoforms, a superfamily of enzymes including more than 70 families of proteins, which play a predominant role in the biotransformation of drugs and xenobiotics. Based on a 'guesstimate' of the number of drug metabolites that are known to be produced by cytochromes P450 isoforms and other oxidoreductases (EC 1), as well as hydrolases (EC 3), and transferases (EC 2), it is supposed that oxidoreductases are the main enzymes responsible for the formation of toxic or active metabolites, whereas transferases play the major role in producing inactive and nontoxic metabolites (Testa 2009).

Terfloth et al. (2007) investigated the application of several model-building techniques, such as k-NN, decision trees, Multilayer Perceptron as Neural Networks (MLPNN), Radial Basis Function Neural Networks (RBF-NN), Logistic Regression (LR) and Support Vector Machine (SVM), to predict the isoform specificity for CYP450 3A4, 2D6 and 2C9 substrates (Terfloth et al. 2007). The applied descriptors included simple molecular properties and functional group accounts, topological descriptors, descriptors related to the shape of molecules or the distribution of interatomic distances considering the 3D structures of the molecules. A 9-descriptor model, established by combining automatic variable selection with the SVM technique, gave the best results. The achieved predictivity for an external data set of 233 compounds was equal 83%. Promising results were also obtained for the decision tree based model with three descriptors only, and 80% predictivity for the external data set was achieved. Burton et al. (2006) constructed classification models for human CYP1A2 and CYP2D6 inhibition using binary decision tree. The decision tree for CYP2D6 had sensitivity 88%, specificity 92% and positive predictivity 90%. The external validation hada ccuracy 89%, sensitivity 91%, specificity 92% and precision 90%. For CYP1A2, accuracy was 89%, sensitivity 95%, specificity 83% and precision 85% for the training set while the test set had 81% accuracy, 76% sensitivity, 86% specificity and 85% precision. The authors identified a range of useful descriptors. Van der Waals surface area (VSA) was particularly efficient and allowed to develop models reaching 95% correct classification. 3D descriptors also provided promising results. Sheridan et al. (2007) applied Random Forest (RF) technique for predicting CYP450 (3A4, 2D6, 2C9) sites of the metabolism, using descriptors that describe the environment around each non-hydrogen atom in each molecule. The authors identified several descriptors positively and negatively related to the oxidation sites of molecules. Compared to the results using MetaSite software (Molecular Discovery) of Cruciani et al. (2005), Sheridan's model performed better in the case of CYP3A4. For CYP2D6 and CYP2C9 the predictions of Sheridan's model were only slightly better.

In the case of metabolism, computer-based expert systems have a much broader applicability. Among them MetaSite is widely used (Cruciani et al. 2005). It makes predictions based on the lability of hydrogens and orientation effects derived from the 3D structure of a CYP active site, independently of the availability of pre-existing data. MetaSite can handle 3A4, 2D6, 2C9, 1A2, 2C9, and 2C19 and can be extended to any CYP for which a homology model can be generated. It is advantageous for enzymes such as CYP1A2 and CYP2C19, where there are not currently enough data in the literature to generate a QSAR model. Moreover, the MetaSite methodology is easy to use, fast and fully automated. Other expert systems are MetabolExpert, developed by CompuDrug (Darvas 1988), METEOR (Testa et al. 2005a) COMPACT (Computer-Optimised Molecular Parametric Analysis of Chemical Toxicity) (Lewis et al. 1996; Lewis 2001) and META, implemented in MCASE ADME Module (MultiCASE) (Klopman et al. 1999, 1997; Talafous et al. 1994).

More information about for predicting drug metabolism can be found in a recent review by Kirchmair et al. (2015).

5.2.6 Integrated ADME Prediction Models

In previous sections, separate models for different processes along the drug discovery and development pipeline are discussed. The medicinal chemist team should try to take advantage by applying them in their project compounds, selected by early stage techniques, e.g., virtual screening, structure or ligand based design for the target of interest, drug-like filtering. The multi-objective character of drug development however has challenged the creation of software tools and web platforms mainly for the purpose of integrated ADME and ADME-related predictions. Many of them are commercial. They differ greatly in terms of their capabilities and applications. Prediction software for physicochemical properties like lipophilicity and ionization, related to ADME, has already been discussed in Sect. 4. Solubility is another endpoint of interest for oral absorption as well as for formulation issues. Such predictions serve as inputs to models of key ADME properties, mainly for gastrointestinal absorption, BBB permeability, oral bioavailability (including affinity to uptake or efflux transporter) and plasma protein binding. Predictions of possible metabolite, as well as toxicity endpoints like mutagenicity, carcinogenicity or teratogenicity are also implemented in certain software. Some popular software are Know-it-All (Bio-Rad Laboratories http://www.bio-rad.com/), ADME Boxes (Pharma Algorithms—now included in ACD/ADME Suite), and ADMET Predictor (Simulations Plus Inc. http://www.simulations-plus.com/). VolSurf/VolSurf + (Molecular Discovery and Tripos) also predicts various ADME properties including passive intestinal absorption, blood-brain barrier permeation, solubility, protein binding, volume of distribution, and metabolic stability on the basis of different models based on VolSurf descriptors.

Moreover, there is a trend towards developing more sophisticated, mathematical PBPK models, see also Sect. 5.2.2. In these software tools, in vitro and/or in vivo ADME data are integrated with the results of QSAR/QSPR models (e.g., for percentage plasma protein binding or blood/brain barrier penetration) for organism-based ADME modeling. GastroPlus and Cloe, which mimic the processes inside living organisms, are more commonly used. Simcyp (http://www.simcyp.com/) is a proprietary PBPK simulator that provides a platform for modeling the ADME properties of drugs and their metabolites, as well as drug-drug interactions, in virtual patient populations (Jamei et al. 2009).

It should be noted as a warning for using software for ADME prediction that the results should be considered as rough estimates, useful for screening purposes or as starting points for further modeling or experimental evidence.

6 Conclusions

Drug discovery and development is a complicated multi-objective and expensive enterprise, with drug candidates being a compromise of competing pharmacodynamics and pharmacokinetic processes. In silico predictions along the different stages of the pipeline provide valuable support in the selection of drug candidates with balanced properties, so as to control each stage early enough and reduce failures at clinical phases. High technology provides new endpoints that may serve to establish efficient QSAR and QSPR models, which themselves profit of the evolution in computational and statistical techniques. Local and global models have their own value, dependent on the underlying goal and the timeline. Initial screening, off-target affinities or ADME properties benefit more by global models, while local models are suitable for selected project ligands with potential affinity for a target receptor. Interpretability of models is an important issue. The medicinal chemist is more familiar with models containing well understandable physicochemical or molecular descriptors, which provide an insight in the mechanism of action. However the most accurate model is not always the most interpretable. In such cases the intended use of the model is the determinant factor. Nevertheless, toxicity models for regulatory purpose must have a certain degree of interpretability as required by OECD.

The correct use of the models implies that the user is aware of their merits and pitfalls. Their evaluation should consider the accuracy and range of the endpoints, while external validation with blind test sets is a strict prerequisite in particular for global models. In such cases, determination of their applicability is useful in order to evaluate when predictions are reliable.

In conclusion, the results of the in silico models at the different stages of drug discovery should be taken into consideration for prioritizing the drug candidates, before proceeding to the next step. The ultimate goal is to produce safe and efficient drug candidates, a goal, which can be achieved by finding the golden ratio between affinity to the target receptor, in regard also to off-targets and the appropriate pharmacokinetic properties in compliance with the concept of druglikeness. The tools are available, they need to be properly used.

References

Abad-Zapatero, C. (2007). Ligand efficiency indices for effective drug discovery. *Expert Opinion Drug Discovery, 2,* 469–488. doi:10.1517/17460441.2.4.469.

Abraham, M. H., Ibrahim, A., Zhao, Y., & Acree, W. E. (2006). A data base for partition of volatile organic compounds and drugs from blood/plasma/serum to brain, and an LFER analysis of the data. *Journal of Pharmaceutical Sciences, 95,* 2091–2100.

Abraham, M. H., Ibrahim, A., Zissimos, A. M., Zhao, Y. H., Comer, J., & Reynolds, D. P. (2002). Application of hydrogen bonding calculations in property based drug design. *Drug Discovery Today, 7,* 1056–1063.

Akhondi, S. A., Kors, J. A., & Muresan, S. (2012). Consistency of systematic chemical identifiers within and between small-molecule databases. *Journal of Cheminformatics, 4,* 1.

Anderson, A. C. (2003). The process of structure-based drug design. *Chemistry & Biology, 10,* 787–797.

Andrews, C. W., Bennett, L., & Lawrence, X. Y. (2000). Predicting human oral bioavailability of a compound: Development of a novel quantitative structure-bioavailability relationship. *Pharmaceutical Research, 17,* 639–644.

Artursson, P., Palm, K., & Luthman, K. (2001). Caco-2 monolayers in experimental and theoretical predictions of drug transport. *Advanced Drug Delivery Reviews, 46,* 27–43.

Ashour, M.-B. A., Gee, S. J., & Hammock, B. D. (1987). Use of a 96-well microplate reader for measuring routine enzyme activities. *Analytical Biochemistry, 166,* 353–360.

Avdeef, A. (2012). *Absorption and drug development: solubility, permeability, and charge state*. Wiley.

Balaban, A. T. (Ed.). (1997). From chemical topology to three-dimensional geometry. New York (NY): Plenum Press.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., et al. (2007). NCBI GEO: Mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Research, 35,* D760–D765.

Basant, N., Gupta, S., & Singh, K. P. (2016). Predicting binding affinities of diverse pharmaceutical chemicals to human serum plasma proteins using QSPR modelling approaches. *SAR and QSAR in Environmental Research, 27,* 67–85.

Bergström, C. A., Charman, W. N., & Porter, C. J. (2016). Computational prediction of formulation strategies for beyond-rule-of-5 compounds. *Advanced Drug Delivery Reviews, 101,* 6–21.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry, 4,* 90–98.

Birchall, K., Gillet, V. J., Harper, G., & Pickett, S. D. (2008a). Evolving interpretable Structure-activity relationships. 1. Reduced graph queries. *Journal of Chemical Information and Modeling, 48,* 1543–1557.

Birchall, K., Gillet, V. J., Harper, G., & Pickett, S. D. (2008b). Evolving interpretable structure-activity relationship models. 2. Using multiobjective optimization to derive multiple models. *Journal of Chemical Information and Modeling, 48,* 1558–1570.

Bois, F. Y., & Brochot, C. (2016). Modeling pharmacokinetics. In E. Benfenati (Ed.), *Silico Methods for Predicting Drug Toxicity* (pp. 37–62). New York, NY: Springer New York.

Brandt, T., Holzmann, N., Muley, L., Khayat, M., Wegscheid-Gerlach, C., Baum, B., et al. (2011). Congeneric but still distinct: How closely related trypsin ligands exhibit different thermodynamic and structural properties. *Journal of Molecular Biology, 405,* 1170–1187. doi:10.1016/j.jmb.2010.11.038.

Brown, A. C., & Fraser, T. R. (1868). On the connection between chemical constitution and physiological action; with special reference to the physiological action of the salts of the ammonium bases derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia. *Journal of Anatomy and Physiology, 2,* 224.

Bruce, C. L., Melville, J. L., Pickett, S. D., & Hirst, J. D. (2007). Contemporary QSAR classifiers compared. *Journal of Chemical Information and Modeling, 47,* 219–227.

Burton, J., Ijjaali, I., Barberan, O., Petitet, F., Vercauteren, D. P., & Michel, A. (2006). Recursive partitioning for the prediction of cytochromes P450 2D6 and 1A2 inhibition: importance of the quality of the dataset. *Journal of Medicinal Chemistry, 49,* 6231–6240.

Byvatov, E., Fechner, U., Sadowski, J., & Schneider, G. (2003). Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of Chemical Information and Computer Sciences, 43,* 1882–1889.

Cabrera-Perez, M. A., Bermejo, M., Alvarez, I. G., Alvarez, M. G., Garrigues, T. M., et al. (2012). QSPR in oral bioavailability: Specificity or integrality? *Mini Reviews in Medicinal Chemistry 12*, 534–550.

Caporuscio, F., & Tafi, A. (2011). Pharmacophore modelling: A forty year old approach and its modern synergies. *Current Medicinal Chemistry, 18,* 2543–2553.

Cartmell, J., Enoch, S., Krstajic, D., & Leahy, D. E. (2005). Automated QSPR through competitive workflow. *Journal of Computer-Aided Molecular Design, 19,* 821–833.

Castillo-Garit, J. A., Marrero-Ponce, Y., Torrens, F., & García-Domenech, R. (2008). Estimation of ADME properties in drug discovery: Predicting Caco-2 cell permeability using atom-based stochastic and non-stochastic linear indices. *Journal of Pharmaceutical Sciences, 97,* 1946–1976.

Cereto-Massagué, A., Guasch, L., Valls, C., Mulero, M., Pujadas, G., & Garcia-Vallvé, S. (2012). DecoyFinder: an easy-to-use python GUI application for building target-specific decoy sets. *Bioinformatics, 28,* 1661–1662.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., et al. (2014). QSAR modeling: where have you been? Where are you going to? *Journal of Medicinal Chemistry 57*, 4977–5010.

Chrysanthakopoulos, M., Koletsou, A., Nicolaou, I., Demopoulos, V. J., & Tsantili-Kakoulidou, A. (2009). Lipophilicity studies on pyrrolyl-acetic acid derivatives. Experimental versus predicted logP values in relationship with aldose reductase inhibitory activity. *QSAR & Combinatorial Science, 28,* 551–560.

Clark, D. E. (2003). In silico prediction of blood–brain barrier permeation. *Drug Discovery Today, 8,* 927–933.

Clark, D. E. (1999). Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *Journal of Pharmaceutical Sciences, 88,* 807–814.

Congreve, M., Carr, R., Murray, C., & Jhoti, H. (2003). A "rule of three" for fragment-based lead discovery? *Drug Discovery Today 8*, 876–877.

Consonni, V., Todeschini, R., & Pavan, M. (2002). Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *Journal of Chemical Information and Computer Sciences, 42,* 682–692.

Cox, R., Green, D. V., Luscombe, C. N., Malcolm, N., & Pickett, S. D. (2013). QSAR workbench: Automating QSAR modeling to drive compound design. *Journal of Computer-Aided Molecular Design, 27,* 321–336.

Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society, 110,* 5959–5967.

Cramer, R. D. (2012). The inevitable QSAR renaissance. *Journal of Computer-Aided Molecular Design, 26,* 35–38.

Crivori, P., Cruciani, G., Carrupt, P.-A., & Testa, B. (2000). Predicting blood-brain barrier permeation from three-dimensional molecular structure. *Journal of Medicinal Chemistry, 43,* 2204–2216.

Cruciani, G., Carosati, E., De Boeck, B., Ethirajulu, K., Mackie, C., Howe, T., et al. (2005). MetaSite: Understanding metabolism in human cytochromes from the perspective of the chemist. *Journal of Medicinal Chemistry, 48,* 6970–6979.

Cruciani, G., Crivori, P., Carrupt, P.-A., & Testa, B. (2000). Molecular fields in quantitative structure—Permeation relationships: The VolSurf approach. *Journal of Molecular Structure: THEOCHEM, 503,* 17–30.

Csizmadia, F., Tsantili-Kakoulidou, A., Panderi, I., & Darvas, F. (1997). Prediction of distribution coefficient from structure. 1. Estimation method. *Journal of Pharmaceutical Sciences, 86,* 865–871.

Darvas, F. (1988). Predicting metabolic pathways by logic programming. *Journal of Molecular Graphics, 6,* 80–86.

Dearden, J. C. (2007). In silico prediction of ADMET properties: How far have we come? *Expert Opinion Drug Metabolism Toxicology, 3,* 635–639.

De Benedetti, P. G., & Fanelli, F. (2010). Computational quantum chemistry and adaptive ligand modeling in mechanistic QSAR. *Drug Discovery Today, 15,* 859–866.

De Melo, E. B., Ferreira, M. M. C., et al. (2009). Nonequivalent effects of diverse LogP algorithms in three QSAR studies. *QSAR Comb Sci 28*, 1156–1165.

Di, L., Kerns, E. H., Bezar, I. F., Petusky, S. L., & Huang, Y. (2009). Comparison of blood–brain barrier permeability assays: in situ brain perfusion, MDR1-MDCKII and PAMPA-BBB. *Journal of Pharmaceutical Sciences, 98,* 1980–1991.

Dunn, W. J. (1988). QSAR approaches to predicting toxicity. *Toxicology Letters, 43,* 277–283.

Ecker, G. F., & Noe, C. R. (2004). In silico prediction models for blood–brain barrier permeation. *Current Medicinal Chemistry, 11,* 1617.

Ekins, S., Bravi, G., Binkley, S., Gillespie, J. S., Ring, B. J., Wikel, J. H., et al. (1999). Three-and four-dimensional quantitative structure activity relationship analyses of cytochrome P-450 3A4 inhibitors. *Journal of Pharmacology and Experimental Therapeutics, 290,* 429–438.

Enyedy, I. J., & Egan, W. J. (2008). Can we use docking and scoring for hit-to-lead optimization? *Journal of Computer-Aided Molecular Design, 22,* 161–168. doi:10.1007/s10822-007-9165-4.

Enyedy, I. J., Ling, Y., Nacro, K., Tomita, Y., Wu, X., Cao, Y., et al. (2001). Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening. *Journal of Medicinal Chemistry*, *44*, 4313–4324.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wold, S. (2001). *Multi-and megavariate data analysis: Principles and applications. Umetrics.*

Estrada, E., Uriarte, E., Molina, E., Simón-Manso, Y., & Milne, G. W. (2006). An integrated in silico analysis of drug-binding to human serum albumin. *Journal of Chemical Information and Modeling, 46,* 2709–2724.

Evans, W. E., & Guy, R. K. (2004). Gene expression as a drug discovery tool. *Nature Genetics, 36,* 214–215.

Faulon, J.-L., Brown, W. M., & Martin, S. (2005). Reverse engineering chemical structures from molecular descriptors: how many solutions? *Journal of Computer-Aided Molecular Design, 19,* 637–650.

Filikov, A. V., Mohan, V., Vickers, T. A., Griffey, R. H., Cook, P. D., Abagyan, R. A., et al. (2000). Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR. *Journal of Computer-Aided Molecular Design, 14*, 593–610.

Fujita, T., & Winkler, D. A. (2016). Understanding the roles of the "two QSARs". *Journal of Chemical Information and Modeling, 56,* 269–274.

Ganellin, C. R. (2004). Robin Ganellin gives his views on medicinal chemistry and drug discovery. *Drug Discovery Today, 9,* 158–160.

Gasteiger, J., et al. (2003). Handbook *of chemoinformatics*. Wiley Online Library.

Gaviraghi, G., Barnaby, R. J., & Pellegatti, M. (2001). Pharmacokinetic challenges in lead optimization. *Testa B Van Waterbeemd H folk. G 3–14*.

Gedeck, P., Rohde, B., & Bartels, C. (2006). QSAR-how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *Journal of Chemical Information and Modeling, 46,* 1924–1936.

Ghafourian, T., & Amin, Z. (2013). QSAR models for the prediction of plasma protein binding. *BioImpacts BI, 3,* 21.

Giaginis, C., Theocharis, S., & Tsantili-Kakoulidou, A. (2008). Quantitative Structure-activity relationships for PPAR-γ binding and gene transactivation of tyrosine-based agonists using multivariate statistics. *Chemical Biology & Drug Design, 72,* 257–264.

Giaginis, C., Theocharis, S., & Tsantili-Kakoulidou, A. (2007). A consideration of PPAR-γ ligands with respect to lipophilicity: Current trends and perspectives. *Expert Opinion on Investigational Drugs, 16,* 413–417.

Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2016). BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research, 44,* D1045–D1053.

Godden, J. W., & Bajorath, J. (2000). Shannon entropy—A novel concept in molecular descriptor and diversity analysis. *Journal of Molecular Graphics and Modelling, 18,* 73–76.

Goldmann, D., Montanari, F., Richter, L., Zdrazil, B., & Ecker, G. F. (2014). Exploiting open data: A new era in pharmacoinformatics. *Future Medicinal Chemistry, 6,* 503–514.

Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry, 28,* 849–857.

Gozalbes, R., Doucet, J. P., & Derouin, F. (2002). Application of Topological descriptors in QSAR and drug design: History and new trends. *Current Drug Targets-Infectious Disorders, 2,* 93–102. doi:10.2174/1568005024605909.

Gramatica, P. (2006). WHIM descriptors of shape. *QSAR & Comb. Sci., 25,* 301–415.

Guha, R. (2011). The ups and downs of structure-activity landscapes. *Chemoinformatics and Computational Chemical Biology*, 101–117.

Guha, R., & Jurs, P. C. (2005). Determining the validity of a QSAR model-a classification approach. *Journal of Chemical Information and Modeling, 45,* 65–73.

Hajduk, P. J., Mendoza, R., Petros, A. M., Huth, J. R., Bures, M., Fesik, S. W., et al. (2003). Ligand binding to domain-3 of human serum albumin: A chemometric analysis. *Journal of Computer-Aided Molecular Design, 17,* 93–102.

Hann, M. M. (2011). Molecular obesity, potency and other addictions in drug discovery. *MedChemComm, 2,* 349–355.

Hansch, C. (1969). Quantitative approach to biochemical structure-activity relationships. *Accounts of Chemical Research, 2,* 232–239.

Hansch, C., & Clayton, J. M. (1973). Lipophilic character and biological activity of drugs II: The parabolic case. *Journal of Pharmaceutical Sciences, 62,* 1–21.

Hansch, C., & Fujita, T. (1964). p-σ-π Analysis. A method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society, 86,* 1616–1626.

Hansch, C., Hoekman, D., Leo, A., Zhang, L., & Li, P. (1995a). The expanding role of quantitative structure-activity relationships (QSAR) in toxicology. *Toxicology Letters, 79,* 45–53.

Hansch, C., & Leo, A. (1979). Substituent constants for correlation analysis in chemistry and biology. Wiley.

Hansch, C., Leo, A., Hoekman, D. H., et al. (1995b). Exploring QSAR: Fundamentals and applications in chemistry and biology. Washington, DC: American Chemical Society.

Hansch, C., Muir, R. M., Fujita, T., Maloney, P. P., Geiger, F., & Streich, M. (1963). The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. *Journal of the American Chemical Society, 85,* 2817–2824.

Hanumegowda, U. M., Wenke, G., Regueiro-Ren, A., Yordanova, R., Corradi, J. P., & Adams, S. P. (2010). Phospholipidosis as a function of basicity, lipophilicity, and volume of distribution of compounds. *Chemical Research in Toxicology, 23,* 749–755.

Helgee, E. A., Carlsson, L., Boyer, S., & Norinder, U. (2010). Evaluation of quantitative structure-activity relationship modeling strategies: Local and global models. *Journal of Chemical Information and Modeling, 50,* 677–689.

Helguera A. M., CombesR. D., González M. P., & Cordeiro M. N. (2008). Applications of 2D descriptors in drug design: A DRAGON tale. *Current Topics in Medicinal Chemistry*, *8*, 1628–1655.

Hertzberg, R. P., & Pope, A. J. (2000). High-throughput screening: New technology for the 21st century. *Current Opinion in Chemical Biology, 4,* 445–451.

Hieronymus, H., Lamb, J., Ross, K. N., Peng, X.P., Clement, C., Rodina, A., et al. (2006). Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell 10*, 321–330.

Hillisch, A., Heinrich, N., & Wild, H. (2015). Computational chemistry in the pharmaceutical industry: From childhood to adolescence. *ChemMedChem, 10,* 1958–1962.

Hodos, R. A., Kidd, B. A., Shameer, K., Readhead, B. P., & Dudley, J. T. (2016). In silico methods for drug repurposing and pharmacology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 8,* 186–210.

Hollósy, F., Valkó, K., Hersey, A., Nunhuck, S., Kéri, G., & Bevan, C. (2006). Estimation of volume of distribution in humans from high throughput HPLC-based measurements of human serum albumin binding and immobilized artificial membrane partitioning. *Journal of Medicinal Chemistry, 49,* 6958–6971.

Hopfinger, A. J., Wang, S., Tokarski, J. S., Jin, B., Albuquerque, M., Madhav, P. J., et al. (1997). Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *Journal of the American Chemical Society, 119,* 10509–10524.

Hopkins, A. L., Keserü, G. M., Leeson, P. D., Rees, D. C., & Reynolds, C. H. (2014). The role of ligand efficiency metrics in drug discovery. *Nature Reviews Drug Discovery, 13,* 105–121.

Houston, J. G., & Banks, M. (1997). The chemical-biological interface: Developments in automated and miniaturised screening technology. *Current Opinion in Biotechnology, 8,* 734–740.

Hou, T., Wang, J., Zhang, W., & Xu, X. (2007). ADME evaluation in drug discovery. 6. Can oral bioavailability in humans be effectively predicted by simple molecular property-based rules? *Journal of Chemical Information and Modeling, 47,* 460–463.

Huang, S.-M., Abernethy, D. R., Wang, Y., Zhao, P., & Zineh, I. (2013). The utility of modeling and simulation in drug development and regulatory review. *Journal of Pharmaceutical Sciences, 102,* 2912–2923.

Hughes, J. D., Blagg, J., Price, D. A., Bailey, S., DeCrescenzo, G. A., Devraj, R. V., et al. (2008). Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorganic & Medicinal Chemistry Letters 18*, 4872–4875.

Irvine, J. D., Takahashi, L., Lockhart, K., Cheong, J., Tolan, J. W., Selick, H. E., et al. (1999). MDCK (Madin–Darby canine kidney) cells: A tool for membrane permeability screening. *Journal of Pharmaceutical Sciences, 88,* 28–33.

Irwin, J. J. (2008). Using ZINC to acquire a virtual screening library. *Current Protocols in Bioinformatics*, 14–6.

Ito, K., Iwatsubo, T., Kanamitsu, S., Nakajima, Y., & Sugiyama, Y. (1998). Quantitative prediction of in vivo drug clearance and drug interactions from in vitro data on metabolism, together with binding and transport. *Annual Review of Pharmacology and Toxicology, 38,* 461–499. doi:10.1146/annurev.pharmtox.38.1.461.

Jamei, M. (2016). Recent advances in development and application of physiologically-based pharmacokinetic (PBPK) models: A transition from academic curiosity to regulatory acceptance. *Current Pharmacology Reports, 2,* 161–169.

Jamei, M., Marciniak, S., Feng, K., Barnett, A., Tucker, G., & Rostami-Hodjegan, A. (2009). The Simcyp$^{®}$ population-based ADME simulator. *Expert Opinion on Drug Metabolism & Toxicology, 5,* 211–223.

Jaworska, J., Nikolova-Jeliazkova, N., & Aldenberg, T. (2005). QSAR applicability domain estimation by projection of the training set descriptor space: A review. *Atla-Nottingham- 33*, 445.

Jaworska, J. S., Comber, M., Auer, C., & Van Leeuwen, C. J. (2003). Summary of a workshop on regulatory acceptance of (Q) SARs for human health and environmental endpoints. *Environmental Health Perspectives, 111,* 1358.

Jorgensen, W. L. (2009). Efficient drug lead discovery and optimization. *Accounts of Chemical Research, 42,* 724–733.

Jorgensen, W. L. (2004). The many roles of computation in drug discovery. *Science, 303,* 1813–1818.

Kaldor, S. W., Kalish, V. J., Davies, J. F., Shetty, B.V., Fritz, J.E., Appelt, K., et al. (1997). Viracept (nelfinavir mesylate, AG1343): A potent, orally bioavailable inhibitor of HIV-1 protease. *Journal of Medicinal Chemistry, 40*, 3979–3985.

Kaliszan, R., & Markuszewski, M. (1996). Brain/blood distribution described by a combination of partition coefficient and molecular mass. *International Journal of Pharmaceutics, 145,* 9–16.

Kansy, M., Senner, F., & Gubernator, K. (1998). Physicochemical high throughput screening: Parallel artificial membrane permeation assay in the description of passive absorption processes. *Journal of Medicinal Chemistry, 41,* 1007–1010.

Kariv, I., Cao, H., & Oldenburg, K. R. (2001). Development of a high throughput equilibrium dialysis method. *Journal of Pharmaceutical Sciences, 90,* 580–587.

Katritzky, A. R., Lobanov, V. S., & Karelson, M. (1994). *CODESSA: reference manual*. FL: Univ. Fla. Gainesv.

Keserü, G. M. (2001). A virtual high throughput screen for high affinity cytochrome P450cam substrates. Implications for in silico prediction of drug metabolism. *Journal of Computer-Aided Molecular Design, 15,* 649–657.

Kier, L. B., & Hall, L. H. (1999). *Molecular structure description: The Electrotopological State*. San Diego, CA: Academic Press.

Kim, M. T., Sedykh, A., Chakravarti, S. K., Saiakhov, R. D., & Zhu, H. (2014). Critical evaluation of human oral bioavailability for pharmaceutical drugs by using various cheminformatics approaches. *Pharmaceutical Research, 31,* 1002–1014.

Kirchmair, J., Göller, A. H., Lang, D., Kunze, J., Testa, B., Wilson, I. D., et al. (2015). Predicting drug metabolism: experiment and/or computation? *Nature Reviews Drug Discovery, 14,* 387–404.

Klebe, G. (1998). Comparative molecular similarity indices analysis: CoMSIA. *Perspectives in Drug Discovery and Design, 12*, 87–104.

Klebe, G., Abraham, U., & Mietzner, T. (1994). Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *Journal of Medicinal Chemistry, 37,* 4130–4146.

Klopman, G., & Kalos, A. N. (1985). Causality in structure-activity studies. *Journal of Computational Chemistry, 6,* 492–506.

Klopman, G., Stefan, L. R., & Saiakhov, R. D. (2002). ADME evaluation: 2. A computer model for the prediction of intestinal absorption in humans. *European Journal of Pharmaceutical Sciences, 17,* 253–263.

Klopman, G., Tu, M., & Fan, B. T. (1999). META 4. Prediction of the metabolism of polycyclic aromatic hydrocarbons. *Theoretical Chemistry Accounts, 102,* 33–38.

Klopman, G., Tu, M., & Talafous, J. (1997). META. 3. A genetic algorithm for metabolic transform priorities optimization. *Journal of Chemical Information and Computer Sciences, 37,* 329–334.

Kola, I., & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery, 3,* 711–715. doi:10.1038/nrd1470.

Konovalov, D. A., Coomans, D., Deconinck, E., & Vander Heyden, Y. (2007). Benchmarking of QSAR models for blood-brain barrier permeation. *Journal of Chemical Information and Modeling, 47,* 1648–1656.

Koukoulitsa, C., Tsantili-Kakoulidou, A., Mavromoustakos, Th., & Chinou, I. (2009). PLS analysis for antibacterial Activity of natural coumarins using Volsurf descriptors. *QSAR and Comb. Sci., 28*, 785–789.

Kubinyi, H. (1979). Lipophilicity and drug activity, in: Progress in Drug Research/Fortschritte Der Arzneimittelforschung/Progrès Des Recherches Pharmaceutiques, pp. 97–198. Springer.

Kubinyi, H., & Kehrhahn, O. H. (1978). Quantitative structure-activity relationships. VI. Non-linear dependence of biological activity on hydrophobic character: Calculation procedures for bilinear model. *Arzneimittel-Forschung, 28,* 598–601.

Kubinyi, H., Mannhold, R., Krogsgaard, L. R., & Timmerman, H. E., (1993). In R. Mannhold, Al (Eds.), *Methods and principles in medicinal chemistry*.

Kumar, R., Sharma, A., & Varadwaj, P. K. (2011). A prediction model for oral bioavailability of drugs using physicochemical properties by support vector machine. *Journal of Natural Science, Biology, and Medicine, 2,* 168.

Lambrinidis, G., Vallianatou, T., & Tsantili-Kakoulidou, A. (2015). In vitro, in silico and integrated strategies for the estimation of plasma protein binding. A review. *Advanced Drug Delivery Reviews, 86,* 27–45.

Larregieu, C. A., & Benet, L. Z. (2014). Distinguishing between the permeability relationships with absorption and metabolism to improve BCS and BDDCS predictions in early drug discovery. *Molecular Pharmaceutics, 11,* 1335–1344.

Larregieu, C. A., & Benet, L. Z. (2013). Drug discovery and regulatory considerations for improving in silico and in vitro predictions that use Caco-2 as a surrogate for human intestinal permeability measurements. *American Association of Pharmaceutical Scientists Journal, 15,* 483–497.

Leeson, P. D., & Springthorpe, B. (2007). The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews Drug Discovery, 6,* 881–890.

Leo, A., Hansch, C., & Elkins, D. (1971). Partition coefficients and their uses. *Chemical Reviews, 71,* 525–616. doi:10.1021/cr60274a001.

Levin, V. A. (1980). Relationship of octanol/water partition coefficient and molecular weight to rat brain capillary permeability. *Journal of Medicinal Chemistry, 23,* 682–684.

Lewis, D. F., Ioannides, C., & Parke, D. V. (1996). COMPACT and molecular structure in toxicity assessment: A prospective evaluation of 30 chemicals currently being tested for rodent carcinogenicity by the NCI/NTP. *Environmental Health Perspectives, 104,* 1011.

Lewis, D. F. V. (2001). COMPACT: A structural approach to the modelling of cytochromes P450 and their interactions with xenobiotics. *Journal of Chemical Technology and Biotechnology, 76,* 237–244.

Lima, A. N., Philot, E. A., Trossini, G. H. G., Scott, L. P. B., Maltarollo, V. G., & Honorio, K. M. (2016). Use of machine learning approaches for novel drug discovery. *Expert Opinion on Drug Discovery, 11,* 225–239.

Lind, K. E., Du, Z., Fujinaga, K., Peterlin, B. M., & James, T. L. (2002). Structure-based computational database screening, in vitro assay, and NMR assessment of compounds that target TAR RNA. *Chemistry & Biology, 9*, 185–193.

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). In Vitro Models for Selection of Development CandidatesExperimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews, 23,* 3–25. doi:10.1016/S0169-409X(96)00423-1.

Löfås, S., & Johnsson, B. (1990). A novel hydrogel matrix on gold surfaces in surface plasmon resonance sensors for fast and efficient covalent immobilization of ligands. *Journal of the Chemical Society, Chemical Communications,* 1526–1528.

Luco, J. M. (1999). Prediction of the brain-blood distribution of a large set of drugs from structurally derived descriptors using partial least-squares (PLS) modeling. *Journal of Chemical Information and Computer Sciences, 39,* 396–404.

MACCS. (2011). MACCS structural keys. San Diego, CA: Accelrys.

Mannhold, R., & Dross, K. (1996). Calculation procedures for molecular lipophilicity: A comparative study. *Quantitative Structure-Activity Relationships, 15,* 403–409.

Mannhold, R., Poda, G. I., Ostermann, C., & Tetko, I. V. (2009). Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *Journal of Pharmaceutical Sciences, 98,* 861–893.

Martin, Y. C. (2005). A bioavailability score. *Journal of Medicinal Chemistry 48*, 3164–3170.

Martin, Y. C. (1978). *Quantitative Drug Design*. New York: A Critical Introduction. Marcel Dekker.

Meanwell, N. A. (2016). Improving Drug Design: An Update on Recent Applications of Efficiency Metrics, Strategies for Replacing Problematic Elements, and Compounds in Nontraditional Drug Space. *Chemical Research in Toxicology, 29,* 564–616.

Mekenyan, O., & Bonchev, D. (1986). Oasis method for predicting biological-activity of chemical-compounds. *Acta Pharmaceutica Jugoslavica, 36,* 225–237.

Mitchell, M., (1998). An introduction to genetic algorithms. MIT press.

Moda, T. L., Montanari, C. A., & Andricopulo, A. D. (2007). Hologram QSAR model for the prediction of human oral bioavailability. *Bioorganic & Medicinal Chemistry, 15,* 7738–7745.

MOE. (2016). Molecular operating environment (MOE), 2013.08; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7.

Moura Barbosa, A. J., & Del Rio, A. (2012). Freely accessible databases of commercial compounds for high-throughput virtual screenings. *Current Topics in Medicinal Chemistry, 12,* 866–877.

Mostrag-Szlichtyng, A., & Worth, A. (2010). *Review of QSAR models and software tools for predicting biokinetic properties*. Comm: Luxemb. Eur.

Muir, R. M., Fujita, T., & Hansch, C. (1967). Structure-activity relationship in the auxin activity of mono-substituted phenylacetic acids. *Plant Physiology, 42,* 1519–1526.

Mysinger, M. M., Carchia, M., Irwin, J. J., & Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry, 55,* 6582–6594. doi:10.1021/jm300687e.

Narayanan, R., & Gunturi, S. B. (2005). In silico ADME modelling: Prediction models for blood–brain barrier permeation using a systematic variable selection method. *Bioorganic & Medicinal Chemistry, 13,* 3017–3028.

Navratilova, I., Myszka, D. G., & Rich, R. L. (2007). Probing membrane protein interactions with real-time biosensor technology. *Biophysical Analysis of Membrane Proteins: Investigating Structure and Function,* 121–140.

Netzeva, T. I., Worth, A. P., Aldenberg, T., Benigni, R., Cronin, M. T., Gramatica, P., et al. (2005). Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. *ATLA 33*, 155–173.

Nicolotti, O., Gillet, V. J., Fleming, P. J., & Green, D. V. (2002). Multiobjective optimization in quantitative structure-activity relationships: Deriving accurate and interpretable QSARs. *Journal of Medicinal Chemistry, 45,* 5069–5080.

Oprea, T. I. (2000). Property distribution of drug-related chemical databases. *Journal of Computer-Aided Molecular Design, 14,* 251–264.

Owens, P. K., Raddad, E., Miller, J. W., Stille, J. R., Olovich, K. G., Smith, N. V., et al. (2015). A decade of innovation in pharmaceutical R&D: The chorus model. *Nature Reviews Drug Discovery, 14,* 17–28.

Pajouhesh, H., & Lenz, G. R. (2005). Medicinal chemical properties of successful central nervous system drugs. *NeuroRx, 2,* 541–553.

Papadatos, G., Gaulton, A., Hersey, A., & Overington, J. P. (2015). Activity, assay and target data curation and quality in the ChEMBL database. *Journal of Computer-Aided Molecular Design, 29,* 885–896.

Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, et al. (2010). ArrayExpress update—An archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Research,* gkq1040.

Paterlini, S., & Minerva, T. (2010). Regression model selection using genetic algorithms. In *Proceedings of the 11th WSEAS International Conference on Nural Networks and 11th WSEAS International Conference on Evolutionary Computing and 11th WSEAS International Conference on Fuzzy Systems,* pp. 19–27.

Pastor, M., Cruciani, G., Mclay, I., Pickett, S., & Clementi, S. (2000). Grid-independent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *Journal of Medicinal Chemistry, 43,* 3233–3243.

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., et al. (2010). How to improve R&D productivity: The pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery, 9,* 203–214.

Pham-The, H., González-Álvarez, I., Bermejo, M., Garrigues, T., Le-Thi-Thu, H., & Cabrera-Pérez, M. Á. (2013). The use of rule-based and QSPR approaches in ADME profiling: A case study on caco-2 permeability. *Molecular Informatics, 32,* 459–479.

Platts, J. A., Abraham, M. H., Zhao, Y. H., Hersey, A., Ijaz, L., & Butina, D. (2001). Correlation and prediction of a large blood–brain distribution data set-an LFER study. *European Journal of Medicinal Chemistry, 36,* 719–730.

Pliška, V., Testa, B., & van de Waterbeemd, H. (1996). Lipophilicity: The empirical tool and the fundamental objective. an introduction. In V. Pliška, B. Testa, P. -D. H. van de Waterbeemd (Eds.), *Lipophilicity in drug action and toxicology* (pp. 1–6). Wiley-VCH Verlag GmbH.

Polanski, J. (2009). Receptor dependent multidimensional QSAR for modeling drug-receptor interactions. *Current Medicinal Chemistry, 16,* 3243–3257.

Puzyn, T., Leszczynski, J., & Cronin, M. T. (2010). *Recent advances in QSAR studies: Methods and applications*. Springer Science & Business Media.

Qiu, T., Qiu, J., Feng, J., Wu, D., Yang, Y., Tang, K., et al. (2016). The recent progress in proteochemometric modelling: Focusing on target descriptors, cross-term descriptors and application scope. *Briefings in Bioinformatics.* bbw004.

Rekker, R. F., & Mannhold, R. (1992). *Calculation of drug lipophilicity: The hydrophobic fragmental constant approach.* Wiley-VCH.

Rich, R. L., & Myszka, D. G. (2000). Advances in surface plasmon resonance biosensor analysis. *Current Opinion in Biotechnology, 11,* 54–61.

Rodgers, S. L., Davis, A. M., Tomkinson, N. P., & van de Waterbeemd, H. (2011). Predictivity of simulated ADME AutoQSAR models over time. *Molecular Informatics, 30,* 256–266.

Rodgers, S. L., Davis, A. M., & van de Waterbeemd, H. (2007). Time-series QSAR analysis of human plasma protein binding data. *QSAR & Combinatorial Science, 26,* 511–521.

Rogge, M. C., & Taft, D. R. (Eds.). (2010) preclinical drug development second edition. In *Drugs and the pharmaceutical sciences* (Vol.187). CRS Press, Taylor and Francis Group.

Rowley, M., Kulagowski, J. J., Watt, A. P., Rathbone, D., Stevenson, G. I., Carling, R. W., et al. (1997). Effect of plasma protein binding on in vivo activity and brain penetration of glycine/NMDA receptor antagonists. *Journal of Medicinal Chemistry, 40,* 4053–4068. doi:10.1021/jm970417o.

Roy, K., Mitra, I., Kar, S., Ojha, P. K., Das, R. N., & Kabir, H. (2012). Comparative studies on some metrics for external validation of QSPR models. *Journal of Chemical Information and Modeling, 52,* 396–408. doi:10.1021/ci200520g.

Roy, P. P., Paul, S., Mitra, I., & Roy, K. (2009). On two novel parameters for validation of predictive QSAR models. *Molecules, 14,* 1660–1701.

Rücker, C., Rücker, G., & Meringer, M. (2007). y-Randomization and its variants in QSPR/QSAR. *Journal of Chemical Information and Modeling, 47,* 2345–2357.

Rutenber, E. E., & Stroud, R. M. (1996). Binding of the anticancer drug ZD1694 to E. coli thymidylate synthase: Assessing specificity and affinity. *Structure, 4,* 1317–1324.

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules, 17,* 4791–4810.

Saiakhov, R. D., Stefan, L. R., & Klopman, G. (2000). Multiple computer-automated structure evaluation model of the plasma protein binding affinity of diverse drugs. *Perspectives in Drug Discovery and Design, 19,* 133–155.

Sakiyama, Y. (2009). The use of machine learning and nonlinear statistical tools for ADME prediction. *Expert Opinion on Drug Metabolism & Toxicology, 5,* 149–169.

Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., & Sherman, W. (2013). Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design, 27,* 221–234.

Satyanarayanajois, S.D. (2011). *Drug design and discovery: Methods and protocols*. Humana Press.

Schindler, T., Bornmann, W., Pellicena, P., Miller, W. T., Clarkson, B., & Kuriyan, J. (2000). Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. *Science, 289,* 1938–1942.

Schneider, G. (2010). Virtual screening: An endless staircase? *Nature Reviews Drug Discovery, 9,* 273–276.

Sedykh, A., Zhu, H., Tang, H., Zhang, L., Richard, A., Rusyn, I., et al. (2011). Use of in vitro HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity. *Environmental Health Perspectives, 119,* 364.

Sherer, E. C., Verras, A., Madeira, M., Hagmann, W. K., Sheridan, R. P., Roberts, D., et al. (2012). QSAR Prediction of passive permeability in the LLC-PK1 cell line: Trends in molecular properties and cross-prediction of caco-2 permeabilities. *Molecular Informatics, 31,* 231–245.

Sheridan, R. P. (2014). Global quantitative structure-activity relationship models vs selected local models as predictors of off-target activities for project compounds. *Journal of Chemical Information and Modeling, 54,* 1083–1092.

Sheridan, R. P., Korzekwa, K. R., Torres, R. A., & Walker, M. J. (2007). Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9. *Journal of Medicinal Chemistry, 50,* 3173–3184.

Sheridan, R. P., McMasters, D. R., Voigt, J. H., & Wildey, M. J. (2015). eCounterscreening: Using QSAR predictions to prioritize testing for off-target activities and setting the balance between benefit and risk. *Journal of Chemical Information and Modeling, 55,* 231–238.

Sohn, Y. S., Park, C., Lee, Y., Kim, S., Thangapandian, S., Kim, Y., et al. (2013). Multi-conformation dynamic pharmacophore modeling of the peroxisome prolifera-tor-activated receptor γ for the discovery of novel agonists. *Journal of Molecular Graphics and Modelling, 46,* 1–9.

Speck-Planche, A., & Cordeiro, M. N. D. S. (2015). Multitasking models for quantitative structure—Biological effect relationships: Current status and future perspectives to speed up drug discovery. *Expert Opinion on Drug Discovery, 10,* 245–256.

Spowage, B. M., Bruce, C. L., & Hirst, J. D. (2009). Interpretable correlation descriptors for quantitative structure—Activity relationships. *Journal of Cheminformatics, 1,* 22.

Stegmaier, K., Ross, K. N., Colavito, S. A., O'Malley, S., Stockwell, B. R., & Golub, T. R. (2004). Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nature Genetics, 36,* 257–263.

Stuper, A. J., & Jurs, P. C. (1976). ADAPT: A computer system for automated data analysis using pattern recognition techniques. *Journal of Chemical Information and Modeling, 16,* 99–105. doi:10.1021/ci60006a014.

Suenderhauf, C., Hammann, F., & Huwyler, J. (2012). Computational prediction of blood–brain barrier permeability using decision tree induction. *Molecules, 17,* 10429–10445.

Sun, H. (2004). A universal molecular descriptor system for prediction of logP, logS, logBB, and absorption. *Journal of Chemical Information and Computer Sciences, 44,* 748–757.

Swift, R. V., & Amaro, R. E. (2013). Back to the future: Can physical models of passive membrane permeability help reduce drug candidate attrition and move us beyond QSPR? *Chemical Biology & Drug Design, 81,* 61–71.

Talafous, J., Sayre, L. M., Mieyal, J. J., & Klopman, G. (1994). META. 2. A dictionary model of mammalian xenobiotic metabolism. *Journal of Chemical Information and Computer Sciences, 34,* 1326–1333.

Tao, L., Zhang, P., Qin, C., Chen, S. Y., Zhang, C., Chen, Z., et al. (2015). Recent progresses in the exploration of machine learning methods as in-silico ADME prediction tools. *Advanced Drug Delivery Reviews, 86,* 83–100.

Tarcsay, Á., Nyíri, K., & Keserű, G. M. (2012). Impact of lipophilic efficiency on compound quality. *Journal of Medicinal Chemistry, 55,* 1252–1260.

Terfloth, L., Bienfait, B., & Gasteiger, J. (2007). Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *Journal of Chemical Information and Modeling, 47,* 1688–1701.

Testa, B., Balmat, A.-L., & Long, A. (2004). Predicting drug metabolism: Concepts and challenges. *Pure and Applied Chemistry, 76,* 907–914.

Testa, B., Balmat, A.-L., Long, A., & Judson, P. (2005a). Predicting drug metabolism—An evaluation of the expert system METEOR. *Chemistry & Biodiversity, 2,* 872–885.

Testa, B., Vistoli, G., & Pedretti, A. (2005b). Musings on ADME predictions and structure-activity relations. *Chemistry & Biodiversity, 2,* 1411–1427. doi:10.1002/cbdv.200590115.

Testa, B. (2009). Drug metabolism for the perplexed medicinal chemist. *Chemistry & Biodiversity, 6,* 2055–2070.

Tetko, I. V., Poda, G. I., Ostermann, C., & Mannhold, R. (2009). Large-scale evaluation of log P predictors: Local corrections may compensate insufficient accuracy and need of experimentally testing every other compound. *Chemistry & Biodiversity, 6,* 1837–1844.

Tetko, I. V., Tanchuk, V. Y., & Villa, A. E. (2001). Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *Journal of Chemical Information and Computer Sciences, 41,* 1407–1421.

Thiel-Demby, V. E., Humphreys, J. E., St. John Williams, L. A., Ellens, H. M., Shah, N., Ayrton, et al. (2008). Biopharmaceutics classification system: Validation and learnings of an in vitro permeability assay. *Molecular Pharmaceutics 6*, 11–18.

Tian, S., Li, Y., Wang, J., Zhang, J., & Hou, T. (2011). ADME evaluation in drug discovery. 9. Prediction of oral bioavailability in humans based on molecular properties and structural fingerprints. *Molecular Pharmaceutics, 8,* 841–851.

Tilley, J. W., Chen, L., Fry, D. C., Emerson, S.D., Powers, G.D., Biondi, D., et al. (1997). Identification of a small molecule inhibitor of the IL-2/IL-2Rα receptor interaction which binds to IL-2. *Journal of the American Chemical Society 119*, 7589–7590.

Todeschini, R., & Consonni, V. (2009). *Molecular descriptors for chemoinformatics*, Volume 41 (2 Volume Set). Wiley.

Tropsha, A., Gramatica, P., & Gombar, V. K. (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science, 22,* 69–77. doi:10.1002/qsar.200390007.

Tsantili-Kakoulidou, A., & Agrafiotis, D. K. (2011). The 18th european symposium on quantitative structure-activity relationships. *Expert Opinion on Drug Discovery, 6,* 453–456.

Tsopelas, F., Vallianatou, T., & Tsantili-Kakoulidou, A. (2016a). The potential of immobilized artificial membrane chromatography to predict human oral absorption. *European Journal of Pharmaceutical Sciences, 81,* 82–93.

Tsopelas, F., Vallianatou, T., & Tsantili-Kakoulidou, A. (2016b). Advances in immobilized artificial membrane (IAM) chromatography for novel drug discovery. *Expert Opinion on Drug Discovery, 11,* 473–488.

Ursu, O., Rayan, A., Goldblum, A., & Oprea, T. I. (2011). Understanding drug-likeness. *Wiley Interdisciplinary Reviews: Computational Molecular Science, 1,* 760–781.

Usansky, H. H., & Sinko, P. J. (2005). Estimating human drug oral absorption kinetics from Caco-2 permeability using an absorption-disposition model: Model development and evaluation and derivation of analytical solutions for ka and Fa. *Journal of Pharmacology and Experimental Therapeutics, 314,* 391–399.

Vallianatou, T., Lambrinidis, G., Giaginis, C., Mikros, E., & Tsantili-Kakoulidou, A. (2013). Analysis of PPAR-α/γ Activity by Combining 2-D QSAR and Molecular Simulation. *Molecular Informatics, 32,* 431–445.

van de Waterbeemd, H., Camenisch, G., Folkers, G., & Raevsky, O. A. (1996). Estimation of Caco-2 cell permeability using calculated molecular descriptors. *Quantitative Structure-Activity Relationships, 15,* 480–490.

van de Waterbeemd, H., & Smith, D. A., (2001). Relations of molecular properties with drug disposition: The cases of gastrointestinal absorption. *Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies,* 51.

Van de Waterbeemd, H., & Testa, B. (1987). The parametrization of lipophilicity and other structural properties in drug design. *Advances in Drug Research, 16,* 85–225.

Varghese, J. N. (1999). Development of neuraminidase inhibitors as anti-influenza virus drugs. *Drug Development Research, 46,* 176–196.

Vastag, M., & Keseru, G. M. (2009). Current in vitro and in silico models of blood–brain barrier penetration: a practical view. *Current Opinion in Drug Discovery & Development, 12,* 115–124.

Vasudevan, S. R., & Churchill, G. C. (2009). Mining free compound databases to identify candidates selected by virtual screening. *Expert Opinion on Drug Discovery, 4,* 901–906.

Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., & Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry, 45,* 2615–2623.

Vedani, A., Briem, H., Dobler, M., Dollinger, H., & McMasters, D. R. (2000). Multiple-conformation and protonation-state representation in 4D-QSAR: the neurokinin-1 receptor system. *Journal of Medicinal Chemistry, 43,* 4416–4427.

Vedani, A., Dobler, M., & Lill, M. A. (2005). Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. *Journal of Medicinal Chemistry, 48,* 3700–3703.

Veerasamy, R., Rajak, H., Jain, A., Sivadasan, S., Varghese, C. P., & Agrawal, R. K. (2011). Validation of QSAR models-strategies and importance. *International Journal of Drug Design & Discovery, 3,* 511–519.

Vilar, S., Chakrabarti, M., & Costanzi, S. (2010). Prediction of passive blood–brain partitioning: straightforward and effective classification models based on in silico derived physicochemical descriptors. *Journal of Molecular Graphics and Modelling, 28,* 899–903.

Volpe, D. A. (2008). Variability in Caco-2 and MDCK cell-based intestinal permeability assays. *Journal of Pharmaceutical Sciences, 97,* 712–725.

Votano, J. R., Parham, M., Hall, L. M., Hall, L. H., Kier, L. B., Oloff, S., et al. (2006). QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. *Journal of Medicinal Chemistry, 49,* 7169–7181. doi:10.1021/jm051245v.

Wager, T. T., Villalobos, A., Verhoest, P. R., Hou, X., & Shaffer, C. L. (2011). Strategies to optimize the brain availability of central nervous system drug candidates. *Expert Opinion on Drug Discovery, 6,* 371–381.

Wang, N.-N., Dong, J., Deng, Y.-H., Zhu, M.-F., Wen, M., Yao, Z.-J., et al. (2016). ADME properties evaluation in drug discovery: Prediction of caco-2 cell permeability using a combination of NSGA-II and boosting. *Journal of Chemical Information and Modeling, 56,* 763–773.

Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., Zhou, Z., et al. (2012). PubChem's BioAssay database. *Nucleic Acids Research 40*, D400–D412.

Wessel, M. D., Jurs, P. C., Tolan, J. W., & Muskal, S. M. (1998). Prediction of human intestinal absorption of drug compounds from molecular structure. *Journal of Chemical Information and Computer Sciences, 38,* 726–735.

Wiesmann, C., Christinger, H. W., Cochran, A. G., Cunningham, B. C., Fairbrother, W. J., Keenan, C. J., et al. (1998). Crystal structure of the complex between VEGF and a receptor-blocking peptide. *Biochemistry (Mosc), 37,* 17765–17772.

Willett, P. (2004). Evaluation of molecular similarity and molecular diversity methods using biological activity data in methods in molecular biology. In J. Bajorath (Ed.), *Chemoinformatics: Concepts, methods, and tools for drug discovery* (Vol. 275). Totowa, N.J: Humana Press Inc.

Williams, G. (2012). A searchable cross-platform gene expression database reveals connections between drug treatments and disease. *BMC Genomics, 13,* 1.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58,* 109–130.

Wong, W. W., & Burkowski, F. J. (2009). A constructive approach for discovering new drug leads: Using a kernel methodology for the inverse-QSAR problem. *Journal of Cheminformatics, 1,* 1.

Worth, A. P., Hartung, T., & Van Leeuwen, C. J. (2004). The role of the European centre for the validation of alternative methods (ECVAM) in the validation of (Q) SARs. *SAR and QSAR in Environmental Research, 15,* 345–358.

Yee, S. (1997). In vitro permeability across caco-2 cells (colonic) can predict in vivo (small intestinal) absorption in man—Fact or myth. *Pharmaceutical Research, 14,* 763–766.

Yera, E. R., Cleves, A. E., & Jain, A. N. (2014). Prediction of off-target drug effects through data fusion. In *Pacific Symposium on Biocomputing*. NIH Public Access, p. 160.

Yusof I., & Segall, M. D. (2013). Considering the impact drug-like properties have on the chance of success. *Drug Discovery Today 18*, 659–66.

Zhao, P., Rowland, M., & Huang, S.-M. (2012). Best practice in the use of physiologically based pharmacokinetic modeling and simulation to address clinical pharmacology regulatory questions. *Clinical Pharmacology and Therapeutics, 92,* 17–20.

Zhao, Y. H., Le, J., Abraham, M. H., Hersey, A., Eddershaw, P. J., Luscombe, C. N., Boutina, D., et al. (2001). Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. *Journal of Pharmaceutical Sciences 90*, 749–784.

Zsila, F. (2013). Subdomain IB is the third major drug binding region of human serum albumin: Toward the three-sites model. *Molecular Pharmaceutics, 10,* 1668–1682.