

Ligand-based drug designing

Suchitra M. Ajjarapu^{1,2}, Apoorv Tiwari^{1,3}, Pramod Wasudeo Ramteke^{3,4,5}, Dev Bukhsh Singh⁶ and Sundip Kumar¹

¹Bioinformatics Sub-DIC, Molecular Biology & Genetic Engineering, College of Basic Science and Humanities, G.B. Pant University of Agriculture and Technology, Pantnagar, India, ²Department of Biotechnology, Andhra University, India, ³Department of Computational Biology and Bioinformatics, Jacob Institute of Biotechnology and Bio-Engineering, Sam Higginbottom University of Agriculture, Technology and Sciences, India, ⁴Department of Molecular Biology & Genetic Engineering, RTM Nagpur University, India, ⁵Department of Life Sciences, Mandsaur University, India, ⁶Department of Biotechnology, Institute of Biosciences and Biotechnology, Chhatrapati Shahu Ji Maharaj University, Kanpur, India

15.1 Introduction

The present pharmaceutical and biomedical researches have faced a tougher challenge, and recent technological advances have revolutionized the field. The high-throughput screening (HTS) has been strengthening with emerging strategies of genomics, proteomics, chemical biology and chemical modeling, and molecular modeling. The indirect drug design or ligand-based drug designing (LBDD) helps the recognition of different new small compounds (ligands) that bind with the target structure of the protein. The in silico procedures engaged here are found to be convenient in recognizing the drug molecules supported by bioinformatics tools and examined further on different platforms for predicting potential active sites, producing similar structures, performing molecular docking with active ligands, analyzing binding interactions, and optimizing the lead ones to increase its binding properties, efficacy, and safety (Singh & Dwivedi, 2016). Therefore the current topics structure-based drug designing (SBDD) and LBDD have been studied for drug development as well as to distinguish for its drug-likeness properties. This point of view has gained ever popularity because of its novel algorithm developments for software tools and resources applied in drug discovery (Ferreira, Dos Santos, Oliva, & Andricopulo, 2015; Singh & Pathak, 2020). In the SBDD, three-dimensional (3D) structures are prepared by using three different approaches: (1) sketching macromolecular structures in the fragment libraries using 3D graphics interface of 3D structure; (2) distance geometry, quantum, or molecular mechanics by applying mathematical techniques; and (3) building databases of 3D structures using preprogrammed methods.

The absence of reliable 3D macromolecular structure has led a path for target recognition of compounds in a very systematic way by using distinct advanced computational tools. Fig. 15.1 elucidates the representation of basic methods

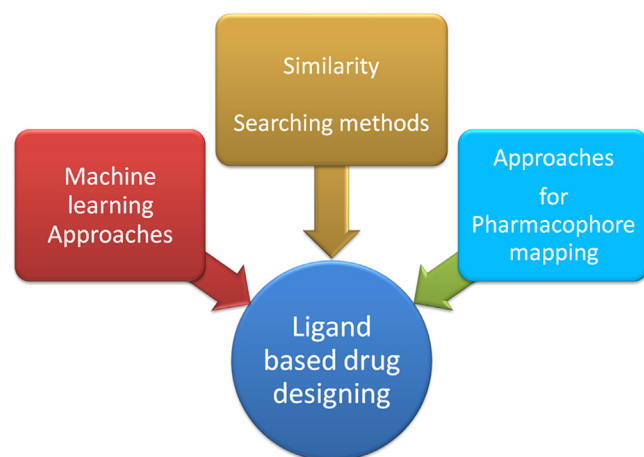


FIGURE 15.1 Basic methods involved in ligand-based drug designing.

and approaches involved in LBDD. Quantitative structure–activity relationship (QSAR) models, Free-Wilson models, and matched molecular pair (MMP) analysis are based on experimental data sets and are used to predict the potency of a lead compound along with absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties (Singh, 2020). Conformational analysis, quantum mechanics, geometry, and optimization play important role in determining the flexibility of compounds and their behavior for the bioactive conformation. Fragment-based replacement and scaffold hopping tools are very effective in addressing the metabolic instability of compounds as well as also guide the use of unused features in developing the models. Pharmacophoric techniques and electrostatic similarity alignment methods based on the criteria of ligand shape are used to compare different sets of compound data through linking their structural activity relationship (SAR) for virtual screening (VS) (Sliwoski, Kothiwale, Meiler, & Lowe, 2013).

15.2 Pharmacophore

Pharmacophore modeling is an approach useful for lead optimization and particularly used in medicinal and computational chemistry. The pharmacophore model is derived from a set of molecules using structural alignment and used for prediction of activity and 3D database searching. A given hypothesis with known/unknown activity data is used to develop 3D QSAR to identify overall aspects of the molecular structure and its activity toward the target. However, the pharmacophore model has wide applications in ADMET profile determining, estimation of side effect, off-target prediction, and target prediction. Consequently, pharmacophore knowledge with molecular docking simulations is used to improve VS. Table 15.1 describes the different tools involved in developing pharmacophore models.

15.2.1 Build pharmacophore hypothesis

This hypothesis laid on the concept of molecular recognition by a set of compounds elucidated by a common feature that interacts with the complementary site on the biological targets (Wermuth, 2006). The 3D structures play a crucial role in pharmacophore modeling. Each hypothesis is built by the superimposition of active ligands and their spatial arrangement and is proposed to understand the key interactions involved in ligand binding. A high scoring hypothesis is well supported by new potentially active molecules or used in the alignment of ligands in creating 3D QSAR models. The pharmacophore model explicates the steric and electronic points for optimal interaction with the drug target. The pharmacophore model depends on the ligand type, its size, and structural diversity.

TABLE 15.1 Software tools used for pharmacophore modeling and other applications.

S. no.	Tool name	Description	Link
1.	Pharmer	Efficient Pharmacophore tool for virtual screening	http://zincpharmer.csb.pitt.edu/
2.	Pharmapper	It uses the knowledge of 7000 targets based pharmacophore models. It searches for the best matches for an input ligand against pharmacophore-based models.	http://lilab-ecust.cn/pharmapper/index.html
3.	PharmaGist	It searches pharmacophore from a set of ligand molecules without prior knowledge of the target	http://bioinfo3d.cs.tau.ac.il/PharmaGist
4.	Pharmit	It is a tool used for pharmacokinetic drug monitoring	http://pharmit.csb.pitt.edu/
5.	Discovery studio	It is a therapeutic drug monitoring tool	https://discover.3ds.com/discovery-studio-visualizer-download
6.	Ligandscout	It allows creating 3D pharmacophore models from structural data or test/training set of molecules	https://LigandScout-for-Linux/3000-6677_4-75305808.html
7.	ICM-Chemist	A standalone suit with a list of programs for editing and chemical drawing, clustering, and enumeration	http://www.molsoft.com/icm-chemist.html
8.	Phase	A tool for ligand and structure-based drug design	https://www.schrodinger.com/Phase/

15.2.2 Alignment of molecules

15.2.2.1 Superimposition based on atom overlapping

Superimposition is performed by an atom-atom pairing between the molecules used for alignment. It is widely used in the pharmacophore model and focuses on the alignment of the atom. Superimposition can find the extent of dissimilarity between molecules based on the alignment, but cannot be applied to the molecules with diverse structural types. The binding sites are based on the superimposing of molecules, which can be elaborated as the molecular alignment, and therefore obtained by projecting the molecule to the active sites or residues in the receptor molecule, which interact with the compound (ligands). This is trusted to be more conceivable, in spite of issues in the conformational analysis due to concerned atoms or molecules increased degree of freedom. Fields/pseudo field-based method performs superimposition by comparing the similarities with respect to interaction energy fields for corresponding molecules. Electrostatic similarity along with the surface similarity of the molecule has been used for molecular alignment (Cleves, Johnson, & Jain, 2019). Target template is used as the base platform for the superimposition of pharmacophore. The 3D QSAR approach of COMPASS acts by selecting the best bioactive conformation from a set of initial poses with optimal alignment.

15.2.3 Similarity search methods

Similarity search methods work on the principle that similar molecules tend to have similar properties, and can be further used for similarity-based VS. This starts with using a reference active compound as input and then searching other similar from the compound database based on the similarity principle. The top-ranking molecules are selected for biological testing, although there is no single measure of similarity (Mate, Hofmann, Wenzel, & Heermann, 2014). The components of similarity measure fall under three categories: calculation of molecular descriptors, similarity coefficient, and a weighing function that enables both equal and nonequal contributions from all parts of given data.

The screening of new active molecules uses computational representations for various descriptors. This method reveals the structural characteristics of the retrieved molecules by accepting the fact of the online database. Here, a prepared molecular file of compounds is used as input for calculating the different types of descriptors. In automated full QSAR modeling, computers use different fingerprints and then subjected to descriptor calculations as including Morgan, Feat Morgan, AtomPar, Torsion, RDkit, Layered, MACS, and Pattern algorithms (Landrum, 2017).

15.2.4 2D finger prints

Here, molecules are represented as a binary vector and each bit in the string represents the binary vector of one molecular fragment. The typical length would vary ~ 1000 bits. The binary code for the presence of molecule represents 1 whereas 0 for the absence of each fragment in the molecule. This binary system works efficiently with high accuracy in substructure search. If the query substructure presents in the database, then the molecule is set to 1 in the query line and must be set to 1 in the database as well. The principle of similarity depends on determining the number of bits that exist in common to any two structures. Examples are dictionary-based fingerprints where predefined fragments are searched and each of which maps to every single bit (Willett, 2010).

This function is based on one-to-many mapping of a fragment and bits used in the bit string. Each of the fragments is generated using divergent hashing functions, and each of which sets a single bit in the fingerprint. Here, fragments are produced algorithmically without a dictionary.

The Tanimoto coefficient is also used as a similarity coefficient for binary bit strings.

$$\text{SIMRD} = \frac{C}{R} + D - C$$

where C bits are set in common for the reference and database structure, R and D bits are set in database and reference structure, respectively. This represents a more complex form for use with nonbinary data, such as physiochemical property vectors. This coefficient applies to the cosine coefficient, Euclidean distance, and Tversky index.

15.3 3D fingerprints

A 3D fingerprint is based on the presence or absence of geometric features and functions that follow conditions, such as pairs of atoms at a given distance range, triplets of atoms, and associated distance. It functions by the pairing of pharmacophore as triplets (donors, acceptors, and aromatic centers), valence angles, and torsion angles.

15.3.1 3D similarities depend on the alignment

Rapid overlay of chemical structures where molecules in 3D with similarity scores based on common value are aligned.

$$\text{SIM} = \frac{V_c}{V_a} + V_b - V_c$$

15.3.2 Conformational flexibility

The different forms of conformations arise from rotation around single bonds (torsion angles) with varied ring conformations. These rotations result in the formation of N rotatable bonds by $(360 \text{ degrees}/\theta) \times N$ of combinatorial explosion. If we select a small increment in torsion angle, then a molecule with some rotatable bonds will represent a large number of conformations (Wahi, Freyss, & von Korff, 2019).

15.3.3 Scaffold hopping

The process of developing novel compounds by modifying the molecule from the central core can be called scaffold hopping. Descriptors, such as reduced graphs, provide summary representations of chemical structures, topological pharmacophore keys, and 3D descriptors.

15.3.4 Fragment-based drug design

This application has been well established for the rational development in drug discovery (Price, Howard, & Cons, 2017). Fragments with higher affinity are grown, and as well linked to adjacent binding compounds that show higher affinity. Table 15.2 represents the fundamental steps to be followed while performing the pharmacophore model hypothesis.

15.4 Pharmacophore mapping

15.4.1 Diverse conformation generation

The bioactive conformation of the selected model is important due to the conformational space of a molecule. The conformations similar to bioactive conformation are searched in four ways: (1) systematic search in the torsional space: systematic search algorithms are used in minimizing rotatable bonds, (2) clustering method, (3) stochastic methods, such as Monte Carlo, and (4) molecular dynamics (Smellie, Teig, & Towbin, 1995). Conformational search methods provide vast coverage of the conformational space within the accessible region (Kristam, Gillet, Lewis, & Thorne, 2005).

Generation of the FAST conformation accounts for the torsional space with less time consumption. The BEST conformation channel preserves rigorous energy minimization, and optimization in both torsional, and the poling

TABLE 15.2 Fundamental steps of pharmacophore modeling.

S. no.	Major steps
1.	A chemically diverse set of structures with defined experimental activity
2.	Conformation generation
3.	Generation of 3D pharmacophore model employing the training set of compounds:(a) Constructive phase: pharmacophore generated by common active molecules(b) Subtractive phase: removes pharmacophore which is not likely to be useful(c) Optimization phase: optimize the model and generate a hypothesis
4.	Quality assessment: select pharmacophore hypothesis based on the cost function
5.	Validation of the selected hypothesis
6.	Fisher's randomization test for validation of pharmacophore hypothesis
7.	Mapping of the test set of compounds on hypothesis and prediction

conformation algorithm by CAESAR. The CAESAR counts on the platform of division and conquers, and recursive approach by combining rotational and topological symmetry.

15.4.2 Generation of 3D pharmacophore

Hypogen and HipHop are the two algorithms used for the generation of the 3D pharmacophore model (Accelyrs, 2010; Sutter, Guner, Hoffman, & Waldman, 2016). HypoGen algorithm uses activity values for training set compounds to generate or build a 3D pharmacophore with up to five features. The three phases of modeling a pharmacophore are (1) constructive phase, (2) subtractive phase, and (3) an optimization phase.

15.4.2.1 Constructive phase

HipHop algorithm works on the feature-based principle, which obtains information from an active set of compounds. This phase does not require a template, rather each molecule is considered as a template. The template molecules have different configurations obtained through a reduced exhaustive search, which functions by utilizing a small set of molecules until no further larger configuration is obtained. The HipHop advent defines the number of molecules required to form a complete or partial model of pharmacophore configuration.

15.4.2.2 Subtractive phase

This pharmacophore phase eliminates the unnecessary data structures.

15.4.2.3 Optimization phase

The annealing algorithm is used in this phase. The different features depicted in the models are:

1. Hydrogen bond acceptor: Does not work for primary, secondary, and tertiary amines protonated at physiological pH.
2. Hydrogen bond donor: It does not match with electron-rich pyridines, and imidazole remains protonated, but the nitrogen has been protonated with high basicity.
3. Hydrophobic or hydrophilic: A neighboring set of atoms that are charged to electronegative atoms has surface accessibility including phenyl, cycloalkyl, and methyl groups.
4. Negative charge: These are not adjacent to a positive charge.
5. Negatively invisible: It matches many atoms or groups likely to be deprotonated at physiological pH.
6. Positive charge: These are not placed adjacent to negative charges.
7. Ring aromatic: The entire hypothesis is analyzed with their correlation coefficient with the cost function value.

15.4.3 Validation of the pharmacophore model

15.4.3.1 Analysis of Fisher's randomization test

The structure–activity relationship is known to correlate with the statistical significance considering cross-validation. This is achieved by assigning new values to the activity data in the training and test set. Generation of the original model by randomized data set results has to correlate with the original model of pharmacophore.

15.4.3.2 Test set prediction

This approach is used to measure the activity for a training set of compounds using an external data source. The conformers have been generated based on the set of test molecules. The prediction of activity for a training set of compounds is performed for internal validation of the QSAR model. The correlation between training and test set of molecules is predicted based on their fitting to the designed pharmacophore model. External independent data sources of compounds can be used for model validation.

15.4.3.3 Guner-Henry scoring method

The Guner-Henry (GH) scoring estimates the quality of the pharmacophore models (Guner & Henry, 2000). The GH score is used to quantify the selectivity precision of hits from actives of useful decoys set (DUD) (Huang, Shoichet, & Irwin, 2006). The backing ways are the metrics used for analyzing hits based on the database search by a pharmacophore model.

15.5 Pharmacophore classifications

In ligand-based pharmacophore (LBP), superimposition of a set of active molecules is done to find the chemical features, such as ring, hydrogen bond donor, and acceptor that are crucial for maintaining the biological activity of the compounds. Structure-based pharmacophore (SBP) modeling is performed using interaction points between macromolecular targets with their respective ligands.

15.5.1 Ligand-based pharmacophore modeling

In the absence of a 3D macromolecular structure, the drug discovery process is carried out by searching a known set of ligands that interacts with macromolecular targets. Considering this issue, pharmacophore development for a set of compounds faces two major challenges, the first issue represents a large number of conformation for a ligand due to rotational flexibility and their alignment in case of multiple ligands, while the second issue is to determine the prerequisite common chemical features for building the pharmacophore model. Molecular alignment and conformational analysis tools solve many complexities in LBP modeling.

Characteristics of a good conformation for a molecule should hold the following properties:

1. Low time for conformational calculations;
2. Low energy conformations as a measure of preciseness; and
3. Generates conformations of small molecules which is biologically active to interact with a target.

Let us throw some light on few challenges despite many improvements in LBP modeling:

15.5.1.1 The first problem lies with the ligand flexibility

There is a method called preremunerating in which conformations of the ligand are precomputed and saved in the database (Poptodorov, Luu, & Hoffmann, 2006). In this way, the computational complexity of generating conformers for molecular alignment significantly reduces and also solves the need for a storage capacity.

15.5.1.2 Molecular alignment in LBP modeling

Molecular alignment methods fall under two categories, that is, point-based and property-based approaches. The point-based approach is based on the alignment of atoms, fragments, and chemical features (Dror, Shulman-Peleg, Nussinov, & Wolfson, 2006). Here, a couple of atoms, fragments, or chemical features is superimposed using the least-squares fitting approach (Fig. 15.2). The property-based algorithms utilize molecular fields, such as Gaussian functions to generate an alignment. There has been a drastic advancement in the algorithm used for alignment, which includes stochastic proximity embedding (Bandyopadhyay & Agrafiotis, 2008), atomic property fields (Totrov, 2008), fuzzy pattern recognition (Nettles et al., 2007), and grid-based interaction energies (Baroni, Cruciani, Sciabola, Perruccio, & Mason, 2007).

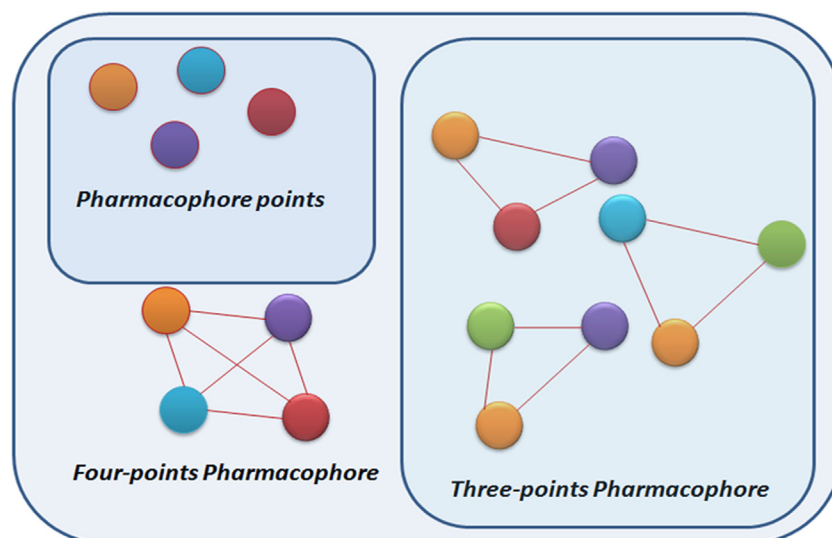


FIGURE 15.2 The circular-shaped identities represent points used in point-based pharmacophore.

15.5.1.3 Selection of a training set of compounds

In the LBP modeling, the selection of a training set affects the finally developed pharmacophore model, and the selection of training set compounds is very important for the development of a better and precise model (Poptodorov et al., 2006).

15.5.2 Structure-based pharmacophore modeling

This approach depends on the 3D structure of the macromolecular receptor or macromolecule-ligand complex. This approach is very similar to the docking procedure which provides a similar level of knowledge. This is considered as one of the complementary techniques to the docking protocol that considers the information of the active site with their dimensional relationship to build a pharmacophore model.

15.5.2.1 Macromolecule ligand complex

The ligand–target complex determines core interactions between the ligand and the target in the binding site. LigandScout (Wolber & Langer, 2005) is promising software that is based on the macromolecular-ligand complex scheme. Software tools, such as Pocket v.2 (Chen & Lai, 2008) and GBPM (Ortuso, Langer, & Alcaro, 2006), work on the same approach. The SBP method is a typical example of a target-based pharmacophore model development (Accelrys, 2010). The most significant obstacle is that the acquired pharmacophore model cannot be used to replicate the information for the QSAR model because the SBP works with either a selected macromolecule-ligand complex or a macromolecule.

15.5.2.2 Feature-based outlook

This combines both the macromolecule-ligand-based features along with unoccupied binding site information. It analyses the 3D structure of target–ligand complex and trains a complete set of chemical and geometric rules to generate an accurate pharmacophore model (Friesner, Murphy, & Repasky, 2006). In the case of apo-structure, ligands are docked into an apo-binding site using the docking tools, such as the Glide XP program (Salam, Nuti, & Sherman, 2009).

15.6 Application of pharmacophore in virtual screening and de novo design

15.6.1 Virtual screening

VS deals with the process of searching the ligands from the 3D chemical database based on the pharmacophore model. The pharmacophore hypothesis is taken as a template and compounds searched and screened based on feature matching with the template. Screening of large databases is a big task, yet it handles some major obstacles:

1. Distinct molecules can be retrieved in VS based on a single pharmacophore model from different hypothesis, which could be the reason for higher false-positive/false-negative rates in many cases.
2. Robust development and pharmacophore validation is the most critical because it requires inclusive validation and optimization.
3. Recent attempts (Gimenez-Oya, Villacan, Fernández-Busquets, & Rubio-Martinez, 2009) to involve molecular dynamics simulations into pharmacophore modeling stated that trajectories obtained from multi-dimensional scaling represent the flexibility of pharmacophore.

Operating the conformational flexibility and pattern identification of small molecules holds the most important factor in VS design and development.

Conformational flexibility has two objectives:

15.6.1.1 Conformer selection

Low energy conformational analysis is performed to a wide range of a diverse set of molecules when registered in a database. Each such conformer becomes a rigid molecule.

15.6.1.2 Exploration of conformational space

Here, the triangle smoothing approach can be used to recognize the minimum–maximum distance between each atom pair. It creates a distance-range apart from the distance graph for the structure of compounds. The screen and graph

search is made by minimum–maximum distance using altered algorithms. Therefore the last conformational analysis of the compounds results from the search/graph using the torsional angles.

15.6.2 De novo pathway

The present idea has been designed to identify novel candidate drugs that are similar to the chosen pharmacophore model. The pharmacophore-based de novo design began with NEWLEAD software (Tschinke & Cohen, 1993). Disconnected molecular fragments are used as molecular inputs for the given pharmacophore model. These disconnected sets of pharmacophore models are linked using atoms, ring moieties, or chains. The main drawback of this software is that it limits only to the functional groups rather than chemical features. Pharmacophore-based de novo ligand designing of the drug-like molecules can be used to solve the issues that arise with other ligand designing software tools (Huang et al., 2010). It creates drug-like compounds for a given input and builds the pharmacophore hypothesis.

15.7 Advancement in exploring 3D pharmacophore principles over the above limitations

The 3D pharmacophore models are generated using an atomistic platform of macromolecules or by the superimposition of conformation of multiple ligands that are used to estimate the structure–activity relationships or compounds screening. Some advanced approaches useful in pharmacophore modeling are listed in Table 15.3.

15.8 Quantitative structure–activity relationship

The Field of QSAR modeling has pioneered effectively in the present era of computer-aided drug design. QSAR tool in the fundamental science evolved through mathematical QSAR to measure accuracy in course of predicting the potency/toxicity/activity. The web of science core has published approximately 6000 papers on QSAR which then evolved

TABLE 15.3 Advanced methods used in 3D pharmacophore analysis.

Category	Approach	Description
MD integration	Hydration-site restricted pharmacophore (Lee, Krieger, Li, & Bahar, 2020)	Thermodynamic properties of water are used to reduce the number of features for model development
	SILCS-Pharm (Sato, Honma, & Yokoyama, 2010)	Probe-related binding hot-spots are used for 3D pharmacophore generation
	Dynophore (Jimenez, Doerr, Martínez-Rosell, Rose, & De Fabritiis, 2017)	It uses 3D pharmacophore features and MD simulation statistics
	Common hits approach (Jimenez, Škalič, Martínez-Rosell, & De Fabritiis, 2018)	This considers 3D pharmacophore models from MD simulations, which can be grouped according
	MYSHAPE (Skalic, Jiménez, Sabbadin, & De Fabritiis, 2019)	MD simulation-based interactions are used to refine shared-featured pharmacophores
	GRAIL (Skalic et al., 2019)	The molecular interaction fields are used in model development
Machine learning	Water Pharmacophore (Schneidman, Dror, Inbar, Nussinov, & Wolfson, 2008)	Water thermodynamics, docking analysis, and molecular interaction fields are used
	HSPHarm (Wang, Shen, & Wang, 2017)	Random forest decision trees trained over pharmacophore fingerprints and to reduce the number of features used in model developments
Web applications	DeepSite and related software (Koes & Camacho, 2012)	The neural network approaches are used to train the pharmacophoric descriptors related to cavities (binding affinities)
	PharmaGist (Koes & Camacho, 2012)	3D pharmacophore development based on ligand
	PharmMapper (Koes & Camacho, 2012)	It determines potential drug targets based on the pharmacophore model
	Pharmer-related Applications (Koes & Camacho, 2012)	The 3D pharmacophore generation and virtual screening of small molecule databases

substantially in the past five years. The classical methods and models developed into advanced versions, and they are being used for the development of drugs/chemicals, agrochemical (herbicides), and cosmetic products. However, novel ways and applied areas also have emerged for prediction and optimization of process, design, and synthesis, and thus the models become a fundamental part of the drug discovery process which provides significant direction in the planning of the experiments. QSAR modeling works by exploiting the chemical structures, and their biological activity to meet the demand for novel drug compounds. In chemoinformatics, compounds are represented in different ways using mathematical and statistical methods for analyzing properties (Lo, Rensi, Torng, & Altman, 2018). QSAR model development and applications have the following steps: preparation of data, analysis of data, model development, validation of the model, and VS of chemical databases (Fig. 15.3).

15.8.1 Designation of QSAR

There are different dimensions of QSAR based on criteria and features used for building the model.

1. 1D: Molecular properties, such as pKa and log P with molecular activity.
2. 2D: Connectivity indices and 2D pharmacophore are considered with structural patterns.
3. 3D: Noncovalent interactions fields shielded through molecules.
4. 4D: Correlation between ligand configurations.
5. 5D: Correlation of induced fit model in 4D QSAR.
6. 6D: Varied solvation models in 5D QSAR.

15.8.2 Backbone of chemical similarity

Traditional QSAR models have followed linear (regression) models obtained through a diverse set of compounds (ligand) with similar biological properties. QSAR model is expected to predict a change in the biological activity of a compound as a result of structural modifications. Through the advancement of machine learning (ML), a nonlinear regression model was developed as the most important landmark in QSAR analysis. Bioactivity data can be determined for the compounds that share a common core (scaffold) structure with varied R groups at more than one site. QSAR model plans on the fundamental platform of structurally similar compounds with the same biological function. The

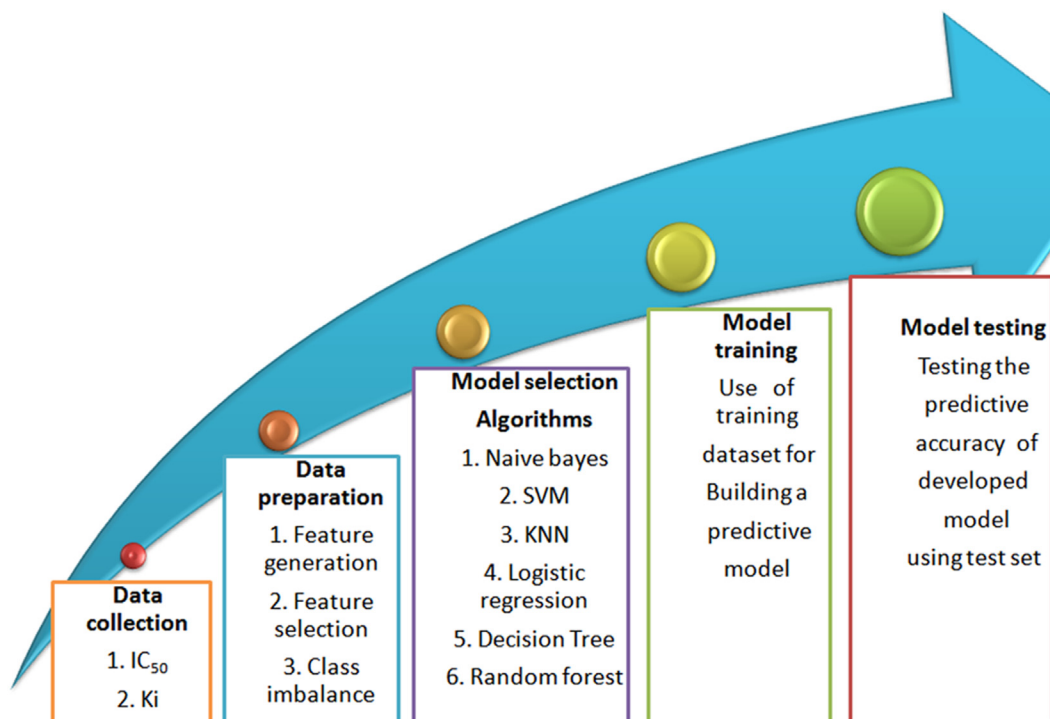


FIGURE 15.3 Representation of general workflow of quantitative structure-activity relationship model.

change in the structure of a compound would bring a gradual change to the biological activity in a linear fashion (regression). Functional groups attached to compounds or their analogs suggest that the effect of structural variations can be predicted to ensure potency (Leelananda & Lindert, 2016).

The feature is used to recognize various compound responses in SAR continuous versus discontinuous. This plays a very crucial role in understanding structural features and their biological response. The numerical measures are used to assess the similarity of molecular descriptors (Willett, 2006). The Tanimoto similarity calculations are performed using molecular fingerprints, mostly in ML. The Tanimoto coefficient provides similarity in the interval of 0, 1. Another point of view is a classification based on similarity. It is performed by clustering small molecules of data sets on hierarchical scaffolds extracted from them. These scaffolds are used as core structures and substructure-based similarity for compounds is searched by calculating the number of maximum common substructure between the query and hit compounds. On a large scale, this can be determined through MMP.

15.8.3 QSAR data

The analysis depends on the size, quality, and diversity of chemical structures. Advancements in the field of robotics and miniaturization have increased the diversity of data sets in recent years. The validation begins with data preparation, quality, and data curation including molecular structures to the biological data for a particular lead compound. The feature selection process follows by identifying a nonredundant set of variables for choosing the best model. Consequently, it helps in understanding the generated data and improves the prediction of relevant predictors. The data have been constructed by retrieving information from an online databank repository or build an individual (private) data set with measurements from lab experiment values or different sources.

The biological activity data for the compounds/drugs should be measured using uniform and standard protocols and preferably from a point source, such as organism, tissue, or protein. The 3D QSAR model utilizes the most accurate activity data for the development of a good and robust model. The QSAR model should be designed using diverse data sets for maintaining the accuracy of the model (Kim, 2001; Sybyl, 2005). The flow of steps for the 3D QSAR model has been shown in Fig. 15.4. A QSAR model is build using a training set of compounds, and then, it is validated either by taking the test set from the parent data or by taking external data for a new experimental source. The following points can be considered for accurate biological data.

1. The compounds need to undergo the binding mode through the same mechanism of action of repeated actions.
2. A compound's biological activity correlates with its binding affinity.
3. The biological data are supposed to be distributed symmetrically around its mean.

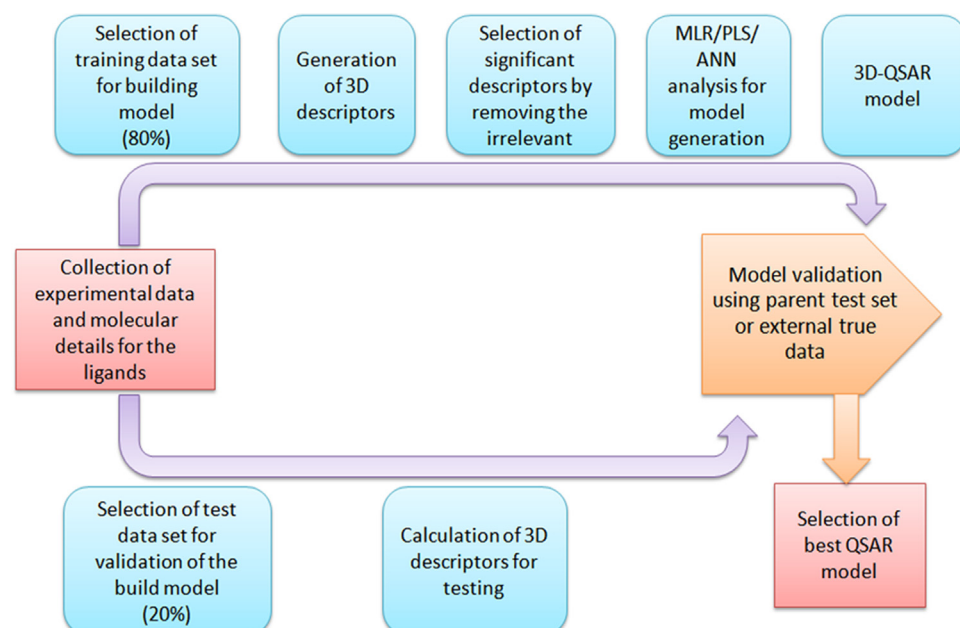


FIGURE 15.4 Flow of steps involved in building 3D quantitative structure–activity relationship model.

15.8.4 Classification of 3D QSAR approaches

The 3D QSAR is categorized into three basic types based on alignment dependent/independent, linear/nonlinear, and ligand/receptor-based criteria (Fig. 15.5). The QSAR model development is based on the following criteria:

1. alignment mode;
2. chemometric techniques; and
3. based on the receptor and ligand molecules.

15.8.5 Molecular interactions and energies

The superimposition of molecules has been overlaid by a set of molecules and the center of grid box known to calculate the interaction energies between the ligand and the receptor molecule with different probe atoms when placed at each intersection of the lattice (Allen, 2002; Kim, 2001). The methodologies for CoMFA are as follows: the interaction energy plays a paramount role in the grid spacing (0.5 Å), during molecular superimposition for a reference (most active) compound based on the fields. The centroid of grid box size measures approximately 3–4 Å with the surface of overlaid molecules. The electrostatic energies among lattice points nearby and inherent correlation with a similar grid box are used for the calculation of steric and Van der Waals interactions. Grid box location affects the significant features in the CoMFA model and models are developed to find the best grid position. Probes are used for the calculation of interaction energies in CoMFA. The probes measure interactions for substances like water or a chemical fragment like ethyl/methyl group necessarily functional group significantly. The intersection points of the lattice of every probe positioned in turn are the energies obtained for the compounds and calculated by molecular force fields. A mathematical equation for the force field combines with parameters, such as bond length, bond angles, dihedral angles, distance along with the coordinates, and other similar groups, that fit empirically with the potential energy surface. The Standard Lennard-Jones function is used to model the Van der Waals interactions while Coulomb's law supports electrostatic charges (Bell et al., 2008). The deepest steep of the slope depicts the Van der Waals surface that represents the potential energy to the proximity of the surface at lattice points.

15.8.5.1 Description of molecular shape analysis

Molecular shape analysis can be described as a ligand-based QSAR that uses the advantage of merging conformational analysis with the classical Hansch perspective (Hopfinger, 1980). The workflow involves the molecule for all the torsion angles within a fixed geometry by an intramolecular conformational analysis. The conformational analysis uses molecular mechanics to find the steric, electrostatic, and hydrogen bonding within the fixed geometry.

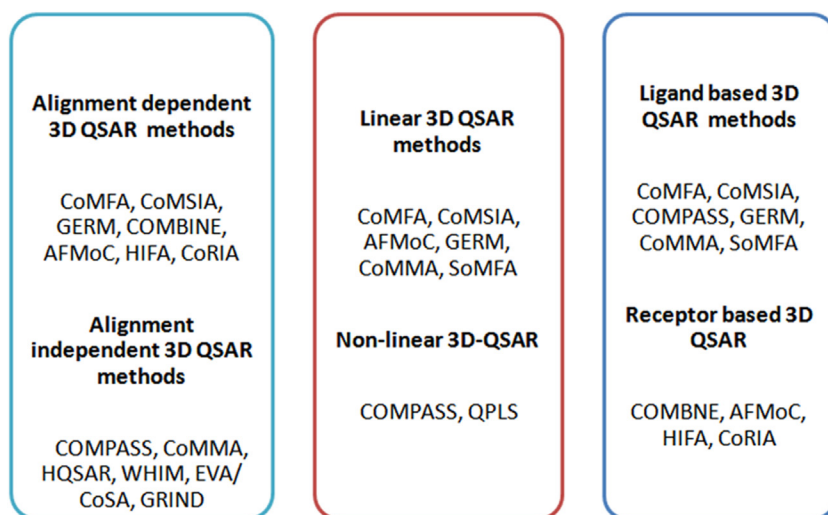


FIGURE 15.5 Classification of interaction energies on different criteria.

15.8.5.2 GRID

The molecular field analysis and interaction energy are calculated and that exploits nonsecured interactions between 3D structures and a probe settled at the sample positions on a lattice throughout the supermolecule. The program computes steric and static potentials, conjointly calculates the gas bonding potential employing a DRY probe. The water probe is supposed to feature and calculate hydrophobic interactions (Kim, 2001) and can settle for electrons; thus it is represented through hydrophobic interaction energy like logP. Various energy levels at each grid are displayed in conjunction with the macromolecule structure. The negative energy levels of the contours describe the regions at which binding ought to be favored, whereas the positive energy levels determine the form of the target. GRID considers surface properties of energetically favorable binding sites for building pharmacophore models, and rational style of potent inhibitors of GRID maps is being used as descriptors (Pastor, Cruciani, & Watson, 1997). The GRID software package has been provided by Molecular Discover Ltd (GRID, 2009).

15.8.5.3 Determination of comparative molecular moment analysis

The comparative molecular moment analysis (CoMMA) computes the similarity in descriptors at the molecular level on the standards of molecular mass and charge distribution as well as a second-order of reactions (Silverman & Platt, 1996). The molecular structure has two mathematician reference frames, out of which is the principal mechanical phenomenon quadrupole axes are calculated with respect to the molecular center of the dipole. This technique has been designed to develop QSAR of animal products or metabolites (CoMMA, 2009; Kovatcheva et al., 2004). Programs CoMMA and CoMMA2 are available for descriptor analysis (Silverman, 2000).

15.8.5.4 Detailed analysis of comparative molecular surface analysis

Comparative molecular surface analysis (CoMSA) represents a nongrid three-dimension QSAR strategy (Polanski & Walczak, 2000; Polanski, Gieleciak, Bak, 2002). The Kohonen's self-organizing map (SOM, a sort of neural network) remodels by deriving features to the points in Cartesian coordinates, converting 3D molecules into 2D geography maps, which are sampled haphazardly in the least of the Van der Waals of the molecules. The topological map remains as the projections of the partial atomic charges. The points are collected from each grid and calculated as the molecular electric potentials at the surface itself. CoMSA is used for the modeling of pKa values, virtual combinatorial library screening, and 3D QSAR (Gieleciak & Polanski, 2007; Jojart, Martinek, & Marki, 2005).

15.8.5.5 Comparative residue interaction analysis

Comparative residue interaction analysis is a 3D QSAR method that exploits the descriptors which describe the physical science parameters in ligand interaction (Datar, Khedkar, Malde, & Coutinho, 2006). The descriptors, such as strain energy, binding, area, lipophilicity, molecular volume, and molar refractivity, account for the correlation with the biological activities of the molecules using the generic version of the partial least squares (PLS) technique (G/PLS). This methodology has been used as the solution for the complication of peptide QSAR (Dhaked, Verma, Saran, & Coutinho, 2009).

15.8.5.6 Adaptation of fields for molecular comparison

It is a 3D QSAR approach based on fields derived from the macromolecule environment, henceforth it is known as the "reverse" CoMFA [adaptation of fields for molecular comparison (AFMoC)] approach (Gohlke & Klebe, 2002). This procedure begins by putting a grid into the receptor-binding site. This atom-based interaction field has to correlate with neighbor grid values with interaction fields, which are then correlated with binding affinities. AFMoC-derived fields are represented using different maps and binding affinities for a ligand can be shown.

15.8.5.7 Representation of COMFA results

The interaction energy at every grid point exhibits a correlation with biological properties and activity by generating an equation. Contour plots are being used to depict the regions around the molecules that are significant for biological activity. The positive contours are shown for steric fields are depicted in green color as bulk and most favored regions, while yellow color stands for negative contours favoring less favored region, and blue color represents negative contours that favor electropositive substituent. Score plots are used between biological activity and latent variables/descriptors to represent the structure–activity relationship (Worley & Powers, 2013).

COMFA map for chalcone derivatives as anticancer agents has been generated to represent the effect of steric and electrostatic features on the activity of the molecules (ElMchichi, Belhassan, Lakhliifi, & Bouachrine, 2020). In Fig. 15.6A, the CoMFA contour map for the steric region contains two green regions which indicate that bulky groups at this position will increase the inhibitory activity of the compound, whereas two yellow regions suggest that a bulky group at this position will decrease the inhibition activity. In the CoMFA electrostatic contour map Fig. 15.6B, the blue map shows the region where negative electrostatic potential will decrease the activity of the compound, whereas the red contour map suggests that electronegative groups at this position will improve the activity of inhibitors.

15.8.5.8 Concluding remarks for interaction energies

Drug discovery methods have been widely accepted during this era in major fields despite their pitfalls and caveats. The QSAR guidelines given by the community of OECD are supported by well-established principles. The reliability of QSAR models depends on the training set, molecular alignment, conformational analysis, and used mathematical model. The accuracy of final models is validated using an internal training set as well as also by using an external validation approach with an external source experimental data. Moreover, QSAR results are used for approval or disapproval of biological testing of a compound. During this era of ever-increasing demand for the drug discovery process, QSAR models are a very important part of the analysis (Verma, Khedkar, & Coutinho, 2010).

15.8.6 QSAR modeling

The mathematical modeling-based relationship is calculated between molecular descriptors to their biological activities. QSAR modeling uses multiple linear regression (MLR) or nonlinear modeling algorithms, such as artificial neural networks (ANN). The QSAR modeling is based on either linear or nonlinear or combinations of both that are expressed in the form of the mathematical equation using molecular descriptors. The QSAR models are categorized as (1) regression-based QSAR models, (2) classification-based models, and (3) development of focused libraries on the criteria of features for 3D pharmacophore by VS of libraries. Evaluation of model building is based on feature selection. The model without feature selection is assessed on a complete set of descriptors removing irrelevant or noninformative variables to improve the predictability for the given model.

15.8.6.1 Determination of multitask modeling

The traditional way of the QSAR model predicts a single compound and its biological activity. In recent times, multiple activities are used for prioritizing compounds based on multiparameter optimization or multitask optimization. This is achieved through a single task model by using a neural net or techniques based on ANN and deep learning. Multitask optimization features as the most active area in the development of QSAR modeling. The fitness of the QSAR model depends on the accuracy of input data, approaches used for descriptors' selection, as well as the statistical and mathematical methods used for building the model.

15.8.6.2 Validation of QSAR modeling

The organization for economic cooperation and development has proposed five validity principles that every QSAR model has to satisfy for regulatory purposes. In proportion to guiding principles, the QSAR models should have defined

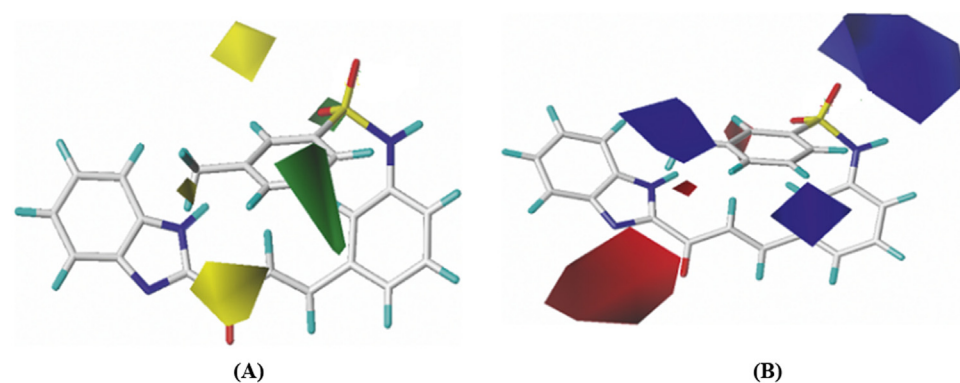


FIGURE 15.6 (A) Steric contour maps and (B) electrostatic contour maps of the CoMFA model.

endpoints, unambiguous algorithms, and pharmacological interpretability. The data set includes the larger number of independent variables (x). To extract a stable and the best QSAR model, a wide range of descriptors can be selected using PLS to interpret biological activities in the best way. The other methods, such as MLR, PA, and PCR, have been modeled to obtain linear relationships. Henceforth, the connecting link of dependent (y) and independent (x) variables is either leaves behind a great ambiguity (Lambrinidis & Tsantili-Kakoulidou, 2018).

The external test set is used for validating the QSAR model. The complete data are classified as a training set, and a test set, out of which test set is used for prediction, and a metric set is then checked for accuracy. Several strategies have been developed for QSAR validation. A scatter plot showing the correlation between predicted and observed activity of the training set and a test set of molecules has been shown in Fig. 15.7. For a good-quality model, there should be a high degree of correlation between predicted activity (using QSAR model) and observed activity of the compounds for both sets training as well as test.

15.8.6.3 Different validation methods

Internal validation takes account of the training set of compounds whereas external validation test set of compounds. However, the calibration subset constructs the model while the validation subset predicts the model for the new data. In the case of external validation, a chemical set of the same compounds will be considered in this approach. Distribution of the data files into training. In random selection, the selection of a complete set is classified as training and test set randomly. A crucial part of a statistically significant QSAR model is the size of the training and test set. Training and test set selection is done by using approaches: (1) k-clustering: it is one of the nonhierarchical clustering technique; (2) Kohonen's SOM selection; (3) statistical molecular design; (4) Kennard-stone selection; (5) sphere exclusion, and (6) extrapolation-oriented test set selection.

15.8.7 Concept of applicability domain and QSAR approaches

Applicability domain (AD) tool makes the center of QSAR. Several modes for AD are outlined based on the similarity principle (Nikolova & Jaworska, 2003). Therefore AD is portrayed as the space valued for the building of the model. The problem of AD estimation remains still active and ever-growing in the research field. Many QSAR/QSPR models have been derived through different hypotheses and algorithms. Hanser, Barber, Marchal, and Werner (2016) highlighted AD in three different aspects, such as applicability, reliability, and decidability. For a better reliable model, test set predictions should be close enough to experimental observation. Algorithms can be assessed for the AD of QSAR/QSPR models in search of chemical compounds, or chemical reactions that describe the quality of prediction. Different approaches used for building the QSAR model are shown in Fig. 15.8.

Regression models are easily interpretable methods for the correlation of an independent variable (x) with a dependent variable (y). Simple linear regression generates a standard linear regression model that calculates the output (y) for QSAR equations of single independent variables (x). It is expressed $y = a + bx$, where y denotes a dependent variable, x means independent variable and represents the constant of the intercept, and regression coefficient is b . This method is used to develop simple relationships between structures and biological activity using fundamental descriptors governing

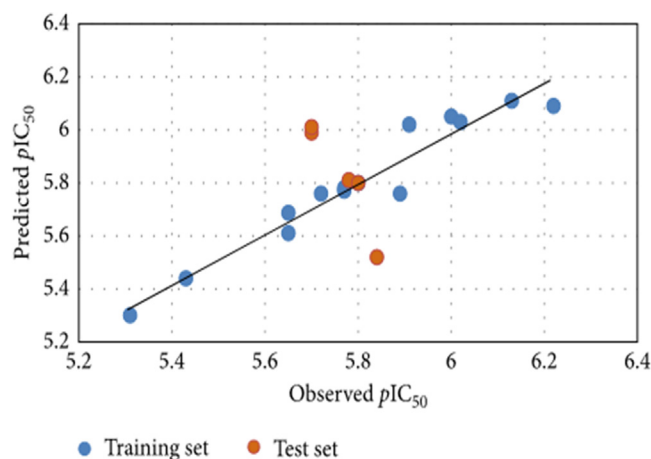


FIGURE 15.7 Scatter plot showing predicted versus the actual pIC₅₀ values for training and test data set based on CoMSIA model.

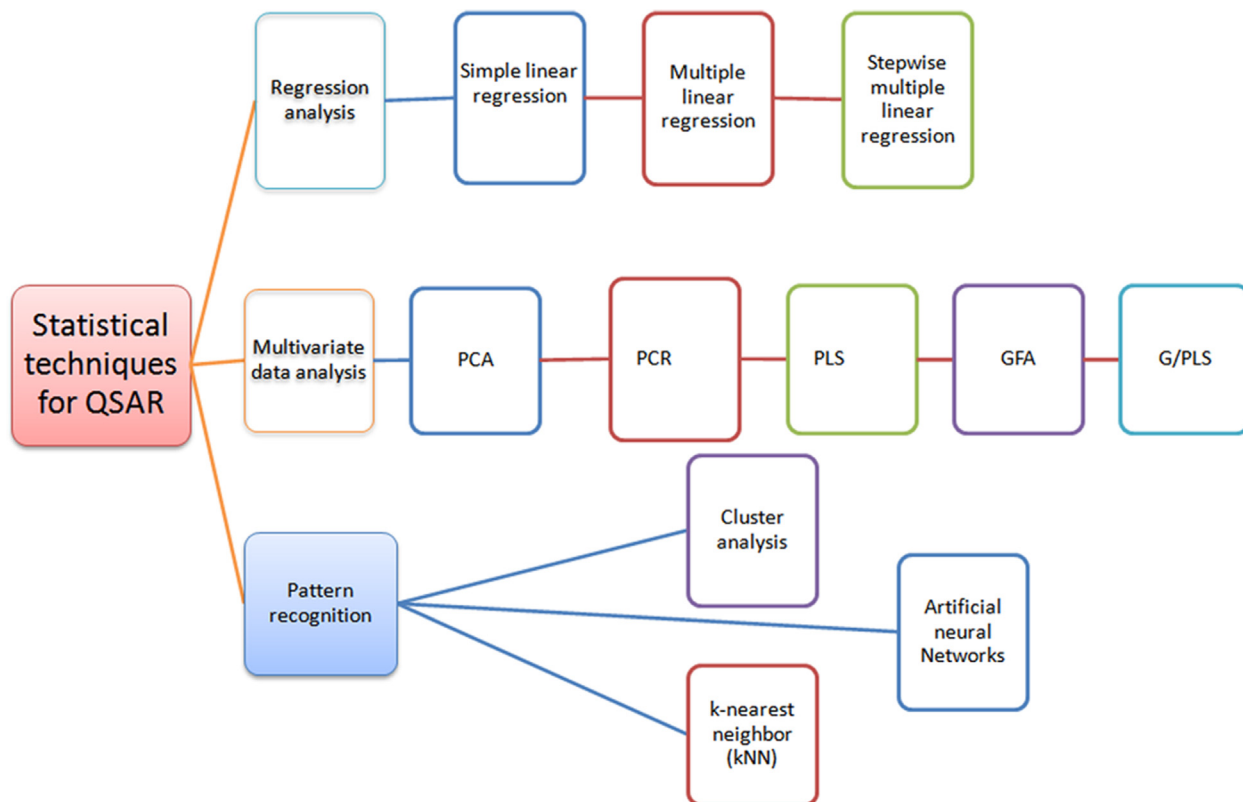


FIGURE 15.8 Diagrammatic representation of statistical techniques used in the building of the quantitative structure–activity relationship model.

the activity as input values. The MLR method is used for calculations of druggable properties using standard multivariable regression and finds the dependency on the whole of the given descriptors. The correlation is measured by the parameters of dependent factors such as multiple correlation coefficient, Fisher F ratio, t -test, standard deviation, and also independent tests, such as the leave-one-out (LOO) method. Classification of QSAR models based on linear, non-linear, and mixed approaches are shown in Fig. 15.9.

In the case of a large number of observations, the principle of PLS confers a statistical solution for independent variables. Therefore the PLS is used for the selection of significant descriptors as all descriptors are not relevant for the model. Principal components analysis (PCA) is used to find the significant component by reducing the dimension of larger data sets based on relationships among independent variables (Dunteman, 1989). This opens a new door for a new set of components called orthogonal descriptors, principal components, to find the information present in the independent variable. PCA is used to deduce a multivariate data set of descriptors dimensionality (Dunteman, 1989). The independent variables become the fundamental components to execute a linear regression.

ANN is defined as parallel computational systems of highly interconnected processing elements called neurons (Baskin, Palyulin, & Zefirov, 2008). The initial layer is known as an input layer, and each neuron is made to receive data from outside/user as inputs for QSAR. Consequently, there are hidden layers that are present one to many from the input layer. Every input descriptor value is represented by the weight according to its significance in the prediction. The inputs weighed are added and have been supplied to the hidden layers, which consider the processing of a nonlinear transfer function. The output layer is compelled to the neurons where results are generated.

The most commonly classified technique to describe a new pattern (a molecule) in ML is k-nearest neighbor (Ajmani, Jadhav, & Kulkarni, 2006). The Euclidean distance matrix is used to find the similarity measure between the molecules using the structural descriptors. The training sets are computed using Euclidean distance by connecting through an unknown object (u), with all the known objects. The calculated distance, k objects from the training set chosen by the user would be similar to the objects. A test set of samples are selected by leave one out cross-validation with optimal k value through categorization. The decision tree comes under the nonparametric learning method for classification and regression analysis. The algorithm works through dividing and conquering the portion of data into subsets with the similarity principle applied in values.

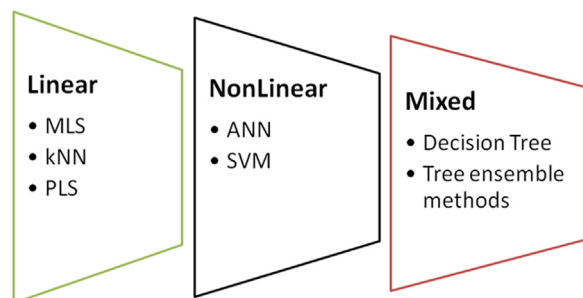


FIGURE 15.9 Diagrammatic representation of the classification of models for quantitative structure–activity relationship.

15.9 Development of new QSAR: HQSAR

Hologram quantitative structure activity relationship (HQSAR) exploits the molecular substructures in binary patterns (i.e., fingerprints) as descriptors in QSAR models. This approach uses all 2D structures, and biological activity as inputs and these are converted to linear, branched, overlapping fragments. These fragments make use of integer values to test the algorithm, thus making an integer array. These arrays can also be called molecular holograms, and spaces occupied by holograms are used as descriptors. It is built by using PLS regression and then validated by the LOO method. HQSAR is preferred over 2D/3D techniques because of the merits: (1) the discrepancy of activity with the shift of a substituent can be readily identified, (2) differentiate fragments using stereochemical and hybridization state, (3) correlations are comparable to 3D QSAR techniques, reduce the time, generate conformers, and mutual alignment in 3D space, and (4) encodes all possible fragments, and subfragments to encode imperative specificity for the region information about fragments. The HQSAR approach helps in the distinction of the parameters, such as toms, bonds, connections, hydrogens, chirality, donor, and acceptor (Hasegawa & Funatsu, 2000).

1. Atoms: The atom parameter allows fragments to differentiate between elemental atom types (selecting O distinguished from N).
2. Bonds: This parameter uses bond order for differentiating fragments (pentane differs from pentene in the absence of one hydrogen).
3. Connections: The connection parameter is a measure for atomic hybridization in the patterns for fragments. Thus HQSAR connects constituents of atoms, their bond distribution, and bond orders.
4. Hydrogens: HQSAR access hydrogen bonding patterns during the generation of fragments.
5. Chirality: Stereochemistry differentiates between *cis* and *trans* forms of structures of fragments based on atomic and bond orders.
6. Donor/acceptor: HQSAR help in the distinction between hydrogen donor and hydrogen acceptor (Fig. 15.10).

15.10 Application of QSAR/SAR

To date, QSAR methodologies have also been applied to regenerative medicine and biomaterials modeling, which are further classified into major groups. This study differentiated between sparse and nonsparse feature selections that are meant to reduce the complexity of materials along with biological system interaction. Different tools/software widely used for QSAR analysis are listed in Table 15.4.

15.10.1 Synthetic organic chemistry and QSAR

The practical implementation of QSAR faces a new challenge in predictive computational chemistry. The area of robotics has gained rapid growth in the development in exploring cheminformatics tools for the efficient synthesis of lead molecules. The specific building blocks and retro-synthesis (specific target molecules) are the most widely used synthesis planning strategies. Synthetic routes consider kinetic parameters that are to be predicted by models. In a selected elementary reaction, reaction conditions account for solvent, catalyst, and temperature that leads to high yield that has to be suggested by the algorithm. Cheminformatics tools use a wide range of such parameters currently used in computer-aided synthesis design. The chemical reaction comprises a very elaborative modeling problem in cheminformatics. Depending on ML methods, SMILES (sequence-to-sequence models) are encoded by chemical structures (Hoonakker, Lachiche, Varnek, & Wagner, 2010; Polishchuk, 2017).

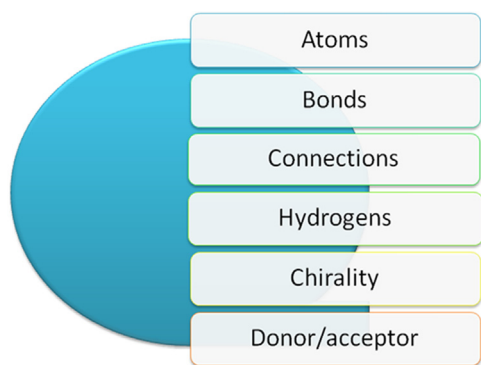


FIGURE 15.10 Schematic representation of various descriptors HQSAR technique.

TABLE 15.4 Software/tools related to QSAR study.

S. no.	Tool name	Description	External source
1.	LiSIs	Workflow system for virtual screening	http://lisis.cs.ucy.ac.cy
2.	ChemDes	Web-based platform for molecular descriptors and fingerprint computation	http://www.scbdd.com/chemdes
3.	Gusar	QSAR/modeling based on the training set, prediction, and validation	http://pharmaexpert.ru/GUSAR/antitargets.html
4.	DWFS	Feature selection tool based on a parallel genetic algorithm	https://www.cbrc.kaust.edu.sa/dwfs/
5.	VIDEAN	Visual and interactive descriptor analysis	http://lidecc.cs.uns.edu.ar/VIDEAN/
6.	BioPPSy	An open-source platform for QSAR/QSPR analysis (ADMET)	https://sourceforge.net/projects/bioppsy/
7.	LogS	ADMET prediction model for water solubility (logS) database several others, logBB, P-gp, Caco-2, oral absorption prediction	http://modem.ucsd.edu/adme/databases/databases_logS.htm
8.	QSAR Toolbox	QSAR software for the large and small dataset based on different approaches, such as MLR, PLS, and classification	http://teqip.jdvu.ac.in/QSAR_Tools/

QSAR, Quantitative structure–activity relationship.

15.10.2 Prediction of kinetic and thermodynamic parameters

The logarithm of the reaction rate constant ($\log k$) is a common endpoint in QSAR modeling. The application of QSAR to different modes accounts for the solvent effects, and temperature for chemical reactions using neural network approaches (Engkvist, 2018). The current platform of QSAR is performed on small and large datasets that account for solvent effects and varied temperatures for chemical reactions using neural network approaches (Engkvist, 2018). Descriptors computed are sequenced along with descriptors of solvent and temperature in this model. This technology demands the knowledge of the order of reactants in the rate equation, which develops an automatized QSRR workflow problematic.

15.10.3 Drug development and other applications

The important challenge of QSAR is not about finding a target but rather utilizes all the pieces of data needed to draw the additional reliable evaluation. This methodology takes the advantage of all the contributions from in vivo, in silico, and in vitro experiments. Because of this, it has become predominant to know and implement the knowledge offered by various QSAR models. To conduct all in vivo experiments would be very time taking and costly, but QSAR methods can do the analysis easily at a low cost of study. Many animal testing is required to draw the necessary conclusion for a drug compound. This raises many ethical issues and requires concerning permissions from the ethical committee. Some risky drug compounds have to be thoroughly checked with regulators and a dossier submission required by the

concerned industry. The QSAR approach estimates all the testing properties in the initial screening study and identifies the compounds that may be toxic.

The concept of LBDD is based mostly on drug planning, which is to develop the simplest drug molecule once a 3D structure is unknown through trial and error methods. This approach reduces the time for drug development and also improves the biological activities of a drug. The computational studies have been increased rapidly to handle complexity, produce accurate results, enhance reproducibility, lower cost, and identify the novel target. The innovations in computational tools have improved the VS and de novo drug designing, along with protein-ligand interactions studies. Therefore the constant efforts by QSAR experts have improved the process of in silico drug discovery with good accuracy (Sliwoski et al., 2013).

15.11 Conclusion

The process of modern drug discovery utilizes the knowledge of genomics, bioinformatics, computational chemistry, and combinatorial chemistry for VS, HTS, de novo ligand design, in silico pharmacokinetic screening, and in vitro studies. This chapter has addressed major issues in ligand-based drug design and QSAR analysis. QSAR analysis estimates the biological activities of compounds based on the model developed by deriving the chemical descriptors. QSAR-based prediction also guides the changes required in the lead compound to attain a better binding affinity, efficacy, and selectivity to a drug target.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Accelrys Inc. (2010). *Discovery studio 2.1*. San Diego, CA: Accelrys Inc.
- Ajmani, S., Jadhav, K., & Kulkarni, S. A. (2006). Three-dimensional QSAR using the k-nearest neighbor method and its interpretation. *Journal of Chemical Information and Modeling*, *46*, 24–31.
- Allen, F. (2002). The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Crystallographica Section B*, *58*, 380–388.
- Bandyopadhyay, D., & Agrafiotis, D. K. (2008). A self-organizing algorithm for molecular alignment and pharmacophore development. *Journal of Computational Chemistry*, *29*, 9658.
- Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F., & Mason, J. S. (2007). A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): Theory and application. *Journal of Chemical Information and Modeling*, *47*, 27994.
- Baskin, I. I., Palyulin, V. A., & Zefirov, N. S. (2008). Neural networks in building QSAR models. *Methods in Molecular Biology*, *458*, 137–158.
- Bell, P. H., et al. (2008). Studies in chemotherapy. A theory of the relation of structure to activity of sulfanilamide type compounds. *Journal of the American Chemical Society*, *64*, 2905–2917.
- Chen, J., & Lai, L. H. (2008). Pocket v.2: Further developments on receptor-based pharmacophore modeling. *Journal of Chemical Information and Modeling*, *46*, 268491.
- Cleves, A. E., Johnson, S. R., & Jain, A. N. (2019). Electrostatic-field and surface-shape similarity for virtual screening and pose prediction. *Journal of Computer-Aided Molecular Design*, *33*, 865–886. Available from <https://doi.org/10.1007/s10822-019-00236-6>.
- CoMMA. (2009). IBM Bioinformatics Group. <http://cbcsrv.watson.ibm.com/Tco.html>.
- Datar, P. A., Khedkar, S. A., Malde, A. K., & Coutinho, E. C. (2006). Comparative residue interaction analysis (CoRIA) a 3D-QSAR approach to explore the binding contributions of active site residues with ligands. *Journal of Computer-Aided Molecular Design*, *20*, 343–360.
- Dhaked, D. K., Verma, J., Saran, A., & Coutinho, E. (2009). Exploring the binding of HIV-1 integrase inhibitors by comparative residue interaction analysis (CoRIA). *Journal of Molecular Modeling*, *15*, 233–245.
- Dror, O., Shulman-Peleg, A., Nussinov, R., & Wolfson, H. (2006). Predicting molecular interactions in silico. I. An updated guide to pharmacophore identification and its applications to drug design. *Frontiers in Medicinal Chemistry*, *3*(551), 84.
- Dunteman, G. H. (1989). Basic concepts of principal components analysis. In G. H. Dunteman (Ed.), *Principal components analysis* (pp. 15–22). London: SAGE Publications Ltd.
- EIMchichi, L., Belhassan, A., Lakhlifi, T., & Bouachrine, M. (2020). 3D-QSAR study of the chalcone derivatives as anticancer agents. *Journal of Chemistry*. Available from <https://doi.org/10.1155/2020/5268985>.
- Engkvist, O., Norrby PO, Selmi N, Lam YH, Peng Z, Sherer EC, Amberg W, Erhard T, Smyth LA. (2018). Computational prediction of chemical reactions: current status and outlook. *Drug Discovery Today*, *23*, 1203–1218.
- Ferreira, L. G., Dos Santos, R. N., Oliva, G., & Andricopulo, A. D. (2015). Molecular docking and structure-based drug design strategies. *Molecules (Basel, Switzerland)*, *20*(7), 13384–13421. Available from <https://doi.org/10.3390/molecules200713384>.

- Friesner, R. A., Murphy, R. B., Repasky, M. P., et al. (2006). Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of Medicinal Chemistry*, 49(21), 6177–6196. Available from <https://doi.org/10.1021/jm051256o>.
- Gieleciak, R., & Polanski, J. (2007). Modeling robust QSAR Iterative variable elimination schemes for CoMSA: Application for modeling benzoic acid pKa values. *Journal of Chemical Information and Modeling*, 47, 547–556.
- Gimenez-Oya, V., Villacañas, O., Fernández-Busquets, X., & Rubio-Martinez, J. (2009). Imperial S. Mimicking direct proteinprotein, and solvent mediated interactions in the CDP-methylerythritol kinase homodimer: A pharmacophore-directed virtual screening approach. *Journal of Molecular Modeling*, 15(8), 9971007.
- Gohlke, H., & Klebe, G. (2002). DrugScore meets CoMFA: Adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. *Journal of Medicinal Chemistry*, 45, 4153–4170.
- GRID. (2009). *Molecular Discovery Ltd.* http://www.moldiscovery.com/soft_grid.php.
- Guner, O. F., & Henry, D. R. (2000). Metric for analyzing hit lists and pharmacophores. In O. F. Guner (Ed.), *Pharmacophore perception, development, and use in drug design, IUL biotechnology series* (p. 191212). La Jolla, CA: International University Line.
- Hanser, T., Barber, C., Marchal, & Werner, S. (2016). Applicability domain: Towards a more formal definition. *SAR and QSAR in Environmental Research*, 27, 893–909.
- Hasegawa, K., & Funatsu, K. (2000). Partial least squares modeling and genetic algorithm optimization in quantitative structure-activity relationships. *SAR and QSAR in Environmental Research*, 11(3–4), 189–209. Available from <https://doi.org/10.1080/10629360008033231>.
- Hoonakker, F., Lachiche, N., Varnek, A., & Wagner, A. (2010). *International Journal on Artificial Intelligence Tools*, 20, 253–270.
- Hopfinger, A. J. (1980). A QSAR investigation of dihydrofolate reductase inhibition by Baker triazines based upon molecular shape analysis. *Journal of the American Chemical Society*, 102, 7196–7206.
- Huang, N., Shoichet, B. K., & Irwin, J. J. (2006). Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry*, 49, 6789801.
- Huang, Q., et al. (2010). PhDD: A new pharmacophore-based de novo design method of drug-like molecules combined with assessment of synthetic accessibility. *Journal of Molecular Graphics & Modelling*, 28(8), 77587.
- Jimenez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S., & De Fabritiis, G. (2017). DeepSite: Protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics (Oxford, England)*, 33(19), 3036–3042.
- Jimenez, J., Škalič, M., Martínez-Rosell, G., & De Fabritiis, G. K. (2018). DEEP: Protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks. *Journal of Chemical Information and Modeling*, 58(2), 287–296. Available from <https://doi.org/10.1021/acs.jcim.7b00650>.
- Jojart, B., Martinek, T. A., & Marki, A. (2005). The 3D structure of the binding pocket of the human oxytocin receptor for benzoxazine antagonists, determined by molecular docking, scoring functions and 3D-QSAR methods. *Journal of Computer-Aided Molecular Design*, 19, 341–356.
- Kim, K. H. (2001). Thermodynamic aspects of hydrophobicity and biological QSAR. *Journal of Computer-Aided Molecular Design*, 15, 367–380.
- Koes, D. R., & Camacho, C. J. (2012). ZINCPharmer: Pharmacophore search of the ZINC database. *Nucleic Acids Research*, 40(W1), W409–W414.
- Kovatcheva, A., Golbraikh, A., Oloff, S., Xiao, Y. D., Zheng, W., Wolschann, P., Buchbauer, G., & Tropsha, A. (2004). Combinatorial QSAR of ambergris fragrance compounds. *Journal of Chemical Information and Computer Sciences*, 44, 582–595.
- Kristam, R., Gillet, V. J., Lewis, R. A., & Thorner, D. (2005). Comparison of conformational analysis techniques to generate pharmacophore hypotheses using catalyst. *Journal of Chemical Information and Modeling*, 45(2), 46176.
- Lambrinidis, G., & Tsantili-Kakoulidou, A. (2018). Challenges with multi-objective QSAR in drug discovery. *Expert Opinion on Drug Discovery*, 13(9), 851–859. Available from <https://doi.org/10.1080/17460441.2018.1496079>.
- Landrum, G. (2017). RDKit documentation. *Release*, 2017, 1–125. Available from <http://www.rdkit.org>.
- Lee, J. Y., Krieger, J. M., Li, H., & Bahar, I. (2020). Phammaker: Pharmacophore modeling and hit identification based on druggability simulations. *Protein Science: A Publication of the Protein Society*, 29(1), 76–86. Available from <https://doi.org/10.1002/pro.3732>.
- Leelananda, S. P., & Lindert, S. (2016). Computational methods in drug discovery. *Beilstein Journal of Organic Chemistry*, 12, 2694–2718. Available from <https://doi.org/10.3762/bjoc.12.267>, PMID: 28144341; PMCID: PMC5238551.
- Lo, Y. C., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 8, 1538–1546. Available from <https://doi.org/10.1016/j.drudis.2018.05.010>.
- Mate, G., Hofmann, A., Wenzel, N., & Heermann, D. W. (2014). A topological similarity measure for proteins. *Biochimica et Biophysica Acta (BBA)—Biomembranes*, 1838(4), 1180–1190. Available from <https://doi.org/10.1016/j.bbamem.2013.08.019>.
- Nettles, J. H., et al. (2007). Flexible 3D pharmacophores as descriptors of dynamic biological space. *Journal of Molecular Graphics & Modelling*, 26, 62233.
- Nikolova, N., & Jaworska, J. (2003). Approaches to measure chemical similarity: A review. *QSAR & Combinatorial Science*, 22, 1006–1026. Available from <https://doi.org/10.1002/qsar.200330831>.
- Ortuso, F., Langer, T., & Alcaro, S. G. B. P. M. (2006). GRID based pharmacophore model. Concept and application studies to proteinprotein recognition. *Bioinformatics*, 22(12), 144955.
- Pastor, M., Cruciani, G., & Watson, K. A. (1997). A strategy for the incorporation of water molecules present in a ligand binding site into a three-dimensional quantitative structure–activity relationship analysis. *Journal of Medicinal Chemistry*, 40, 4089–4102.
- Polanski, J., Gieleciak, R., & Bak, A. (2002). The comparative molecular surface analysis (COMSA)—A nongrid 3D QSAR method by a coupled neural network and PLS system: Predicting pK(a) values of benzoic and alkanolic acids. *Journal of Chemical Information and Computer Sciences*, 2002(42), 184–191.

- Polanski, J., & Walczak, B. (2000). The comparative molecular surface analysis (COMSA): A novel tool for molecular design. *Computers & Chemistry*, 24, 615–625.
- Polishchuk. (2017). Interpretation of quantitative structure-activity relationship models: Past, present, and future. *Journal of Chemical Information and Modeling*, 57(11), 2618–2639.
- Poptodorov, K., Luu, T., & Hoffmann, R. D. (2006). Methods, and principles in medicinal chemistry. In T. Langer, & R. D. Hoffmann (Eds.), *Pharmacophores, and pharmacophores searches* (Vol. 2). Weinheim, Germany: Wiley-VCH.
- Price, A. J., Howard, S., & Cons, B. D. (2017). Fragment-based drug discovery and its application to challenging drug targets. *Essays in Biochemistry*, 61(5), 475–484. Available from <https://doi.org/10.1042/EBC20170029>.
- Salam, N. K., Nuti, R., & Sherman, W. (2009). Novel method for generating structure-based pharmacophores using energetic analysis. *Journal of Chemical Information and Modeling*, 49(10), 2356–2368. Available from <https://doi.org/10.1021/ci900212>.
- Sato, T., Honma, T., & Yokoyama, S. (2010). Combining machine learning and pharmacophore-based interaction fingerprint for *in silico* screening. *Journal of Chemical Information and Modeling*, 50(1), 170–185. Available from <https://doi.org/10.1021/ci900382>.
- Schneidman, D. D., Dror, O., Inbar, Y., Nussinov, R., & Wolfson, H. J. (2008). PharmaGist: A webserver for ligand-based pharmacophore detection. *Nucleic Acids Research*, 36, W223–W228. Available from <https://doi.org/10.1093/nar/gkn187>.
- Silverman, B. D. (2000). Three-dimensional moments of molecular property fields. *Journal of Chemical Information and Computer Sciences*, 40, 1470–1476.
- Silverman, B. D., & Platt, D. E. (1996). Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *Journal of Medicinal Chemistry*, 39, 2129–2140.
- Singh, D. B. (Ed.) (2020). *Computer-aided drug design*, Singapore: Singapore. https://doi.org/10.1007/978-981-15-6815-2_7.
- Singh, D. B., & Dwivedi, S. (2016). Docking and molecular dynamics simulation study of inhibitor 2-Fluoroaristeromycin with anti-malarial drug target PfSAHH. *Network Modeling and Analysis in Health Informatics and Bioinformatics*, 5, 16.
- Singh, D. B., & Pathak, R. K. (2020). *Computational approaches in drug designing and their applications. Experimental protocols in biotechnology* (pp. 95–117). New York, NY: Humana.
- Skalic, M., Jiménez, J., Sabbadin, D., & De Fabritiis, G. (2019). Shape-based generative modeling for de novo drug design. *Journal of Chemical Information and Modeling*, 59(3), 1205–1214. Available from <https://doi.org/10.1021/acs.jcim.8b00706>.
- Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W., Jr (2013). Computational methods in drug discovery. *Pharmacological Reviews*, 66(1), 334–395. Available from <https://doi.org/10.1124/pr.112.007336>.
- Smellie, A., Teig, S., & Towbin, P. (1995). Poling: Promoting conformational variation. *Journal of Computational Chemistry*, 16, 17187.
- Sutter, J., Guner, O. F., Hoffman, R., & Waldman, M. (2016). In O. F. Guner (Ed.), *Pharmacophore perception, development, and use in drug design*. La Jolla, CA: International University Line, 2000.
- Sybyl. (2005). *Sybyl version 7.1*. St. Louis, MO: Tripos Associates Inc.
- Totrov, M. (2008). Atomic property fields: Generalized 3D pharmacophoric potential for automated Ligand superposition, pharmacophore elucidation and 3D QSAR. *Chemical Biology & Drug Design*, 71, 1527.
- Tschinke, V., & Cohen, N. (1993). The NEWLEAD program: A new method for the design of candidate structures from pharmacophoric hypotheses. *Journal of Medicinal Chemistry*, 36, 386370.
- Verma, J., Khedkar, V. M., & Coutinho, E. C. (2010). 3D-QSAR in drug design—A review. *Current Topics in Medicinal Chemistry*, 10(1), 95–115. Available from <https://doi.org/10.2174/156802610790232260>.
- Wahi, J., Freyss, J., von Korff, M., et al. (2019). Accuracy evaluation and addition of improved dihedral parameters for the MMFF94s. *Journal of Cheminformatics*, 11, 53. Available from <https://doi.org/10.1186/s13321-019-0371-6>.
- Wang, X., Shen, Y., Wang, S., et al. (2017). PharmMapper update: A web server for potential drug target identification with a comprehensive target pharmacophore database. *Nucleic Acids Research*, 45(W1), W356–W360.
- Wermuth, C. G. (2006). Pharmacophores: Historical perspective and viewpoint from a medicinal chemist. In T. Langer, & R. D. Hoffmann (Eds.), *Pharmacophores and pharmacophore searches*. Weinheim: Wiley-VCH-313.
- Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today*, 11(23–24), 1046–1053.
- Willett, P. (2010). Similarity searching using 2D structural fingerprints. In J. Bajorath (Ed.), *Cheminformatics and computational chemical biology. Methods in molecular biology (methods and protocols)* (Vol. 672). Totowa, NJ: Humana Press. Available from https://doi.org/10.1007/978-1-60761-839-3_5.
- Wolber, G., & Langer, T. (2005). LigandScout: 3-D pharmacophores derived from protein bound ligands and their use as virtual screening filters. *Journal of Chemical Information and Modeling*, 45(1), 1609.
- Worley, B., & Powers, R. (2013). Multivariate analysis in metabolomics. *Current Metabolomics*, 1(1), 92–107. Available from <https://doi.org/10.2174/2213235X11301010092>.