Computational and Structural Biotechnology Journal 23 (2024) 2964–2977

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Review Article

Artificial intelligence in antidiabetic drug discovery: The advances in QSAR and the prediction of α-glucosidase inhibitors

Adeshina I. Odugbemi [a,b,d], Clement Nyirenda [c], Alan Christoffels [a,e], Samuel A. Egieyeh [b,d,*]

[a] South African Medical Research Council Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, Bellville, Cape Town 7535, South Africa
[b] School of Pharmacy, University of the Western Cape, Bellville, Cape Town 7535, South Africa
[c] Department of Computer Science, University of the Western Cape, Cape Town 7535, South Africa
[d] National Institute for Theoretical and Computational Sciences (NITheCS), South Africa
[e] Africa Centres for Disease Control and Prevention, African Union, Addis Ababa, Ethiopia

ARTICLE INFO

Keywords:
QSAR
Molecular descriptors
Machine learning
Deep learning
Diabetes
α-glucosidase

ABSTRACT

Artificial Intelligence is transforming drug discovery, particularly in the hit identification phase of therapeutic compounds. One tool that has been instrumental in this transformation is Quantitative Structure-Activity Relationship (QSAR) analysis. This computer-aided drug design tool uses machine learning to predict the biological activity of new compounds based on the numerical representation of chemical structures against various biological targets. With diabetes mellitus becoming a significant health challenge in recent times, there is intense research interest in modulating antidiabetic drug targets. α-Glucosidase is an antidiabetic target that has gained attention due to its ability to suppress postprandial hyperglycaemia, a key contributor to diabetic complications. This review explored a detailed approach to developing QSAR models, focusing on strategies for generating input variables (molecular descriptors) and computational approaches ranging from classical machine learning algorithms to modern deep learning algorithms. We also highlighted studies that have used these approaches to develop predictive models for α-glucosidase inhibitors to modulate this critical antidiabetic drug target.

1. Introduction

Drug discovery is an important process in biomedical research, by which new candidate medications are discovered or developed. It plays a vital role in addressing various medical challenges such as diabetes, cardiovascular diseases, cancer, and neurodegenerative disorders. The discovery of new drugs enables the development of innovative therapies that manage symptoms, slow disease progression, and potentially cure these conditions.

Drug discovery has historically been marked by exorbitant costs and protracted timelines, particularly during the hit identification phase of therapeutic compounds [1,2]. Fortunately, technological advancements such as computational modelling and computer-aided drug design approaches have made the process more efficient and cost-effective. With the emergence of Artificial Intelligence (AI), drug discovery has been revolutionized by AI-powered algorithms, accelerating the identification of promising drug candidates, and reducing time and costs [3,4]. AI is a vast subject of computer science that seeks to develop intelligent systems capable of doing activities that would normally need human intelligence [5]. These tasks include recognizing patterns, making decisions, and learning from experience. AI encompasses various subfields and techniques, including Machine Learning (ML) and Deep Learning (DL).

One critical tool in the computer-aided drug design toolkit that uses ML in its process is quantitative structure-activity relationship (QSAR) analysis [6,7]. The ML used enables learning and predictions from input data without explicit programming. Essentially, QSAR allows researchers to predict the biological activity of new compounds based on their chemical structure, making it a useful tool in the search for effective compounds to modulate drug targets for disease conditions such as diabetes. Diabetes has become one of the most challenging health problems of this century with a rapid increase in its global prevalence [8–10]. Among the many potential antidiabetic targets for QSAR predictions, α-glucosidase has garnered significant attention due to its ability to effectively suppress postprandial hyperglycaemia, a significant contributor to the progression of diabetic complications [11]. The

* Corresponding author at: School of Pharmacy, University of the Western Cape, Bellville, Cape Town 7535, South Africa.
E-mail address: segieyeh@uwc.ac.za (S.A. Egieyeh).

https://doi.org/10.1016/j.csbj.2024.07.003
Received 16 April 2024; Received in revised form 3 July 2024; Accepted 3 July 2024
Available online 6 July 2024

development of QSAR models to predict inhibitors of α-glucosidase could rebound on the progress towards novel antidiabetic therapies.

To develop QSAR models, researchers use various strategies for generating input variables and computational approaches ranging from the classical machine learning algorithms such as Multiple Linear Regression (MLR), Partial Least Squares (PLS), Decision Trees (DT), Random Forests (RF) and Support Vector Machines (SVM), to modern deep learning algorithms such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Graph Neural Networks (GNNs) [6,12,13]. The excitement and buzz surrounding AI today are notably concentrated on DL [14,15]. Deep Learning is a specialized area within Machine Learning that deals with deep neural networks and has been particularly successful in handling complex tasks involving large datasets. It is important to note that there are several strategies in QSAR to generate the input data (independent variables) for the ML algorithms, which includes 2D-, and 3D-QSAR. These strategies largely depend on the dimensionality of the compounds' molecular descriptors, which are numerical representations of compound structures [16]. In essence, the descriptors are derived from 2D or 3D structural information of chemical compounds. QSAR models have been instrumental in accurately predicting the biological activity of new compounds and identifying promising drug candidates, reducing the time, costs, and risks associated with drug discovery.

Different reviews have discussed the role of QSAR in drug discovery from various perspectives [6,17–22]. For instance, Dudek et al. [18] provided a comprehensive overview of QSAR, focusing on generating molecular descriptors and the methodologies for activity prediction. Wang et al. [22] elaborated on the basic principles of QSAR model development, highlighting the current applications and challenges of QSAR in different fields. Abuhammad et al. [17] specifically examined QSAR research in type II diabetes, detailing the QSAR strategies employed to model molecular bioactivities towards antidiabetic targets.

While these reviews have covered QSAR, they have yet to specifically examine the application in α-glucosidase antidiabetic target, to our knowledge. In the current review, we delve into QSAR strategies in the context of molecular descriptors and the machine learning algorithms that drive model predictions (ranging from classical to modern algorithms), further exploring their application to α-glucosidase predictive models as a case study. We examined studies on the Scopus database from 2011 to 2023 that developed QSAR models for predicting α-glucosidase inhibitors. Additionally, we discussed the strengths and limitations of various classes of molecular descriptors, machine learning algorithms, and prominent QSAR software platforms used by these studies for model building.

2. Quantitative structure-activity relationship

The concept of Quantitative Structure-Activity Relationship (QSAR) was initially introduced by [23]. This pioneering work established numerical substituent constants that represent the specific contributions of distinct molecular fragments to a compound's overall activity. In contemporary drug design, QSAR plays a crucial role, since it can be used to quickly and efficiently screen large numbers of compounds for potential drug candidates as opposed to high throughput in vitro screening which can be expensive to set up and maintain [24]. The foundational principle of QSAR lies in the understanding that a compound's biological activity is closely tied to its molecular structure. This structural information is encoded into molecular descriptors, and the QSAR model establishes mathematical relationships between these descriptors and biological activities.

From a broad perspective, the fundamental premise of QSAR requires the translation of molecular structures into quantitative parameters known as molecular descriptors. The central objective within the QSAR framework is to determine an appropriate mathematical function that connects these quantitative descriptors with experimental activity data. However, the field of QSAR has witnessed innovations from researchers, breaking the boundaries of quantitative structural descriptors to explore diverse qualitative descriptors focused on capturing the presence or absence of structural features (such as structural fingerprints) rather than precise numerical measurements. [25].

The classification of QSAR methodologies is based on how descriptor values and structural representations are derived. This categorization has evolved over time, progressing from the early 1- or 2-dimensional (1D or 2D) linear free energy relationships proposed by Hansch and Free-Wilson [23,26], to Crammer's 3-dimensional (3D) QSAR approach [27]. Subsequent advancements include Hopfinger's extension to the 4th dimension (4D) [28] and Vedani's exploration of the 5th and 6th dimensions (5D and 6D) [29–31]. In the context of antidiabetic QSAR studies, the predominant focus remains on 2D and 3D approaches.

3. Molecular descriptors in QSAR, the independent variable for machine learning

A molecular descriptor represents the outcome of a systematic logical and mathematical process that converts the chemical information ingrained within a symbolic portrayal of a molecule into valuable numerical parameters [32]. These descriptors are essential for characterizing the molecular structure accurately, which is crucial for establishing a meaningful correlation between the structure and its expected properties such as biological activity. Molecular descriptors are the critical, independent variables that train machine learning (ML) algorithms in QSAR model development. They are the foundation for ML algorithms to build and identify potential drug candidates. By transforming complex chemical information into a format that ML algorithms can understand, molecular descriptors enable the algorithms to sift through vast amounts of data, identify patterns, and predict how new compounds might behave. This prediction makes ML-driven drug discovery promising; it allows researchers to pinpoint which molecules are worth investigating further in the lab, thus speeding up the drug discovery process and making it more efficient. Consequently, molecular descriptors are considered the cornerstone of QSAR models, as they provide the core elements necessary for developing reliable models for predicting these expected properties.

While it is widely acknowledged that molecules sharing structural similarities tend to exhibit similar properties, it is important to acknowledge the phenomenon of "activity cliffs." Activity cliffs refer to instances where molecules with closely resembling structures demonstrate vastly divergent biological activities [33]. Although they have the potential to challenge conventional QSAR machine learning modelling assumptions, activity cliffs offer valuable insights into the intricacies of molecular interactions and the limitations of descriptor-based predictions. Despite the presence of some activity cliffs, QSAR models utilizing molecular descriptors remain a powerful tool for identifying trends and guiding lead optimization in drug discovery.

Molecular descriptors are derived by applying principles from several different theories, such as graph theory, quantum-chemistry, information theory and organic chemistry [34]. Together, these theories provide a comprehensive toolkit for developing and using molecular descriptors.

In general, a QSAR model's dimensionality matches that of the molecular descriptors utilized. For example, a 2D QSAR model would employ 2D molecular descriptors to represent the chemical structures, while a 3D QSAR model would use 3D descriptors [35,36].

3.1. 0D and 1D molecular descriptors

0D descriptors are the simplest representations of molecular structures and can be derived solely from the chemical formula of a molecule. These descriptors capture essential bulk properties without delving into details of molecular connectivity or bonding patterns [37]. These descriptors are very simple to compute and interpret but have a low information content and a high degree of degeneration, having equal

values for isomers.

On the other hand, 1D molecular descriptors provide a more detailed glimpse into molecular structures by representing molecules as sets of substructures or fragments. These descriptors go beyond the bulk properties captured by 0D descriptors and focus on encoding specific substructures within a molecule, thus offering insights into specific structural features influencing biological activity. This encoding may be in binary form, indicating the presence or absence of a particular substructure, or in frequency representation, reflecting the rate of occurrence of these substructures [16]. For example, a molecule can be represented by the occurrence of functional groups like hydroxyl (-OH), amino (-NH_2), or carbonyl (-C═O) groups. Also, fragments such as benzene rings, alkyl chains, or heterocycles can be encoded as part of 1D descriptors. Similar to 0D descriptors, 1D descriptors also have high degeneracy with different molecules having the same descriptor values (e.g., isomers).

While 0D and 1D descriptors offer a basic level of molecular representation, their limitations necessitate the use of higher-dimensional descriptors like 2D and 3D, which capture information about atom connectivity and spatial arrangements.

3.2. 2D molecular descriptors

2D molecular descriptors represent an advancement over 1D descriptors by incorporating information about how atoms are connected within a molecule, taking into account the presence and nature of chemical bonds. This representation is often achieved through a molecular graph, where atoms are depicted as vertices and bonds as edges [38]. From this graph representation, various numerical quantifiers of molecular topology, known as topological indices (TIs), are derived in a direct and unambiguous manner. These TIs are often referred to as topological descriptors. They allow capturing information about adjacency, connectivity, branching, cyclicity, and symmetry.

Topological indices can be logically categorized into two main groups: topostructural indices and topochemical indices. Topostructural indices encode information solely about adjacency and through-bond distances between atoms [39]. They essentially capture information about the size, shape, and branching of the molecule but do not consider the chemical nature of the atoms or bonds. Topochemical indices, on the other hand, go beyond just connectivity and incorporate information about the chemical properties of the atoms and bonds, such as electronegativity, hydrophobicity, atomic mass, and the presence of hydrogen bond donors/acceptors. Examples of TIs include Wiener index, Randic index, Connectivity index, Schultz index, Zagreb index, and McGowan volume [34].

In addition to molecular graphs, linear notation systems such as Wiswesser line notation, Simplified Molecular Input Line Entry System (SMILES), and SMiles ARbitrary Target Specification (SMARTS) offer alternative 2D representations of molecules. These notations encode molecular structures in a linear format, providing a concise and standardized way to represent complex molecules [40]. While these linear representations capture connectivity information in a compact form, they are often not used as independent variables for QSAR model development for some reason. For example, SMILES can present ambiguity in their molecular representation, where different SMILES strings can represent the same molecular structure. This ambiguity may lead to inconsistencies in model inputs and potentially affecting the reliability and reproducibility of QSAR-ML models.

In essence, 2D molecular descriptors encompass both the structural connectivity of atoms and certain chemical properties, providing a more detailed representation of molecular structure and reactivity. Overall, 2D molecular descriptors play a crucial role in understanding structure-activity relationships and predicting molecular properties. They offer a comprehensive view of molecular structure by considering both the topological arrangement of atoms and certain chemical properties, facilitating more accurate and interpretable computational analysis.

3.3. 3D Molecular descriptors

3D descriptors are numerical values that quantify various aspects of a molecule's three-dimensional structure. These descriptors provide information about the spatial arrangement of atoms, bonds, and other molecular features in three-dimensional space.

3.3.1. Electronic descriptors

Electronic descriptors encompass valuable information about the electronic properties of a molecule, describing its electronic nature. These descriptors contain information regarding the atomic net and partial charges. Furthermore, electronic descriptors quantify specific electronic properties of a molecule, such as electron density, electronic energies, molecular orbital energies, and electron delocalization. Derivation of these descriptors is often obtained through quantum mechanical calculations or empirical models that consider the electronic structure and behaviour of the molecule [41].

Quantum-chemical descriptors are a subset of electronic descriptors, particularly obtained through quantum mechanical calculations, and take into account the full wavefunction of the molecule [42]. They provide information about the distribution of electrons, electron cloud interactions, and electronic properties relevant to chemical reactivity and physical behaviour. This implies that quantum-chemical descriptors are capable of providing a more comprehensive and accurate depiction of the electronic properties of a molecule when compared to electronic descriptors calculated using empirical and statistical methods [43]. However, the computational expense of quantum-chemical descriptors is considerably higher, which limits their applicability to small molecules.

3.3.2. Geometrical descriptors

Geometrical descriptors take into account the three-dimensional structure of the molecule, focusing not only the positions of the atoms, but also the connections among them. Since a geometrical representation involves the knowledge of the relative positions of the atoms in 3D space, geometrical descriptors usually provide more information and discrimination power for similar molecular structures and molecule conformations than topological descriptors [38]. Despite their high information content, geometrical descriptors usually show some drawbacks. They require geometry optimization and, therefore, the cost to calculate them. Moreover, for flexible molecules, several molecule conformations can be available; on one hand, new information is available and can be exploited, but, on the other hand, the problem complexity can significantly increase [34]. Two of the most known classes of three-dimensional descriptors are the Weighted Holistic Invariant Molecular (WHIM) descriptors and the Geometry, Topology, and Atom-Weights Assembly (GETAWAY) descriptors.

WHIM descriptors encode three-dimensional information on size, shape, symmetry, and atomic property distribution. They do not take into account the connections among atoms but only their position in the three-dimensional space [44]. GETAWAY descriptors encode information about the influence that each atom has in determining the whole shape in the molecule and evaluate the interactions among atoms with respect to their geometrical position in the three-dimensional space. For their calculation molecular matrix, representations are considered [45].

3.3.3. Alignment dependency in 3D descriptors

There are two common broad approaches to calculating 3D-descriptors: alignment-dependent and alignment-independent methods.

3.3.3.1. Alignment-dependent method. This method places a critical focus on molecular alignment prior to the calculation of 3D descriptors. Prominent examples of this approach are Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA). In CoMFA, all aligned ligands are placed in an energy

grid, and at each lattice point, a charged probe atom is positioned to calculate energy. The energy calculated corresponds to electrostatic (Coulombic) and steric (Van der Waals) potentials [46,47]. These values serve as descriptors for further analysis. CoMSIA is similar to CoMFA, but unlike CoMFA, CoMSIA calculates the similarity indices of a molecule at a regular grid of points in space. The similarity indices are a measure of how similar the molecule is to a reference molecule [48]. Oftentimes, the alignment-dependent approach has many limitations as it may be time consuming; this can introduce user biasness and it may affect the sensitivity of the resultant model. Moreover, this approach may be impracticable to compound set with diverse scaffolds.

3.3.3.2. Alignment-independent method. This approach is invariant to molecule rotation and translation in space, and thus, requires no superposition of compounds [49]. Here, geometrical and electronic descriptors are calculated without focusing on the alignment of molecules.

3.4. Higher dimensional QSAR

Ligands, in many cases, are not static and can have considerable conformational flexibility, and this can affect the performance of 3D-QSAR models, which rely on rigid representations. In a way to overcome the limitations of 3D descriptors and associated 3D QSAR models, the 4D QSAR methodology was developed [28]. It is designed to account for the variations in the conformational space of the ligands by using ensemble sampling. As a ligand can have many conformers, such detail in chemical representation and simulation requires a lot more computational power than traditional 2D and 3D QSAR. A step further beyond 4D QSAR is the 5D QSAR, which takes into account the dynamic changes in the molecular structure of targets upon ligand binding in an "induced fit" concept [30,31]. Unlike rigid docking models, which assume a fixed structure for both ligands and targets, induced fit considers the flexibility and adaptability of molecules during the binding process. 6D QSAR goes even further by adding an extra dimension that captures the influence of solvent environments on molecular interactions [29].

While 4D to 6D descriptors in QSAR modelling may provide better results than the 2D and 3D approaches, their high computational cost makes them less suitable for high-throughput virtual screening.

3.5. Fingerprint descriptors

Molecular fingerprints are binary or integer arrays that represent the presence or absence of certain chemical features or substructures within a molecule. These features can range from simple binary bits corresponding to the presence or absence of specific atoms or bonds to more complex representations capturing molecular topology [50,51].

The goal of generating molecular fingerprints is to produce a compact representation that retains important structural information while minimizing the computational burden. This is particularly important when dealing with large databases of compounds or conducting high-throughput virtual screening experiments.

Some of the most used fingerprint types are described as follows.

3.5.1. Structural key fingerprints

Structural key fingerprints are binary fingerprints that encode the presence or absence of specific substructures within a molecule which are predefined. They typically represent the presence of certain chemical fragments or functional groups, such as hydroxyl groups, amines, or aromatic rings [51,52]. PubChem fingerprints, and Molecular Access System (MACCS) keys are popular examples of structural key fingerprints.

3.5.2. Topological fingerprints

Topological fingerprints, also known as path-based fingerprints, encode the connectivity information of a molecule in a binary or integer format. They often represent the occurrence of specific atom types or bond types at defined distances from each other. Topological fingerprints are sensitive to molecular size and symmetry. Topological Torsion (TT) and Daylight fingerprints are well-known examples of topological fingerprints [53].

3.5.3. Circular fingerprints

Circular fingerprints represent a distinct type of topological fingerprints, which are characterized by their use of hashed topological fingerprints [54]. In contrast to conventional topological fingerprints, which rely on path analysis within a molecule, circular fingerprints utilize a record of the molecular environment surrounding each atom within a specified radius. Consequently, these fingerprints cannot be employed for substructure inquiries, as identical fragments may possess distinct environments [52]. Nonetheless, they are extensively utilized in full structure similarity searches. Extended-Connectivity Fingerprints (ECFP), Functional-Class Fingerprints (FCFP), and Molprint2D are prominent examples of circular fingerprints.

3.5.4. Pharmacophore fingerprints

Pharmacophore fingerprints are designed to encode the spatial arrangement of specific chemical features essential for molecular recognition, such as hydrogen bond donors/acceptors, hydrophobic regions, and aromatic rings, that are necessary for a molecule to interact with a biological target [55].

3.5.5. Protein–ligand interaction fingerprints (PLIF)

Protein-ligand interaction fingerprints are used to represent intramolecular interactions by analyzing and extracting the binding patterns, mainly non-covalent interactions, between receptors and ligands. Such fingerprints can use information regarding molecular docking or structure-based experimental data to convert 3D protein–ligand interactions into 1D bitstrings, which are subsequently used to train machine learning models [56].

The summary of the strengths and challenges associated with each type of descriptor, progressing from lower- to higher-dimensional space, is presented in Table 1.

4. Machine learning algorithms in QSAR

Machine learning algorithms are essential for developing predictive models to accurately predict outcomes by analysing data. In QSAR, these algorithms use various molecular descriptors of compounds as independent variables to train the ML model. The molecular descriptors can be obtained from either 2D or 3D structural information. The bioactivities derived from experiments are considered the dependent variables in model development. By leveraging ML, researchers can create models that accurately predict the bioactivity of a given compound [57,58]. We can subjectively categorize ML algorithms used in QSAR into two categories: Classical and Modern ML algorithms. While several examples fall into these two categories, we have only discussed a few algorithms commonly used in QSAR model development.

4.1. Classical machine learning algorithms

Classical ML encompasses well-established techniques and algorithms, such as multiple linear regression, partial least squares, decision trees, random forest and support vector machines, often requiring manual feature engineering. It tends to excel in scenarios with interpretable models, smaller datasets, and a reliance on domain expertise for feature selection.

4.1.1. Linear models

Linear models are one of the fundamental building blocks of ML. They create a function that fits a line to the training data. This line tries to capture the relationship between the independent and the dependent

Table 1
Summary of descriptor categories in QSAR.

Descriptor Type	Description	Strengths	Challenges
0D	Represent simplest molecular properties, derived solely from chemical formulas	Easy to compute, simple interpretation	Low information content, high degree of degeneracy, identical values for isomers
1D	Represent molecules as sets of substructures or fragments	More detailed than 0D, captures specific substructures influencing activity	Still limited in complexity, high degeneracy, poor differentiation of isomers
2D	Incorporate information on how atoms are connected within a molecule, often using molecular graphs	Captures connectivity and topological information	Cannot capture 3D spatial information, some ambiguity in linear notation systems like SMILES
3D	Quantify spatial arrangement of atoms and bonds, including electronic and geometrical descriptors	Provides detailed spatial and electronic information, important for modeling 3D interactions	Computationally intensive, requires geometry optimization, handling multiple conformations can increase complexity
Higher-dimensional Descriptors (4D, 5D, 6D)	Account for conformational flexibility, dynamic changes in molecular structures, and solvent effects	Captures dynamic and environmental influences, potentially more accurate predictions	Extremely high computational cost, increased complexity, less practical for high-throughput screening

variables. Several examples of linear models exist, but we only describe a few common to QSAR model development, such as multiple linear regression, Partial Least Squares, and Binary Logistic Regression.

4.1.1.1. Multiple linear regression. Multiple Linear Regression (MLR) is a statistical technique that helps establish a linear relationship between a dependent variable such as biological activity and two or more independent variables like molecular descriptors. It is an extension of simple linear regression, which has only one independent variable, and it is useful in situations where multiple factors may influence the outcome. MLR is a foundational and interpretable technique in QSAR modelling [59]. However, it is important to note that MLR is a linear model, which means it can only capture linear relationships between the dependent and independent variables. If the relationship between the dependent and independent variables is non-linear, then MLR may not be the most suitable method for developing a QSAR model.

4.1.1.2. Partial least squares. Partial Least Squares (PLS) is a widely used multivariate statistical technique in QSAR studies. It is suitable for regression analysis of high-dimensional data, that is, data with a high number of independent variables [60]. PLS works by discovering a set of latent variables that are linear combinations of the original descriptors. Latent variables are new variables generated by combining the original independent variables in a way that captures the most critical information for predicting the dependent variable [61]. PLS iteratively finds pairs of latent variables. The first latent variable is determined by finding the linear combination of independent variables that has the highest covariance with the dependent variable. The second latent variable is found by identifying the linear combination of independent variables that has the highest covariance with the residual from the first latent variable. This process continues until the desired number of latent variables is reached [62,63]. Compared to other methods, PLS is less sensitive to outliers because it does not give equal weight to all independent variables [63]. Instead, PLS assigns more weight to the independent variables most correlated with the dependent variable. However, the latent variables produced by PLS may not always provide a straightforward interpretation [64].

4.1.1.3. Binary logistic regression. Binary Logistic Regression (BLR) constructs a model that maps the features of compounds to the probability of their biological activity, typically encoded as binary outcomes (e.g., active or inactive). The algorithm employs the logistic function, also known as the sigmoid function, to ensure that the predicted probabilities lie between 0 and 1, making it suitable for binary classification tasks. BLR learns the optimal coefficients that minimize the discrepancy between predicted probabilities and actual class labels in the training data [65]. One of the primary advantages of BLR in QSAR is its interpretability, as it provides insights into the impact of individual features on the probability of influencing biological activity. Additionally, it is computationally efficient, making it suitable for analyzing large datasets. Despite its strengths, BLR assumes a linear relationship between features and outcomes, which may limit its performance when dealing with complex, nonlinear relationships.

4.1.2. Tree-based models

Tree-based ML models are a powerful and popular technique used in various machine learning tasks. These models work by using a series of decision trees to arrive at a final prediction.

4.1.2.1. Decision trees. Decision trees (DT) are a type of machine learning algorithm that is used to solve classification and regression problems. In DT, data is represented as a tree-like structure, where each node represents a decision, and each branch represents a possible outcome of that decision. The algorithm recursively partitions data into subsets based on the values of specific features, selecting the best feature and corresponding threshold at each step to maximize a chosen splitting criterion [66]. This criterion could be Gini impurity for classification or mean squared error for regression [67]. The tree structure continues to grow until stopping conditions are met, such as a maximum depth, a minimum number of samples at a leaf node, or a specific impurity level. The decision tree then serves as a predictive model, guiding new data along a path from the root node to a leaf node, where a final prediction is made.

DT are useful in QSAR because they can handle both quantitative and categorical data, uncover non-linear relationships between molecular descriptors and activities, and are easy to interpret [68]. However, to prevent overfitting, they require careful handling, and complex applications may benefit from pruning or ensemble methods. Overfitting is an undesirable behavior in machine learning where the model performs well on training data but not new data. DT in many cases may not perform optimally, so strategies like class weighting or ensemble methods may be needed [69]. Additionally, single DT can have high variance on test set, meaning they may not generalize well to new, unseen data. This is where ensemble methods, like Random Forests and Gradient Boosting, can be beneficial.

4.1.2.2. Random forests. Random Forests (RF) are a widely used and powerful ensemble learning method in the field of QSAR. They are designed to improve predictive accuracy and reduce overfitting by combining multiple DT [70,71]. RF use a process called bootstrapping to randomly sample the data with replacement, creating multiple datasets. During the construction of each DT, a random subset of descriptors is selected at each node, creating diversity. By combining the predictions of each individual decision tree through majority voting (for classification tasks) or averaging (for regression tasks), RF are able to produce robust and accurate predictions. They excel in handling complex, high-dimensional data and reducing overfitting [72]. However, their ensemble nature makes them less interpretable than individual DT, as

the overall prediction is an aggregate of many trees.

4.1.3. Support vector machines

Support Vector Machines (SVM) are a versatile ML algorithm that can be used for both classification and regression problems. SVM work by transforming the feature space to find a hyperplane that maximizes the margin between classes or regression targets [73,74]. A hyperplane is a subspace of one dimension less than the ambient space. In simpler terms, if a space is 3-dimensional, then its hyperplanes are the 2-dimensional planes, while if the space is 2-dimensional, its hyperplanes are the 1-dimensional lines. SVM find an optimal hyperplane, which acts as the decision boundary, that separates the data into classes with the largest possible margin. The margin is the distance between the hyperplane and the closest points in each class, known as support vectors. SVM can be categorized as linear or nonlinear models in their operation. While simple linear SVMs operate with a linear decision boundary in the feature space, nonlinear SVM can learn more complex decision boundaries by implicitly mapping the data into a higher-dimensional space, where it is easier to find a hyperplane that separates the data along the line of biological activity. The mapping uses kernel functions such as polynomial, radial basis function (RBF), or sigmoid kernels [75,76]. SVM are powerful tools for handling high-dimensional data, and they can capture non-linear relationships between descriptors and biological activity [77,78]. However, their application can be computationally intensive, and the choice of kernel and tuning parameters requires careful consideration [79].

4.1.4. K- Nearest neighbours

K-Nearest Neighbours (kNN) identifies the similarity between the molecular structures of chemical compounds through their descriptors. Each compound's descriptors serve as features, encapsulating various aspects of its molecular makeup. During the training phase, kNN learns the instances of the descriptors of each compound in the dataset and the corresponding biological activities or properties. When a new compound is predicted, kNN calculates the similarity between its descriptors and those of the compounds in the training set using distance metrics such as Euclidean distance [80,81]. The "k" nearest neighbours to the new compound are identified, and their associated labels are used for prediction. The major advantages of kNN in QSAR lie in its simplicity and interpretability. kNN does not require complex assumptions about the underlying data distribution, and by analysing the properties of the nearest neighbours, kNN can provide insights into the structural features influencing biological activity [80]. However, as molecular descriptors increase, the distance metric in the high-dimensional space becomes less meaningful, leading to inaccurate predictions. Moreover, the value of k (number of neighbours) significantly impacts the model's performance. Choosing a small k can lead to overfitting, while a very high k can result in underfitting [82].

4.1.5. Bayesian network

Bayesian Network (BN) works by probabilistically representing the relationships between chemical features and biological activities. The algorithm starts by constructing a directed acyclic graph where nodes represent chemical features and biological activities, and edges represent conditional dependencies between them [83]. The conditional probability distributions associated with each node are learned from the available data through statistical methods or expert knowledge [84]. One key advantage of BN in QSAR modelling is their ability to handle uncertainty effectively. QSAR data often contain noise, missing values, or measurement errors, and BN can explicitly model and propagate uncertainty through the network [85,86], providing more reliable predictions and uncertainty estimates. Additionally, BN offers interpretability through graphical representations of the dependencies between variables [87], which may allow researchers to understand the underlying mechanisms driving compound activity. However, constructing an accurate BN requires sufficient data to estimate the conditional probability distributions. When data is limited, BN may suffer from overfitting or underfitting issues. Moreover, the computational complexity of BN can be high, especially for large datasets with many variables.

4.1.6. Multi-layer perceptrons

Multi-Layer Perceptrons (MLPs) are a type of artificial neural networks (ANNs) that consist of an input layer (matching the number of descriptors), one or more hidden layers with interconnected nodes (neurons), and an output layer (predicting the activity). MLPs function by adjusting weights and biases associated with connections between neurons to minimize the error between the predicted and actual activity [88]. This learning process makes MLPs excel at finding complex, non-linear relationships between molecular descriptors and biological activity. However, MLPs come with common challenges such as overfitting, especially with a high number of neurons and small datasets.

4.2. Modern machine learning algorithms

Modern Machine Learning primarily revolves around deep learning (DL), which is a subfield of ML that focuses on learning complex patterns from data by using ANNs. DL algorithms have been shown in QSAR research to produce models that are more robust and accurate models when compared to the classical algorithms [89,90]. While MLPs are foundational ANNs, they wouldn't be considered the forefront of deep learning algorithms today. Here, we looked at the fundamental principles of the more complex DL algorithms common in drug discovery such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Graph Neural Networks (GNNs), and their advantages in QSAR modelling.

4.2.1. Convolutional neural networks

CNNs are a class of DL models primarily designed for processing grid-like data, such as pixels of data in image processing. However, it has found application in QSAR using fixed molecular fingerprints as input data [91–93]. CNNs use a multi-layered architecture that includes convolutional layers, pooling layers, and fully connected layers. Convolutional layers are used to extract features from the input data. Each convolutional layer contains a set of filters, which are applied to the input data to extract different features. The output of a convolutional layer is a feature map, which is a matrix of values that represent the extracted features. These values are the degree of confidence that the CNNs algorithm has detected a feature. Pooling layers are used to reduce the dimensionality of the feature maps and to make the model more robust to noise [94]. Fully connected layers, similar to the hidden layers in traditional neural networks, are used to combine the extracted features and to make the final prediction [95,96]. During training, CNNs adjust their parameters through backpropagation and optimization algorithms to minimize a chosen loss function [97]. The power of CNNs lies in their ability to automatically learn and extract relevant features from the data, making them suitable for predicting biological activities based on molecular structures. However, CNNs require large amounts of data to train effectively and can be computationally demanding. Additionally, their interpretability can be challenging compared to simpler models. Despite these challenges, CNNs' powerful feature extraction capabilities make them valuable drug discovery tools.

4.2.2. Recurrent neural networks

RNNs are a class of ANNs, designed for handling sequential data. They operate by maintaining a hidden state that evolves with each input element in a sequence, allowing them to capture dependencies across time steps [98]. The hidden state represents the information that the RNNs have learned about the sequential data up to a certain time. At each time step, the RNNs processes an input and updates its hidden state based on the current input and the previous hidden state. Each input corresponds to a specific position in the sequence representing the

molecular structure of a compound, where each element in the sequence corresponds to an atom or a molecular substructure. The RNNs can produce an output at each time step, and it can also provide an output at the end of the sequence. These outputs can be used for predicting the biological activity of compounds. Simplified molecular-input line-entry system (SMILES) strings are the commonly used molecular representation used for RNNs [99,100]. They are easy to parse and process. However, there are some drawbacks to using SMILES strings for RNN which includes their sensitivity to the order of atoms, meaning that two SMILES strings that represent the same molecule can be different if the atoms are in a different order. Canonicalization of SMILES may be needed to mitigate against this drawback, standardising the order of atoms [101,102]. Moreover, SMILES strings do not capture the structural information of molecules, such as the bond lengths and angles between atoms. Another molecular representation that can be used for RNNs is molecular graphs [103,104]. A major advantage of molecular graph representations is that they carry more structural information. However, they require a large amount of storage space and significant memory during computation. RNNs are particularly well-suited for QSAR considering their ability to learn sequential data, such as the order of atoms in a molecule, which can be crucial for determining biological activity. The long-term dependency of RNNs makes them difficult to train, particularly because of the gradient explosion or vanishing problem [105,106]. The long-term dependency of RNNs refers to the ability of RNNs to learn long-range relationships in sequential data. This is in contrast to short-range relationships learning of feedforward neural networks that forms the basis of RNNs. RNNs architectures such as long short-term memory (LSTM) and gated recurrent unit (GRU) have been developed to address these issues and enable RNNs to capture long-term dependencies more effectively [107–109].

4.2.3. Graph neural networks (GNNs)

GNNs are a type of deep learning algorithm that is particularly well-suited for QSAR modelling because they can learn to extract features from molecular graphs [110,111]. Molecular graphs are a way of representing molecules as nodes and edges. Nodes represent atoms, and edges represent bonds between atoms. GNNs work by passing messages between the nodes of a molecular graph [112]. The messages are passed over a number of rounds, and at each round, the nodes update their states based on the messages they receive from their neighbours. These messages refer to the information that is exchanged between the nodes of the graph which can include the features of the nodes and the relationships between the nodes [113]. The messages are encoded in a vector representation for each node and can then be used to predict bioactivity of compounds. The messages can be passed in different ways, but the two major approaches to GNNs are recurrent GNNs and convoluted GNNs. In the recurrent GNNs, node representation is learned via some recurrent neural architectures, such as the GRU. On the other hand, convoluted GNNs work by applying a convolution operation to the nodes of a graph, aggregating the features of a node and its neighbours to produce a new feature vector for the node [114]. One of the key advantages of using GNNs for QSAR is that they do not require the use of handcrafted molecular descriptors, which may not capture all of the relevant features of a molecule [110]. By directly working with molecular graphs, where atoms are nodes and bonds are edges, GNNs can learn complex features that capture the relationships and interactions within a molecule. This capability allows GNNs to model intricate molecular structures more effectively than traditional handcrafted descriptors. However, GNNs require substantial computational power and large datasets for training, and their interpretability can be challenging due to the complexity of the learned features. While GNNs offer some interpretability through attention mechanisms that highlight informative parts of the graph, understanding the inner workings of the message passing process can be challenging [115].

The effectiveness of each machine learning algorithm in QSAR and drug discovery largely depends on the size and complexity of the dataset and the target activity. Simple algorithms like Multiple Linear Regression and Partial Least Squares work well with smaller datasets and linear relationships, giving clear and easy-to-understand results. On the other hand, complex algorithms like Random Forests and Support Vector Machines excel with large, detailed datasets, capturing complicated, non-linear patterns. Deep learning models like Convolutional Neural Networks and Graph Neural Networks also thrive on large, complex datasets. However, they require even more data and computational power to avoid overfitting and perform well with new data. Choosing the right algorithm means balancing the data's size and complexity with the specific needs of the target activity to make accurate predictions and support effective drug discovery.

A summary of the strengths and limitations of the algorithms discussed is presented in Table 2.

5. General approach to build QSAR machine learning models

A systematic approach is necessary to construct reliable QSAR models, including data collection, descriptor generation, preprocessing, feature selection, and model evaluation. This section provides an overview of the general approach to the workflow involved in building a QSAR model and a summary chart is presented in Fig. 1.

Data Collection and Cleaning: The first step in building a QSAR model is to gather relevant data from various sources, such as scientific literature, databases, or experimental studies. It's crucial to ensure that the data is of high quality and covers a diverse range of chemical compounds

Table 2
Strengths and limitations of the machine learning algorithms.

Machine Learning Algorithms	Strengths	Limitations
Multiple Linear Regression (MLR)	Simple and easy to interpret, effective for linear relationships	Cannot capture non-linear relationships
Partial Least Squares (PLS)	Handles high-dimensional data, manages multicollinearity well	Latent variables are complex and less interpretable
Binary Logistic Regression (BLR)	Interpretable, computationally efficient for large datasets	Assumes linear relationships, may struggle with complex non-linear relationships
Decision Trees (DT)	It can model non-linear relationships, easy to interpret	Prone to overfitting
Random Forests (RF)	Reduces overfitting through ensemble learning, handles complex data well	Less interpretable than single decision trees, computationally intensive
Support Vector Machines (SVM)	Effective for high-dimensional data, captures non-linear relationships	Computationally intensive, requires careful tuning of hyperparameters
K-Nearest Neighbors (kNN)	Simple to implement and interpret, no assumptions about data distribution	Poor performance with high-dimensional data, computationally expensive, sensitive to the value of k (neighbors)
Bayesian Networks (BN)	Handles uncertainty well, captures complex dependencies between molecular features	High computational complexity
Multi-Layer Perceptrons (MLP)	Can model complex non-linear relationships, flexible architecture	Prone to overfitting, parameters tuning requires expertise and can be time-consuming
Convolutional Neural Networks (CNNs)	Automatically learns features	Requires large datasets and significant computational power, difficult to interpret
Recurrent Neural Networks (RNNs)	Suitable for sequential data	Vanishing/exploding gradients, requires large datasets and computational power
Graph Neural Networks (GNNs)	Effectively handles molecular graphs, learns complex molecular features	Requires substantial data and computational resources, complex to interpret

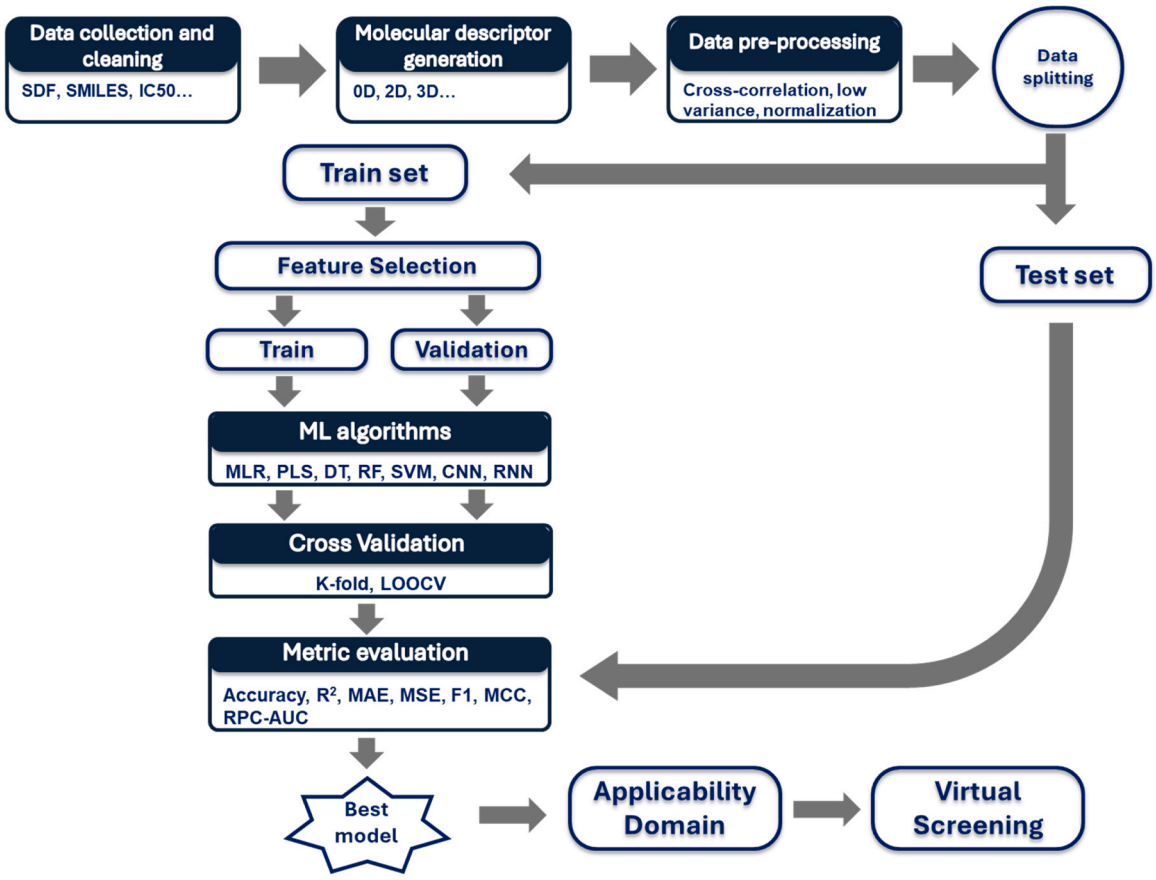

Fig. 1. General workflow for QSAR model development.

and their corresponding activities or properties. Data cleansing involves removing duplicates, handling missing values, and standardizing the format to make it suitable for analysis.

Descriptor Generation: Molecules are complex entities, and their structures must be translated into a numerical format suitable for machine learning algorithms to learn. This translation is achieved through molecular descriptors, as described in Section 3. Depending on the depth of information captured, these descriptors can be 0D, 1D, 2D, or 3D. Several software tools are available to generate descriptors, including PaDEL, RDKit, BlueDesc, Dragon, Mordred, CDK and ChemoPy, among many others.

Data Preprocessing: Preprocessing the data to improve model performance and efficiency is essential before training the model on the generated descriptor data. Data preprocessing approaches include handling issues such as cross-correlation between descriptors, filtering out descriptors with low variance, and normalizing or scaling the data to ensure all features have similar magnitudes.

Data Splitting: To develop a QSAR model, it is a standard practice to divide the dataset into training, validation, and test sets. The training set is used to train the model, the validation set is used to assess model performance during training, and the test set is used to evaluate the model's generalizability and predictive performance on new, unseen data. In many cases, 70 % to 80 % of the entire dataset is used as the training set, while 20 % to 30 % is shared between the validation and test set.

Feature Selection: Feature selection aims to identify the most informative descriptors contributing to the model. Selecting the most informative descriptors can significantly improve model performance and interpretability. Moreover, effective feature selection mitigates the risk of overfitting by focusing on the most informative features while discarding redundant or noisy ones [116]. Depending on the dataset characteristics and modelling objectives, various techniques can be employed for feature selection. The filter and wrapper methods are feature selection strategies often used in QSAR [117]. Filter methods, such as variance thresholding, enable the removal of descriptors with low variance, which are unlikely to capture meaningful variations in the data. Correlation analysis is another example, often used to identify descriptors that exhibit high pairwise correlations. One of the highly correlated pairs is removed to avoid multicollinearity, which may adversely affect model stability. The filter methods are often already captured in the data preprocessing stage. Wrapper methods involve the machine learning algorithm itself in the selection process. Features are added or removed from a subset, and the model's performance is evaluated on each iteration. Prominent examples of wrapper methods include forward selection, backward elimination, and recursive feature elimination [118].

Machine Learning Algorithms: Once the data is prepared and features are selected, various machine learning algorithms can be applied to build QSAR models. These algorithms can include multiple linear regression, partial least squares, decision trees, random forests, support vector machines, and neural networks. The choice of algorithm depends on the nature of the data, the complexity of the relationship between descriptors and activities, and the computational resources available.

Cross Validation and Model Evaluation Metrics: Cross-validation is a technique used to assess the model's generalisation performance and ensure its robustness to variations in the training data. Standard cross-validation methods include k-fold cross-validation and leave-one-out cross-validation (LOOCV). K-fold cross-validation involves dividing the training data into k subsets, training the model on k-1 of these subsets, and assessing the model's performance on the one subset left out. This process is iterated k times, with each subset serving as the validation set once in the entire process. The final performance metric is computed by

averaging the results across all iterations. LOOCV, on the other hand, is a variant of k-fold cross-validation where k equals the total number of data points, meaning each data point serves as a separate validation set in each iteration [119,120]. The performance of the trained model is evaluated using appropriate metrics depending on the specific modelling task and the nature of the predicted activity (e.g., classification or regression). If the model trained tackles a regression problem, metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), R-squared and Q-squared score are often used to assess performance. If, on the other hand, a classification model is involved, metrics such as accuracy, precision, recall, F1 score, Matthews correlation coefficient (MCC), and area under the receiver operating characteristic curve (ROC-AUC) are used, depending on the nature of the prediction task. It is crucial to assess model performance on test sets using these metrics. It is also vital to note that no single metric is universally perfect for all ML tasks, and the choice of metric should be considered based on the specific problem. Moreover, reporting multiple metrics to provide a more comprehensive understanding of the model's performance and potential limitations is often beneficial.

Applicability Domain: The applicability domain defines the chemical space within which the QSAR model's predictions are reliable. It is essential to establish the boundaries of the applicability domain to ensure that the model is not extrapolating beyond its scope [119,121]. The applicability domain can be defined using various methods, with distance-based approaches being particularly common in QSAR. These methods evaluate the similarity between new molecules and those in the training set based on their descriptors. Molecules that surpass a predetermined similarity threshold may fall outside the model's domain [121]. Common similarity measures employed include Euclidean distance and the Tanimoto coefficient, chosen depending on the descriptor characteristics. Leverage analysis is another distance-based method that identifies influential data points within the training set. This analysis, determined using the hat matrix, highlights molecules that disproportionately affect the QSAR model's predictions [122]. High leverage suggests that a molecule resides in a sparsely populated region of the descriptor space, potentially indicating challenges with extrapolation. Establishing the applicability domain is crucial for ensuring QSAR model robustness and validity, thereby enhancing their utility and reliability.

Virtual Screening: Once the QSAR model is trained, validated, and has its applicability domain defined, it can be applied as a virtual screening tool to screen chemical compounds and prioritize candidates for further experimental validation. Virtual screening involves predicting the activity of a large library of compounds and selecting those with the highest predicted potency of the desired activity for further investigation.

Robust QSAR models can be developed following the above-mentioned steps, facilitating drug discovery and molecular property assessments.

The next section highlights studies that have engaged the various input strategies and computational approaches to develop QSAR models for α-Glucosidase inhibitors as a case study.

6. QSAR studies on α-glucosidase inhibitors

Many studies have applied QSAR machine learning to predict modulators of other antidiabetic targets, such as sodium-glucose cotransporter [123,124], dipeptidyl peptidase-4 [125], peroxisome proliferator-activated receptor gamma [126], glycogen synthase kinase 3β [127], and protein tyrosine phosphatase 1B [128]. However, for this review, we have explicitly focused on studies related to α-glucosidase inhibitors. For our literature survey, we conducted a comprehensive search on the Scopus database using the following criteria: (("QSAR" OR "Quantitative Structure-Activity Relationship") AND ("alpha-glucosidase" OR "α-glucosidase")), limited to research articles, and publications between 2011 and 2023. The articles' keywords filter used included QSAR, α-glucosidase, alpha-glucosidase, and Quantitative Structure-Activity Relationship. We considered research articles that we could access, or are open-access.

These studies, employing various QSAR approaches, have contributed to a comprehensive understanding QSAR model application for the prediction of α-glucosidase inhibitors. Some of these studies are discussed in the following paragraphs, and a summary of the QSAR studies is presented.

Wu et al. [129] presented a 2D-QSAR model based on triterpene analogues of ursolic acid, showcasing R^2 values of 0.9986 for training and 0.9996 for the test set. They used PLS as their algorithm of choice for the model building. A simple approach by Dinparast et al. [130] built a QSAR model built with MLR, from molecular weight and density descriptors for benzimidazole derivatives. Their predictions yielded R^2 and Q^2 values of 0.60 and 0.69, respectively. Mora et al. [131] presented a 3D-QSAR analysis on cinnamic acids using the MLR algorithm. They achieved predictive values of R^2 to be 0.901 and Q^2, 0.946.

Using DT, BLR, MLP, k-NN, and SVM, Diéguez-Santana et al. in 2017 [132] and 2019 [133] developed QSAR models for classifying ChEMBL compounds targeting α-amylase and α-glucosidase inhibition. The resulting QSAR models demonstrated an accuracy score of 82.66 % and 84.47 % for the training and test sets, for the 2019 study which utilized DT algorithm. Among the many algorithms used in the 2017 study, they reported k-NN as their top performing algorithm.

In another approach, Joshi et al. [134] used PLS in their QSAR method to predict thiazole-based α-glucosidase inhibitors, achieving cross-validated correlation coefficient Q^2 of 0.800 and non-cross-validated correlation coefficient R^2 of 0.943. Izadpanah et al. [135] contributed a simple mono-descriptor QSAR model for hydrazinyl thiazole-based pyridine derivatives, exhibiting statistical performance with $R^2 = 0.888$ and $Q^2 = 0.872$. They used both MLR and SVM for their model training. Conversely, Halim et al. [136] developed a 2D-QSAR model encompassing a diverse array of 36 compounds using PLS, and yielding $R^2 = 0.88$ and $Q^2 = 0.71$. Dahmani et al. [137] delved into quantum chemistry, constructing a QSAR model to correlate quantum chemical descriptors of 1,2,4-triazolone derivatives with their α-glucosidase inhibitory activity. The constructed model achieved a prediction R^2 of 0.645 using MLR.

Sainy et al. [138] performed 3D QSAR analysis, employing comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) on a series of flavones. MLR analysis was performed using inhibitory activity as the dependent variable and descriptors as the predictor variable. Notably, the derived CoMFA and CoMSIA models displayed cross-validated Q^2 values of 0.742 and 0.759, respectively. Similarly, Liu et al. [139] undertook a similar approach, developing CoMFA and CoMSIA models for myricetin-derived flavonols. However, PLS was the algorithm of choice for training here. Their CoMFA model, characterized by an R^2 value of 1.000 and a Q^2 of 0.598, and the CoMSIA model, featuring an R^2 of 0.998 and Q^2 of 0.730, underscored the effectiveness of these models in predictive analysis.

The contribution from Jia et al. [140] involved the development of 3D-QSAR models through CoMFA and CoMSIA methodologies, applied to the prediction of α-glucosidase inhibitors from dietary flavonoids using PLS. The resulting CoMFA model reveals a non-cross-validated coefficient R^2 of 0.996 and cross-validated correlation coefficient Q^2 of 0.529, and the CoMSIA model achieving an R^2 of 0.997 and Q^2 of 0.515.

An unconventional and noteworthy approach to constructing a 2D-QSAR classifier model for α-glucosidase was undertaken by Diéguez-Santana et al. in 2017 [132] and 2019 [133]. The study's consideration of building models from a large and diverse dataset suggests their QSAR models may capture a wider applicability domain. However, the classification model may not distinguish between strong and weak activities due to the wide range of activity values in the ChEMBL dataset. Moreover, a definite activity threshold was not defined for the active label.

A summary of all the QSAR studies on α-glucosidase inhibitors that we looked at is presented in Table 3. It was observed that none of the studies discussed engaged the use of DL algorithms, which may be attributed in part to the use of small datasets, since DL need large datasets for effective application.

The highlighted studies used various software to develop QSAR models. These tools have unique strengths and limitations that influence their effectiveness in different contexts. Table 4 summarizes these strengths and limitations.

Various studies have explored the landscape of QSAR to establish the predictability of α-glucosidase inhibitors. These studies have shed light on the relationships between molecular structures and inhibitory potentials and have the potential to advance our understanding of α-glucosidase inhibition and facilitate the design of novel inhibitors with enhanced efficacy. However, there are peculiar challenges that need attention in order to have usable models with broader applications. The continued development and application of QSAR methods will be essential for advancing the field of α-glucosidase inhibitor research.

Table 3
QSAR studies on α-glucosidase inhibitors.

Study reference	Descriptors	Algorithm	Dataset	Size
Sainy et al. [138]	3D (CoMFA, CoMSIA)	MLR	Flavones	15
Liu et al. [139]	3D (CoMFA, CoMSIA)	PLS	Flavonols	26
Diéguez-Santana et al.[132]	0-2D	MLP, SVM, BN, BLR, KNN	Diverse set	2055
Joshi et al. [134]	2D and 3D (CoMFA, CoMSIA)	PLS	Thiazole derivatives	46
Izadpanah et al.[135]	2D (Mono-descriptor)	MLR, SVM	Hydrazinyl thiazole-based pyridine derivatives	35
Halim et al. [136]	2D	PLS	Diverse set	36
Dahmani et al.[137]	3D	MLR	1,2,4-triazolone derivatives	18
Jia et al.[140]	3D (CoMFA, CoMSIA)	PLS	Flavonoids	27
Diéguez-Santana et al.[133]	2D	DT	Diverse set	1546
Mora et al. [131]	3D-QSAR	MLR	N-cinnamoyl and hydroxycinnamoyl amides	17
Laoud et al. [141]	3D	PLS	Salacinol analogues	35
Asadollahi-Baboli & Dehnavi [142]	2D and 3D	PLS, SVM	Tetracyclic oxindole derivatives	34
Popović-Djordjević et al.[143]	2D and 3D	PLS	Cyclic urea and carbamate derivatives	13
Channar et al. [144]	3D	MLR	2,6-di (subsituted phenyl) thiazolo[3,2-*b*]− 1,2,4-triazoles	10
Zheng et al. [145]	3D (CoMFA, CoMSIA)	PLS	Xanthone analogues	54
Dinparast et al.[130]	2D and 3D	MLR	Benzimidazole derivatives	14
Zhang et al. [146]	3D	PLS	Diverse set including peptides	19
Liu et al. [147]	3D (CoMFA)	PLS	Phenolic acid amides	42
Wu et al. [129]	2D	PLS	Triterpene analogues of ursolic acid	30
Imran et al. [148]	2D	MLR	Flavone hydrazones	30
Jabeen et al. [149]	3D	PLS	Diverse set	47
Saihi et al. [150]	2D and 3D	MLR	Xanthone and curcuminoid derivatives	57
Masand et al. [151]	2D	MLR, KNN	Xanthone derivatives	43
Moorthy et al. [152]	2D and 3D	MLR	Andrographolide derivatives	19

Table 4
Some prominent software used in the development of QSAR models.

Tools	Strengths	Limitations
WEKA (Waikato Environment for Knowledge Analysis)	– Open-source and free to use – Wide variety of machine learning algorithms – Large user community and support	– Limited built-in chem-informatics functionalities
MOE (Molecular Operating Environment)	– Extensive cheminformatics toolkit – Good visualization tools – Scripting capabilities for automation	– Commercial software with licensing costs – May require significant training
SYBYL	– Good support for 3D-QSAR, including CoMFA and CoMSIA – Established user base in some fields	– Less flexibility for custom model development – Not actively developed anymore
Schrödinger	– Powerful descriptor generation capabilities – Integrated workflow for model building and analysis	– Limited open-source functionality – May require significant training
MATLAB	– Powerful programming language for custom model development – Extensive libraries for scientific computing and data analysis – Highly flexible	– Requires programming expertise – Can be time-consuming to build models from scratch – Requires licensing fee for full features
SPSS	– Good for basic QSAR modeling	– Limited machine learning capabilities – Not specifically designed for cheminformatics
Hyperchem	– Some basic descriptor generation options – Good visualization tools for QSAR analysis	– Limited functionalities for QSAR model building – Not actively developed anymore

7. Challenges and future directions

Amidst the abundance of QSAR studies and predictive models aimed at α-glucosidase prediction, a few challenges hinder the realization of comprehensive and robust models with broad applicability. Although these efforts have made significant contributions to the field, a notable proportion of them have focused on specific compound series and relied on relatively small datasets, which are often unusable for DL. This approach, though may be beneficial in terms of model precision, can inadvertently limit the broader applicability of these models, particularly in the context of screening diverse and structurally varied compound libraries. Only a few studies have considered expanding the library of compounds for training QSAR models. Illustrative cases are the studies conducted by Diéguez-Santana et al. [132,133], which commendably addressed this limitation by delving into the construction of a classification model based on an extensive and diverse ChEMBL dataset. However, it is noteworthy that the models from the study might face challenges when attempting to effectively distinguish among the vast range of activity values inherent in the ChEMBL dataset. The use of large and diverse datasets may be an issue of concern for many researchers, as this data would, in many cases, have to be pulled from different laboratory experiments, resulting in a wide range of varied

IC50 values [153,154]. However, pIC50, achieved by logarithmically transforming IC50, may circumvent this issue by normalizing and enabling direct comparison of values from disparate experiments.

While DL approach in drug discovery is on the rise (Fig. 2), and has recorded some successful applications such as identification of halicin, a novel broad spectrum antibiotic compound [155], it has been largely unexplored for developing predictive models for α-glucosidase inhibitors. For instance, a search using Scopus, an abstract and citation database launched by Elsevier in 2004, with the search terms [("Deep Learning") AND ("alpha-glucosidase" OR "α-glucosidase")], returned one article. However, the article primarily focuses on simulations of induced pluripotent stem cells and differentiated skeletal muscle cells to study the infantile onset of Pompe disease [156]. This underscores the unexplored gap in DL approach for prediction of α-glucosidase inhibitors.

To forge ahead with the development of robust and versatile models for α-glucosidase inhibitors, the key lies in the strategic use of extensive and diverse datasets for model training, alongside a concerted focus on developing DL models that can discern with precision compounds with strong α-glucosidase inhibition. Indeed, DL holds immense potential to revolutionize antidiabetic drug discovery by leveraging vast biological datasets to identify potential drug targets [157,158], repurpose existing drugs [159], predict efficacy and toxicity, perform virtual screening, enable personalized medicine approaches [160], and integrate multi-omics data for disease characterization and treatment optimization [161]. In embracing DL approaches, the identification and development of novel antidiabetic therapies may be expedited.

8. Conclusion

In conclusion, QSAR models offer valuable insights into the complexities of predicting α-glucosidase inhibitors. The combination of molecular descriptors, machine learning algorithms, and data complexity influences the quality of these models. Key descriptors used in QSAR models include 0D descriptors representing simple molecular properties, 1D descriptors capturing information on molecular fragments, 2D descriptors incorporating connectivity and topological information, and 3D descriptors quantifying the spatial arrangement of atoms and bonds. Machine learning algorithms trained on these descriptors range from classical to deep learning techniques. Classical algorithms are generally easier to implement but may struggle with complex, non-linear relationships between descriptors and desired bioactivity. In contrast, deep learning algorithms are highly effective at capturing intricate patterns but require large datasets and significant computational resources.

Many QSAR studies on α-glucosidase inhibitors examined in this article have focused on specific chemical series and used small datasets, limiting model generalizability to a wider chemical space. Aggregating data from these studies can create a more extensive and structurally diverse dataset well-suited for training deep learning algorithms. As QSAR methods advance in α-glucosidase inhibitor research, leveraging diverse datasets becomes crucial for successfully applying deep learning approaches. Embracing these advancements may accelerate progress toward discovering novel α-glucosidase inhibitors and developing practical therapeutic applications.

CRediT authorship contribution statement

Adeshina Isaiah Odugbemi: Writing – review & editing, Writing – original draft, Investigation, Conceptualization. **Clement Nyirenda:** Writing – review & editing. **Alan Christoffels:** Writing – review & editing, Supervision, Conceptualization. **Samuel A. Egieyeh:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization.

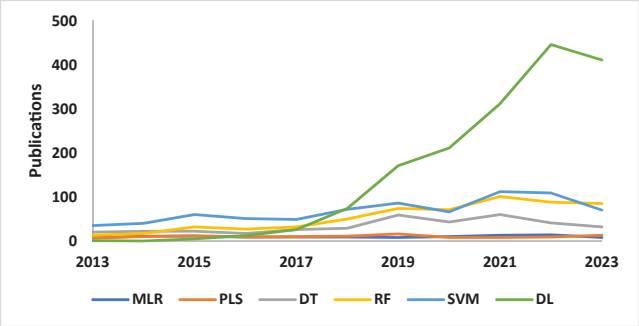

Fig. 2. Trends of algorithms' application in drug discovery using the search term [("*") AND ("drug discovery")], where "* " is Multiple Linear Regression (MLR), Partial Least Squares (PLS), Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM) or Deep Learning (DL).

Declaration of Competing Interest

The authors declare no conflict of interest that could influence this review article.

Acknowledgements

The financial assistance of the National Research Foundation (NRF) towards this study is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

References

[1] Wouters OJ, McKee M, Luyten J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. JAMA 2020;323:844–53. https://doi.org/10.1001/jama.2020.1166.
[2] Rennane S, Baker L, Mulcahy A. Estimating the cost of industry investment in drug research and development: a review of methods and results. 00469580211059731 Inquiry 2022;58. https://doi.org/10.1177/00469580211059731.
[3] Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today 2021;26:80–93. https://doi.org/10.1016/j.drudis.2020.10.010.
[4] Qureshi R, Irfan M, Gondal TM, Khan S, Wu J, Hadi MU, et al. AI in drug discovery and its clinical relevance. Heliyon 2023;9:e17575. https://doi.org/10.1016/j.heliyon.2023.e17575.
[5] Sarker IH. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. SN Comput Sci 2022;3:158. https://doi.org/10.1007/s42979-022-01043-x.
[6] Mao J, Akhtar J, Zhang X, Sun L, Guan S, Li X, et al. Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. iScience 2021;24:103052. https://doi.org/10.1016/j.isci.2021.103052.
[7] Staszak M, Staszak K, Wieszczycka K, Bajek A, Roszkowski K, Tylkowski B. Machine learning in drug design: Use of artificial intelligence to explore the chemical structure–biological activity relationship. WIREs Comput Mol Sci 2022;12:e1568. https://doi.org/10.1002/wcms.1568.
[8] Jaacks LM, Siegel KR, Gujral UP, Narayan KMV. Type 2 diabetes: a 21st century epidemic. Best Pract Res Clin Endocrinol Metab 2016;30:331–43. https://doi.org/10.1016/j.beem.2016.05.003.
[9] Kharroubi AT, Darwish HM. Diabetes mellitus: the epidemic of the century. World J Diabetes 2015;6:850–67. https://doi.org/10.4239/wjd.v6.i6.850.
[10] Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat Rev Endocrinol 2018;14:88–98. https://doi.org/10.1038/nrendo.2017.151.
[11] Ceriello A. Postprandial hyperglycemia and diabetes complications: is it time to treat. Diabetes 2005;54:1–7. https://doi.org/10.2337/diabetes.54.1.1.
[12] Kausar S, Falcao AO. An automated framework for QSAR model building. J Chemin- 2018;10(1). https://doi.org/10.1186/s13321-017-0256-5.
[13] Jiménez-Luna J, Grisoni F, Weskamp N, Schneider G. Artificial intelligence in drug discovery: recent advances and future perspectives. Expert Opin Drug Discov 2021;16:949–59. https://doi.org/10.1080/17460441.2021.1909567.
[14] Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. Artif Intell Healthc 2020:25–60. https://doi.org/10.1016/B978-0-12-818438-7.00002-2.
[15] Xu Y, Liu X, Cao X, Huang C, Liu E, Qian S, et al. Artificial intelligence: a powerful paradigm for scientific research. Innovation 2021;2:100179. https://doi.org/10.1016/j.xinn.2021.100179.

[16] Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, Cedrón F, Novoa FJ, Carballal A, et al. A review on machine learning approaches and trends in drug discovery. Comput Struct Biotechnol J 2021;19:4538–58. https://doi.org/10.1016/j.csbj.2021.08.011.

[17] Abuhammad A, Taha MO. QSAR studies in the discovery of novel type-II diabetic therapies. Expert Opin Drug Discov 2016;11:197–214. https://doi.org/10.1517/17460441.2016.1118046.

[18] Dudek A, Arodz T, Galvez J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. CCHTS 2006;9:213–28. https://doi.org/10.2174/138620706776055539.

[19] Gupta SK, Tripathi PK. CADD studies in the discovery of potential ARI (aldose reductase inhibitors) agents for the treatment of diabetic complications. e180822207672 Curr Diabetes Rev 2023;19. https://doi.org/10.2174/1573399819666220818163758.

[20] Kaur P, Khatik G. An overview of computer-aided drug design tools and recent applications in designing of anti-diabetic agents. Curr Drug Targets 2021;22:1158–82. https://doi.org/10.2174/1389450121666201119141525.

[21] Riyaphan J, Pham D-C, Leong MK, Weng C-F. In silico approaches to identify polyphenol compounds as α-glucosidase and α-amylase inhibitors against type-II diabetes. Biomolecules 2021;11:1877. https://doi.org/10.3390/biom11121877.

[22] Wang T, Wu M-B, Lin J-P, Yang L-R. Quantitative structure–activity relationship: promising advances in drug discovery platforms. Expert Opin Drug Discov 2015;10:1283–300. https://doi.org/10.1517/17460441.2015.1083006.

[23] Hansch C, Maloney PP, Fujita T, Muir RM. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. Nature 1962;194:178–80. https://doi.org/10.1038/194178b0.

[24] Achary PGR. Applications of quantitative structure-activity relationships (QSAR) based virtual screening in drug design: a review. Mini Rev Med Chem 2020;20:1375–88. https://doi.org/10.2174/1389557520666200429102334.

[25] Wu Y, Li M, Shen J, Pu X, Guo Y. A consensual machine-learning-assisted QSAR model for effective bioactivity prediction of xanthine oxidase inhibitors using molecular fingerprints. Mol Divers 2023. https://doi.org/10.1007/s11030-023-10649-z.

[26] Kubinyi H. Quantitative structure-activity relationships. 2. A mixed approach, based on Hansch and Free-Wilson analysis. J Med Chem 1976;19:587–600. https://doi.org/10.1021/jm00227a004.

[27] Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 1988;110:5959–67. https://doi.org/10.1021/ja00226a005.

[28] Hopfinger AJ, Wang S, Tokarski JS, Jin B, Albuquerque M, Madhav PJ, et al. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. J Am Chem Soc 1997;119:10509–24. https://doi.org/10.1021/ja9718937.

[29] Vedani A, Dobler M, Lill MA. Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. J Med Chem 2005;48:3700–3. https://doi.org/10.1021/jm050185q.

[30] Vedani A, Dobler M. Multidimensional QSAR: moving from three- to five-dimensional concepts. Quant Struct-Act Relatsh 2002;21:382–90. https://doi.org/10.1002/1521-3838(200210)21:4<382::AID-QSAR382>3.0.CO;2-L.

[31] Vedani A, Dobler M. 5D-QSAR: the key for simulating induced fit. J Med Chem 2002;45:2139–49. https://doi.org/10.1021/jm011005p.

[32] Consonni V, Ballabio D, Todeschini R. Chapter 12 - Chemical space and molecular descriptors for QSAR studies. In: Roy K, editor. Cheminformatics, QSAR and Machine Learning Applications for Novel Drug Development. Academic Press; 2023. p. 303–27. https://doi.org/10.1016/B978-0-443-18638-7.00022-0.

[33] Wassermann AM, Wawer M, Bajorath J. Activity landscape representations for structure−activity relationship analysis. J Med Chem 2010;53:8209–23. https://doi.org/10.1021/jm100933w.

[34] Consonni V, Todeschini R. Molecular descriptors. In: Puzyn T, Leszczynski J, Cronin MT, editors. Recent Advances in QSAR Studies, vol. 8. Dordrecht: Springer Netherlands; 2010. p. 29–102. https://doi.org/10.1007/978-1-4020-9783-6_3.

[35] Bahia MS, Kaspi O, Touitou M, Binayev I, Dhail S, Spiegel J, et al. A comparison between 2D and 3D descriptors in QSAR modeling based on bio-active conformations. Mol Inform 2023;42:2200186. https://doi.org/10.1002/minf.202200186.

[36] Sato A, Miyao T, Jasial S, Funatsu K. Comparing predictive ability of QSAR/QSPR models using 2D and 3D molecular representations. J Comput Aided Mol Des 2021;35:179–93. https://doi.org/10.1007/s10822-020-00361-7.

[37] Grisoni F, Ballabio D, Todeschini R, Consonni V. Molecular descriptors for structure–activity applications: a hands-on approach. In: Nicolotti O, editor. Computational Toxicology, vol. 1800. New York, NY: Springer New York; 2018. p. 3–53. https://doi.org/10.1007/978-1-4939-7899-1_1.

[38] Mauri A, Consonni V, Todeschini R. Molecular Descriptors. In: Leszczynski J, Kaczmarek-Kedziera A, Puzyn T, Papadopoulos M G, Reis H, Shukla M K, editors. Handbook of Computational Chemistry. Cham: Springer International Publishing; 2017. p. 2065–93. https://doi.org/10.1007/978-3-319-27282-5_51.

[39] Zanni R, Galvez-Llompart M, García-Domenech R, Galvez J. Latest advances in molecular topology applications for drug discovery. Expert Opin Drug Discov 2015;10:945–57. https://doi.org/10.1517/17460441.2015.1062751.

[40] Wigh DS, Goodman JM, Lapkin AA. A review of molecular representation in the age of machine learning. WIREs Comput Mol Sci 2022;12:e1603. https://doi.org/10.1002/wcms.1603.

[41] Hemmateenejad B, Sanchooli M. Substituent electronic descriptors for fast QSAR/QSPR. J Chemom 2007;21:96–107. https://doi.org/10.1002/cem.1039.

[42] Acke G, De Baerdemacker S, Martín Pendás Á, Bultinck P. Hierarchies of quantum chemical descriptors induced by statistical analyses of domain occupation number operators. WIREs Comput Mol Sci 2020;10. https://doi.org/10.1002/wcms.1456.

[43] Karelson M, Lobanov VS, Katritzky AR. Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev 1996;96:1027–44. https://doi.org/10.1021/cr950202r.

[44] Todeschini R, Gramatica P. The Whim theory: new 3D molecular descriptors for qsar in environmental modelling. SAR QSAR Environ Res 1997;7:89–115. https://doi.org/10.1080/10629369708039126.

[45] Consonni V, Todeschini R, Pavan M. Structure/response correlations and similarity/diversity analysis by GETAWAY Descriptors. 1. theory of the novel 3D molecular descriptors. J Chem Inf Comput Sci 2002;42:682–92. https://doi.org/10.1021/ci015504a.

[46] Damale M, Harke S, Kalam Khan F, Shinde D, Sangshetti J. Recent advances in multidimensional QSAR (^{4}D–^{6}D): a critical review. MRMC 2014;14:35–55. https://doi.org/10.2174/1389557513136660104.

[47] Pourbasheer E, Amanlou M. 3D-QSAR analysis of anti-cancer agents by CoMFA and CoMSIA. Med Chem Res 2014;23:800–9. https://doi.org/10.1007/s00044-013-0676-3.

[48] Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 1994;37:4130–46. https://doi.org/10.1021/jm00050a010.

[49] Moumbock AFA, Li J, Mishra P, Gao M, Günther S. Current computational methods for predicting protein interactions of natural products. Comput Struct Biotechnol J 2019;17:1367–76. https://doi.org/10.1016/j.csbj.2019.08.008.

[50] Lim S, Lu Y, Cho CY, Sung I, Kim J, Kim Y, et al. A review on compound-protein interaction prediction methods: Data, format, representation and model. Comput Struct Biotechnol J 2021;19:1541–56. https://doi.org/10.1016/j.csbj.2021.03.004.

[51] Lo Y-C, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. Drug Discov Today 2018;23:1538–46. https://doi.org/10.1016/j.drudis.2018.05.010.

[52] Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods 2015;71:58–63. https://doi.org/10.1016/j.ymeth.2014.08.005.

[53] Yang J, Cai Y, Zhao K, Xie H, Chen X. Concepts and applications of chemical fingerprint for hit and lead screening. Drug Discov Today 2022;27:103356. https://doi.org/10.1016/j.drudis.2022.103356.

[54] Gütlein M, Kramer S. Filtered circular fingerprints improve either prediction or runtime performance while retaining interpretability. J Chemin- 2016;8:60. https://doi.org/10.1186/s13321-016-0173-z.

[55] Beno BR, Mason JS. The design of combinatorial libraries using properties and 3D pharmacophore fingerprints. Drug Discov Today 2001;6:251–8. https://doi.org/10.1016/S1359-6446(00)01665-2.

[56] Zhao Z, Bourne PE. Harnessing systematic protein–ligand interaction fingerprints for drug discovery. Drug Discov Today 2022;27:103319. https://doi.org/10.1016/j.drudis.2022.07.004.

[57] Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, et al. QSAR without borders. Chem Soc Rev 2020;49:3525–64. https://doi.org/10.1039/D0CS00098A.

[58] Patel L, Shukla T, Huang X, Ussery DW, Wang S. Machine learning methods in drug discovery. Molecules 2020;25:5277. https://doi.org/10.3390/molecules25225277.

[59] Liu P, Long W. Current mathematical methods used in QSAR/QSPR studies. Int J Mol Sci 2009;10:1978–98. https://doi.org/10.3390/ijms10051978.

[60] Chung D, Keles S. Sparse partial least squares classification for high dimensional data. Stat Appl Genet Mol Biol 2010;9:17. https://doi.org/10.2202/1544-6115.1492.

[61] Abdi H. Partial least squares regression and projection on latent structure regression (PLS Regression). WIREs Comput Stats 2010;2:97–106. https://doi.org/10.1002/wics.51.

[62] Bentler PM, Huang W. On components, latent variables, PLS and simple methods: reactions to rigdon's rethinking of PLS. Long Range Plann 2014;47:138–45. https://doi.org/10.1016/j.lrp.2014.02.005.

[63] Xie Z, Feng X, Chen X. Partial least trimmed squares regression. Chemom Intell Lab Syst 2022;221:104486. https://doi.org/10.1016/j.chemolab.2021.104486.

[64] Hair JF, Risher JJ, Sarstedt M, Ringle CM. When to use and how to report the results of PLS-SEM. Eur Bus Rev 2019;31:2–24. https://doi.org/10.1108/EBR-11-2018-0203.

[65] Smita M. logistic regression model–A REVIEW. Int J Innov Sci Res Technol 2021;6:1276–80.

[66] Schöning V, Hammann F. How far have decision tree models come for data mining in drug discovery. Expert Opin Drug Discov 2018;13:1067–9. https://doi.org/10.1080/17460441.2018.1538208.

[67] Krzywinski M, Altman N. Classification and regression trees. Nat Methods 2017;14(8):757. https://doi.org/10.1038/nmeth.4370.

[68] Costa VG, Pedreira CE. Recent advances in decision trees: an updated survey. Artif Intell Rev 2023;56:4765–800. https://doi.org/10.1007/s10462-022-10275-5.

[69] Zimmermann A. Ensemble-trees: leveraging ensemble power inside decision trees. Discovery Science. Berlin, Heidelberg: Springer; 2008. p. 76–87. https://doi.org/10.1007/978-3-540-88411-8_10.

[70] Ahn S, Lee SE, Kim M. Random-forest model for drug–target interaction prediction via Kullback–Leibler divergence. J Chemin- 2022;14:67. https://doi.org/10.1186/s13321-022-00644-1.

[71] Kapsiani S, Howlin BJ. Random forest classification for predicting lifespan-extending chemical compounds. Sci Rep 2021;11:13812. https://doi.org/10.1038/s41598-021-93070-6.

[72] Yu F, Wei C, Deng P, Peng T, Hu X. Deep exploration of random forest model boosts the interpretability of machine learning studies of complicated immune responses and lung burden of nanoparticles. Sci Adv 2021;7:eabf4130. https://doi.org/10.1126/sciadv.abf4130.
[73] Montesinos López OA, Montesinos López A, Crossa J. Support Vector Machines and Support Vector Regression. In: Montesinos López OA, Montesinos López A, Crossa J, editors. Multivariate Statistical Machine Learning Methods for Genomic Prediction. Cham: Springer International Publishing; 2022. p. 337–78. https://doi.org/10.1007/978-3-030-89010-0_9.
[74] Tharwat A. Parameter investigation of support vector machine classifier with kernel functions. Knowl Inf Syst 2019;61:1269–302. https://doi.org/10.1007/s10115-019-01335-4.
[75] Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing 2020;408:189–215. https://doi.org/10.1016/j.neucom.2019.10.118.
[76] Karal Ö. Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation. Innov Intell Syst Appl Conf (ASYU) 2020;2020:1–5. https://doi.org/10.1109/ASYU50717.2020.9259880.
[77] Alvarsson J, Lampa S, Schaal W, Andersson C, Wikberg JES, Spjuth O. Large-scale ligand-based predictive modelling using support vector machines. J Chemin-2016;8:39. https://doi.org/10.1186/s13321-016-0151-5.
[78] Rodríguez-Pérez R, Bajorath J. Evolution of support vector machine and regression modeling in chemoinformatics and drug discovery. J Comput Aided Mol Des 2022;36:355–62. https://doi.org/10.1007/s10822-022-00442-9.
[79] Roman I, Santana R, Mendiburu A, Lozano JA. In-depth analysis of SVM kernel learning and its components. Neural Comput Applic 2021;33:6575–94. https://doi.org/10.1007/s00521-020-05419-z.
[80] Abu Alfeilat HA, Hassanat ABA, Lasassmeh O, Tarawneh AS, Alhasanat MB, Eyal Salman HS, et al. Effects of distance measure choice on K-nearest neighbor classifier performance: a review. Big Data 2019;7:221–48. https://doi.org/10.1089/big.2018.0175.
[81] Taunk K, De S, Verma S, Swetapadma A. A brief review of nearest neighbor algorithm for learning and classification. Int Conf Intell Comput Control Syst (ICCS) 2019;2019:1255–60. https://doi.org/10.1109/ICCS45141.2019.9065747.
[82] Wei Z, He Q, Zhao Y. Machine learning for battery research. J Power Sources 2022;549:232125. https://doi.org/10.1016/j.jpowsour.2022.232125.
[83] Han Z-X, Wang C-Y, Wei J-Y, Huang C-J, Zhang W-H, Luo B. Methodological survey of using Bayesian Network for predicting pharmacology-based bioactivities of Chinese medicines: a scoping review. TMR Pharmacol Res 2023;3.
[84] Darwiche A. Chapter 11 Bayesian Networks. In: van Harmelen F, Lifschitz V, Porter B, editors. Foundations of Artificial Intelligence, vol. 3. Elsevier; 2008. p. 467–509. https://doi.org/10.1016/S1574-6526(07)03011-8.
[85] Marcot BG, Penman TD. Advances in Bayesian network modelling: Integration of modelling technologies. Environ Model Softw 2019;111:386–93. https://doi.org/10.1016/j.envsoft.2018.09.016.
[86] Rohmer J. Uncertainties in conditional probability tables of discrete Bayesian Belief Networks: a comprehensive review. Eng Appl Artif Intell 2020;88:103384. https://doi.org/10.1016/j.engappai.2019.103384.
[87] Kyrimi E, McLachlan S, Dube K, Neves MR, Fahmi A, Fenton N. A comprehensive scoping review of Bayesian networks in healthcare: past, present and future. Artif Intell Med 2021;117:102108. https://doi.org/10.1016/j.artmed.2021.102108.
[88] Rojas MG, Olivera AC, Vidal PJ. Optimising multilayer perceptron weights and biases through a cellular genetic algorithm for medical data classification. Array 2022;14:100173. https://doi.org/10.1016/j.array.2022.100173.
[89] Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure–activity relationships. J Chem Inf Model 2015;55:263–74. https://doi.org/10.1021/ci500747n.
[90] Sadeghi F, Afkhami A, Madrakian T, Ghavami R. QSAR analysis on a large and diverse set of potent phosphoinositide 3-kinase gamma (PI3Kγ) inhibitors using MLR and ANN methods. Sci Rep 2022;12:6090. https://doi.org/10.1038/s41598-022-09843-0.
[91] Hentabli H, Bengherbia B, Saeed F, Salim N, Nafea I, Toubal A, et al. Convolutional neural network model based on 2D fingerprint for bioactivity prediction. Int J Mol Sci 2022;23:13230. https://doi.org/10.3390/ijms232113230.
[92] Lee S, Lee M, Gyak K-W, Kim SD, Kim M-J, Min K. Novel solubility prediction models: molecular fingerprints and physicochemical features vs graph convolutional neural networks. ACS Omega 2022;7:12268–77. https://doi.org/10.1021/acsomega.2c00697.
[93] Mendolia I, Contino S, Perricone U, Pirrone R, Ardizzone E. A convolutional neural network for virtual screening of molecular fingerprints. In: Ricci E, Rota Bulò S, Snoek C, Lanz O, Messelodi S, Sebe N, editors. Image Analysis and Processing – ICIAP 2019. Cham: Springer International Publishing; 2019. p. 399–409. https://doi.org/10.1007/978-3-030-30642-7_36.
[94] Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional neural networks. Pattern Recognit 2018;77:354–77. https://doi.org/10.1016/j.patcog.2017.10.013.
[95] Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. Int Conf Eng Technol (ICET) 2017;2017:1–6. https://doi.org/10.1109/ICEngTechnol.2017.8308186.
[96] Lopez Pinaya WH, Vieira S, Garcia-Dias R, Mechelli A. Chapter 10 - Convolutional neural networks. In: Mechelli A, Vieira S, editors. Machine Learning. Academic Press; 2020. p. 173–91. https://doi.org/10.1016/B978-0-12-815739-8.00010-9.
[97] Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging 2018;9:611–29. https://doi.org/10.1007/s13244-018-0639-9.
[98] Tyagi AK, Abraham A. Recurrent neural networks: concepts and applications. CRC Press; 2022.
[99] Grisoni F, Moret M, Lingwood R, Schneider G. Bidirectional molecule generation with recurrent neural networks. J Chem Inf Model 2020;60:1175–83. https://doi.org/10.1021/acs.jcim.9b00943.
[100] Gupta A, Müller AT, Huisman BJH, Fuchs JA, Schneider P, Schneider G. Generative recurrent networks for De Novo drug design. Mol Inf 2018;37:1700111. https://doi.org/10.1002/minf.201700111.
[101] Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent Sci 2018;4:120–31. https://doi.org/10.1021/acscentsci.7b00512.
[102] Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique SMILES notation. J Chem Inf Comput Sci 1989;29:97–101. https://doi.org/10.1021/ci00062a008.
[103] Lai X, Yang P, Wang K, Yang Q, Yu D. MGRNN: structure generation of molecules based on graph recurrent neural networks. Mol Inform 2021;40:2100091. https://doi.org/10.1002/minf.202100091.
[104] D'Souza S, Kv P, Balaji S. Training recurrent neural networks as generative neural networks for molecular structures: how does it impact drug discovery? Expert Opin Drug Discov 2022;17:1071–9. https://doi.org/10.1080/17460441.2023.2134340.
[105] Emmert-Streib F, Yang Z, Feng H, Tripathi S, Dehmer M. An introductory review of deep learning for prediction models with big data. Front Artif Intell 2020;3.
[106] Mikhaeil J, Monfared Z, Durstewitz D. On the difficulty of learning chaotic dynamics with RNNs. Adv Neural Inf Process Syst 2022;35:11297–312.
[107] Chandar S, Sankar C, Vorontsov E, Kahou SE, Bengio Y. Towards non-saturating recurrent units for modelling long-term dependencies. Proc AAAI Conf Artif Intell 2019;33:3280–7. https://doi.org/10.1609/aaai.v33i01.33013280.
[108] Chandra N, Ahuja L, Khatri SK, Monga H. Utilizing gated recurrent units to retain long term dependencies with recurrent neural network in text classification. J Inf Syst Telecom 2021;2:89.
[109] Lee CY, Chen Y-PP. Prediction of drug adverse events using deep learning in pharmaceutical discovery. Brief Bioinforma 2021;22:1884–901. https://doi.org/10.1093/bib/bbaa040.
[110] Hung C, Gini G. QSAR modeling without descriptors using graph convolutional neural networks: the case of mutagenicity prediction. Mol Divers 2021;25:1283–99. https://doi.org/10.1007/s11030-021-10250-2.
[111] Wang Y, Wang J, Cao Z, Barati Farimani A. Molecular contrastive learning of representations via graph neural networks. Nat Mach Intell 2022;4:279–87. https://doi.org/10.1038/s42256-022-00447-x.
[112] Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, et al. Graph neural networks: a review of methods and applications. AI Open 2020;1:57–81. https://doi.org/10.1016/j.aiopen.2021.01.001.
[113] Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, Shao C, et al. Graph neural networks for materials science and chemistry. Commun Mater 2022;3:1–18. https://doi.org/10.1038/s43246-022-00315-6.
[114] Wieder O, Kohlbacher S, Kuenemann M, Garon A, Ducrot P, Seidel T, et al. A compact review of molecular property prediction with graph neural networks. Drug Discov Today: Technol 2020;37:1–12. https://doi.org/10.1016/j.ddtec.2020.11.009.
[115] Zeng X, Xiang H, Yu L, Wang J, Li K, Nussinov R, et al. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. Nat Mach Intell 2022;4:1004–16. https://doi.org/10.1038/s42256-022-00557-6.
[116] Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature selection: a data perspective. ACM Comput Surv 2018;50:1–45. https://doi.org/10.1145/3136625.
[117] Danishuddin, Khan AU. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discov Today 2016;21:1291–302. https://doi.org/10.1016/j.drudis.2016.06.013.
[118] Goodarzi M, Dejaegher B, Heyden YV. Feature selection methods in QSAR studies. J AOAC Int 2012;95:636–51. https://doi.org/10.5740/jaoacint.SGE_Goodarzi.
[119] De P, Kar S, Ambure P, Roy K. Prediction reliability of QSAR models: an overview of various validation tools. Arch Toxicol 2022;96:1279–95. https://doi.org/10.1007/s00204-022-03252-y.
[120] Oselusi SO, Dube P, Odugbemi AI, Akinyede KA, Ilori TL, Egieyeh E, et al. The role and potential of computer-aided drug discovery strategies in the discovery of novel antimicrobials. Comput Biol Med 2024;169:107927. https://doi.org/10.1016/j.compbiomed.2024.107927.
[121] Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. Chemom Intell Lab Syst 2015;145:22–9. https://doi.org/10.1016/j.chemolab.2015.04.013.
[122] Tropsha A, Golbraikh A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Curr Pharm Des 2007;13:3494–504. https://doi.org/10.2174/138161207782794257.
[123] Sharma MC, Sharma S. Molecular modeling studies of thiophenyl C-aryl glucoside SGLT2 inhibitors as potential antidiabetic agents. Int J Med Chem 2014;2014:e739646. https://doi.org/10.1155/2014/739646.
[124] Zhi H, Zheng J, Chang Y, Li Q, Liao G, Wang Q, et al. QSAR studies on triazole derivatives as sglt inhibitors via CoMFA and CoMSIA. J Mol Struct 2015;1098:199–205. https://doi.org/10.1016/j.molstruc.2015.06.004.

[125] Liu R, Cheng J, Wu H. Discovery of food-derived dipeptidyl peptidase IV inhibitory peptides: a review. Int J Mol Sci 2019;20:463. https://doi.org/10.3390/ijms20030463.
[126] Wang Z, Chen J, Hong H. Developing QSAR models with defined applicability domains on PPARγ binding affinity using large data sets and machine learning algorithms. Environ Sci Technol 2021;55:6857–66. https://doi.org/10.1021/acs.est.0c07040.
[127] Cabezas D, Mellado G, Espinoza N, Gárate JA, Morales C, Castro-Alvarez A, et al. In silico approaches to develop new phenyl-pyrimidines as glycogen synthase kinase 3 (GSK-3) inhibitors with halogen-bonding capabilities: 3D-QSAR CoMFA/CoMSIA, molecular docking and molecular dynamics studies. J Biomol Struct Dyn 2023;41:13250–9. https://doi.org/10.1080/07391102.2023.2172457.
[128] Arthur DE, Ejeh S, Uzairu A. Quantitative structure-activity relationship (QSAR) and design of novel ligands that demonstrate high potency and target selectivity as protein tyrosine phosphatase 1B (PTP 1B) inhibitors as an effective strategy used to model anti-diabetic agents. J Recept Signal Transduct 2020;40:501–20. https://doi.org/10.1080/10799893.2020.1759092.
[129] Wu P, Zheng J, Huang T, Li D, Hu Q, Cheng A, et al. Synthesis and evaluation of novel triterpene analogues of ursolic acid as potential antidiabetic agent. PLoS ONE 2015;10. https://doi.org/10.1371/journal.pone.0138767.
[130] Dinparast L, Valizadeh H, Bahadori MB, Soltani S, Asghari B, Rashidi M-R. Design, synthesis, α-glucosidase inhibitory activity, molecular docking and QSAR studies of benzimidazole derivatives. J Mol Struct 2016;1114:84–94. https://doi.org/10.1016/j.molstruc.2016.02.005.
[131] Mora JR, Márquez EA, Calle L. Computational molecular modelling of N-cinnamoyl and hydroxycinnamoyl amides as potential α-glucosidase inhibitors. Med Chem Res 2018;27:2214–23. https://doi.org/10.1007/s00044-018-2229-2.
[132] Dieguez-Santana K, Pham-The H, Rivera-Borroto OM, Puris A, Le-Thi-Thu H, Casanola-Martin GM. A two QSAR way for antidiabetic agents targeting using α-amylase and α-glucosidase inhibitors: model parameters settings in artificial intelligence techniques. Lett Drug Des Discov 2017;14:862–8.
[133] Diéguez-Santana K, Rivera-Borroto OM, Puris A, Pham-The H, Le-Thi-Thu H, Rasulev B, et al. Beyond model interpretability using LDA and decision trees for α-amylase and α-glucosidase inhibitor classification studies. Chem Biol Drug Des 2019;94:1414–21. https://doi.org/10.1111/cbdd.13518.
[134] Joshi D, Yadav S, Sharma R, Pandya M, Bhadauria RS. Molecular modelling studies on thiazole-based α-glucosidase inhibitors using docking and CoMFA, CoMSIA and HQSAR. Curr Drug Discov Technol 2021;18. https://doi.org/10.2174/1570163817666201022111213.
[135] Izadpanah E, Riahi S, Abbasi-Radmoghaddam Z, Gharaghani S, Mohammadi-Khanaposhtanai M. A simple and robust model to predict the inhibitory activity of α-glucosidase inhibitors through combined QSAR modeling and molecular docking techniques. Mol Divers 2021;25:1811–25. https://doi.org/10.1007/s11030-020-10164-5.
[136] Halim SA, Jabeen S, Khan A, Al-Harrasi A. Rational design of novel inhibitors of α-glucosidase: an application of quantitative structure activity relationship and structure-based virtual screening. Pharmaceuticals 2021;14. https://doi.org/10.3390/ph14050482.
[137] Dahmani R, Manachou M, Belaidi S, Chtita S, Boughdiri S. Structural characterization and QSAR modeling of 1,2,4-triazole derivatives as α-glucosidase inhibitors. N J Chem 2021;45:1253–61. https://doi.org/10.1039/d0nj05298a.
[138] Sainy N, Dubey N, Sharma R, Dubey N, Sainy J. 3D QSAR analysis of flavones as antidiabetic agents. Res J Pharm Technol 2022;15:1689–95. https://doi.org/10.52711/0974-360X.2022.00283.
[139] Liu Y, Wang R, Ren C, Pan Y, Li J, Zhao X, et al. Two myricetin-derived flavonols from morella rubra leaves as potent α -glucosidase inhibitors and structure-activity relationship study by computational chemistry. Oxid Med Cell Longev 2022;2022. https://doi.org/10.1155/2022/9012943.
[140] Jia Y, Ma Y, Cheng G, Zhang Y, Cai S. Comparative study of dietary flavonoids with different structures as α-glucosidase inhibitors and insulin sensitizers. J Agric Food Chem 2019;67:10521–33. https://doi.org/10.1021/acs.jafc.9b04943.
[141] Laoud A, Ferkous F, Maccari L, Maccari G, Saihi Y, Kraim K. Identification of novel nt-MGAM inhibitors for potential treatment of type 2 diabetes: Virtual screening, atom based 3D-QSAR model, docking analysis and ADME study. Comput Biol Chem 2018;72:122–35. https://doi.org/10.1016/j.compbiolchem.2017.12.003.
[142] Asadollahi-Baboli M, Dehnavi S. Docking and QSAR analysis of tetracyclic oxindole derivatives as α-glucosidase inhibitors. Comput Biol Chem 2018;76:283–92. https://doi.org/10.1016/j.compbiolchem.2018.07.019.
[143] Popović-Djordjević JB, Jevtić II, Grozdanić ND, Šegan SB, Zlatović MV, Ivanović MD, et al. α-Glucosidase inhibitory activity and cytotoxic effects of some cyclic urea and carbamate derivatives. J Enzym Inhib Med Chem 2017;32:298–303. https://doi.org/10.1080/14756366.2016.1250754.
[144] Channar PA, Saeed A, Larik FA, Rashid S, Iqbal Q, Rozi M, et al. Design and synthesis of 2,6-di(substituted phenyl)thiazolo[3,2-b]-1,2,4-triazoles as α-glucosidase and α-amylase inhibitors, co-relative Pharmacokinetics and 3D QSAR and risk analysis. Biomed Pharm 2017;94:499–513. https://doi.org/10.1016/j.biopha.2017.07.139.
[145] Zheng X, Zhou S, Zhang C, Wu D, Luo H-B, Wu Y. Docking-assisted 3D-QSAR studies on xanthones as α-glucosidase inhibitors. J Mol Model 2017;23. https://doi.org/10.1007/s00894-017-3438-1.
[146] Zhang Y, Wang N, Wang W, Wang J, Zhu Z, Li X. Molecular mechanisms of novel peptides from silkworm pupae that inhibit α-glucosidase. Peptides 2016;76:45–50. https://doi.org/10.1016/j.peptides.2015.12.004.
[147] Liu B, Ma J-M, Chen H-W, Li Z-L, Sun L-H, Zeng Z, et al. α-Glucosidase inhibitory activities of phenolic acid amides with l-amino acid moiety. RSC Adv 2016;6:50837–45. https://doi.org/10.1039/c6ra08330g.
[148] Imran S, Taha M, Ismail NH, Kashif SM, Rahim F, Jamil W, et al. Synthesis of novel flavone hydrazones: in-vitro evaluation of α-glucosidase inhibition, QSAR analysis and docking studies. Eur J Med Chem 2015;105:156–70. https://doi.org/10.1016/j.ejmech.2015.10.017.
[149] Jabeen F, Oliferenko PV, Oliferenko AA, Pillai GG, Ansari FL, Hall CD, et al. Dual inhibition of the α-glucosidase and butyrylcholinesterase studied by molecular field topology analysis. Eur J Med Chem 2014;80:228–42. https://doi.org/10.1016/j.ejmech.2014.04.018.
[150] Saihi Y, Kraim K, Ferkous F, Djeghaba Z, Azzouzi A, Benouis S. Nonlinear qsar study of xanthone and curcuminoid derivatives as α-glucosidase inhibitors. Bull Korean Chem Soc 2013;34:1643–50. https://doi.org/10.5012/bkcs.2013.34.6.1643.
[151] Masand VH, Mahajan DT, Patil KN, Chinchkhede KD, Jawarkar RD, Hadda TB, et al. k-NN, quantum mechanical and field similarity based analysis of xanthone derivatives as α-glucosidase inhibitors. Med Chem Res 2012;21:4523–34. https://doi.org/10.1007/s00044-012-9995-z.
[152] Moorthy HN, Ramos MJ, Fernandes PA. Prediction of the relationship between the structural features of andrographolide derivatives and α-glucosidase inhibitory activity: a quantitative structure-activity relationship (QSAR) Study. J Enzym Inhib Med Chem 2011;26:78–87. https://doi.org/10.3109/14756361003724760.
[153] Lewis RA, Wood D. Modern 2D QSAR for drug discovery. WIREs Comput Mol Sci 2014;4:505–22. https://doi.org/10.1002/wcms.1187.
[154] Maharao N, Antontsev V, Wright M, Varshney J. Entering the era of computationally driven drug development. Drug Metab Rev 2020;52:283–98. https://doi.org/10.1080/03602532.2020.1726944.
[155] Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell 2020;180:688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021.
[156] Esmail S, Danter WR. DeepNEU: artificially induced stem cell (aiPSC) and differentiated skeletal muscle cell (aiSkMC) simulations of infantile onset POMPE disease (IOPD) for potential biomarker identification and drug discovery. Front Cell Dev Biol 2019;7:325. https://doi.org/10.3389/fcell.2019.00325.
[157] Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, et al. Deep-learning-based drug–target interaction prediction. J Proteome Res 2017;16:1401–9. https://doi.org/10.1021/acs.jproteome.6b00618.
[158] You J, McLeod RD, Hu P. Predicting drug-target interaction network using deep learning model. Comput Biol Chem 2019;80:90–101. https://doi.org/10.1016/j.compbiolchem.2019.03.016.
[159] Zhu T, Li K, Herrero P, Georgiou P. Deep learning for diabetes: a systematic review. IEEE J Biomed Health Inform 2021;25:2744–57. https://doi.org/10.1109/JBHI.2020.3040225.
[160] Petrovic K. Deep learning in personalized medicine: advancements and applications. J Adv Anal Healthc Manag 2023;7:34–50.
[161] Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: a review. Biotechnol Adv 2021;49:107739. https://doi.org/10.1016/j.biotechadv.2021.107739.