iScience

CellPress
OPEN ACCESS

Review

Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models

Jiashun Mao,[1,2,3,9] Javed Akhtar,[2,4,9] Xiao Zhang,[5,9] Liang Sun,[6] Shenghui Guan,[2,3] Xinyu Li,[7] Guangming Chen,[2,4] Jiaxin Liu,[8] Hyeon-Nae Jeon,[8] Min Sung Kim,[8] Kyoung Tai No,[1,*] and Guanyu Wang[2,3,4,*]

SUMMARY

Early quantitative structure-activity relationship (QSAR) technologies have unsatisfactory versatility and accuracy in fields such as drug discovery because they are based on traditional machine learning and interpretive expert features. The development of Big Data and deep learning technologies significantly improve the processing of unstructured data and unleash the great potential of QSAR. Here we discuss the integration of wet experiments (which provide experimental data and reliable verification), molecular dynamics simulation (which provides mechanistic interpretation at the atomic/molecular levels), and machine learning (including deep learning) techniques to improve QSAR models. We first review the history of traditional QSAR and point out its problems. We then propose a better QSAR model characterized by a new iterative framework to integrate machine learning with disparate data input. Finally, we discuss the application of QSAR and machine learning to many practical research fields, including drug development and clinical trials.

INTRODUCTION

Machine learning (ML), with historical breakthroughs being made, has been widely used in Big Data and parallel computation applications such as image recognition, knowledge representation, robotics, autonomous driving, and drug development (Freitag, 2000; Krizhevsky et al., 2017; Kukar et al., 1999; Lecun et al., 2015; Cho et al., 2005). The latter is notoriously inefficient mainly owing to the high development cost (approximately US$2.6 billion for each newly approved drug), low clinical trial success rate (less than 12%), and low return on investment (Wenz, 1982). Computer-aided drug development (design, screening, and testing) may reduce the costs and increase the success rate and investment return (Cheng et al., 2012; Lu et al., 2006; Jain, 2017; Lill and Danielson, 2011; Mcinnes, 2007). Since the 1990s, techniques such as homology modeling, molecular docking, quantitative structure-activity relationship (QSAR) modeling, and molecular dynamics (MD) simulation have been used to research on drug activity mechanisms (Capener et al., 2000; Cavasotto and Phatak, 2009; Edwards et al., 2016; Ewing et al., 2001; Gombar et al., 2004; Hartmman et al., 2013; Kitchen et al., 2004; Krieger et al., 2003; Kwon et al., 2007; Lampi et al., 2010; Morris and Lim-Wilby, 2008; Ohashi and Tanaka, 2010; Rapaport, 2004; Rapaport et al., 2002; Thangapandian et al., 2013; Zheng et al., 2013). Artificial-intelligence (AI)-based big data analyses and high-performance computations have greatly improved the efficiency of drug development, especially in the drug discovery stage. More and more pharmaceutical companies are investing in AI technology. Currently, the value of the medical AI market is approximately US$700 million and is expected to grow at a compound annual rate of 53%, reaching US$8 billion by 2022. More than 35% of this AI market share is taken by drug discovery (Clancey and Shortliffe, 1985; Sondak, 1990). To improve the efficiency of drug discovery and increase the success rate of drug synthesis, many ML companies become specialized in serving pharmaceutical companies; their services include disease target identification, compound screening, *de novo* drug design, clinical image recognition (Secco et al., 2016), toxicity prediction, and the prediction of absorption, distribution, metabolism, and excretion (ADME) (Hou and Xu, 2002; Kassel, 2004; Li, 2001; Yang et al., 2004). A drug repurposed by the AI company Benevolent (Stephenson et al., 2019), for example, is now in the second phase of clinical trials and is being tested by Johnson & Johnson. However, it is important to be aware of the pros and cons of different AI systems, as they are optimized for specific purposes. By investing in different AI systems, pharmaceutical companies not only can engage multiple areas ranging from drug discovery to clinical trials but also may discover breakthrough treatments for complex diseases.

[1]The Interdisciplinary Graduate Program in Integrative Biotechnology and Translational Medicine, Yonsei University, Incheon 21983, Republic of Korea

[2]Department of Biology, School of Life Sciences, Southern University of Science and Technology, 1088 Xueyuan Avenue, Shenzhen, Guangdong 518055, China

[3]Guangdong Provincial Key Laboratory of Computational Science and Material Design, Shenzhen, Guangdong 518055 China

[4]Guangdong Provincial Key Laboratory of Cell Microenvironment and Disease Research, Shenzhen, Guangdong 518055, China

[5]Shanghai Rural Commercial Bank Co., Ltd, Shanghai 200002, China

[6]Department of Physics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong, China

[7]School of Life and Health Sciences and Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen 518172, China

[8]Biotechnology, College of Life Science and Biotechnology, Yonsei University, Seoul 03722, Republic of Korea

[9]These authors contribute equally

*Correspondence: ktno@yonsei.ac.kr (K.T.N.), wanggy@sustech.edu.cn (G.W.)

https://doi.org/10.1016/j.isci.2021.103052

AI technologies are driven by new ML algorithms, advances in computational power, and ever-increasing experimental data (Bazoon et al., 2002; Pasquier and Hamodrakas, 2009). Advanced biotechnologies such as next-generation sequencing (Dijk et al., 2014; Kim, 2019), cryo-electron microscopy, high-throughput screening (Shin et al., 2016), medical digitization, and internet-of-things infrastructure (Adrian et al., 1984; Atzori et al., 2010; Cai et al., 2019; Dubochet et al., 1988; Feng et al., 2012; Gupta et al., 2009; Li et al., 2013; Song et al., 2019) provide high-quality Big Data for the application of ML-based drug development. ML algorithms can be roughly divided into three kinds: unsupervised, supervised, and reinforced (Ertel, 2017; Figueiredo and Jain, 2002; Kaelbling et al., 1996; Le, 2013; Piotr et al., 2006; Radford et al., 2015; Sahami, 1997; Sasakawa et al., 2010; Turian et al., 2010; Wang et al., 2019; Zhu and Goldberg, 2009). Unsupervised ML can be used to find hidden patterns in medical and biological data and to identify new disease targets (Bailey and Elkan, 1995; Wiskott and Sejnowski, 2002). Supervised ML can be used to predict drug activity, toxicity, and ADME from the existing data of drugs and clinical trials (Carneiro et al., 2007; Conneau et al., 2017; Igual and Seguí, 2017; Møller, 1993). Deep learning (DL), as advanced ML, has unprecedented power to scale up the capabilities of ML. AI is expected to become a major cost-effective, low-risk method in drug virtual screening.

QSAR, first established by Corwin Hansch (Hansch et al., 1962), was a natural extension of physical chemistry into the field of virtual drug screening. After more than 50 years of development with interdisciplinary breakthroughs and community promotion, QSAR has transformed from simple regression analysis (which can only handle similar compounds) to multiple statistics ML technique (which can analyze a very large data set of molecular structures). QSAR models have been widely used to model the biophysical properties of many chemicals (Hopfinger et al., 1997; Jaworska et al., 2005; Sabljić et al., 1995; Wold, 2010) and to assess the potential impacts of medicines, chemicals, and nanomaterials on human health and ecosystems (Alves et al., 2016; In et al., 2012; Karelson et al., 2010; Kim et al., 2015; Shin et al., 2017, 2018; Svetnik et al., 2003). For example, Lee et al. established a QSAR system called MS-HEMs to manage high-energy molecules (Lee et al., 2012). In the field of computer-aided drug design (CADD), QSAR has long been recognized as an effective method of structure- and knowledge-based drug design and optimization (Cramer, 2012). In this article, we review the merits, reliability, and limitations of QSAR to generate new insights into ML-based QSAR models, their integration with experimental or simulation data, and their potential applications (Zhao, 2003).

Structure-based drug design

The key to molecular targeted therapy is the discovery of lead drugs that can inhibit the targeted proteins, which usually entails the screening of a large number of small molecule compounds. Drug screening can be achieved by complex experiments (Larios et al., 2012; Lutz et al., 1996; Pan et al., 2005; Strasser et al., 2003), but their high cost and slow speed prevent high-throughput realization. CADD, through software such as GOLD, SYBYL, DiscoveryStudio, Autodock (Ash et al., 1997; Gao and Huang, 2011; Hashmi, 2007; Trott and Olson, 2010; Verdonk et al., 2010; Wang et al., 2015a; Yang et al., 2011), can rapidly search common libraries (e.g., proprietary libraries, Maybridge commercial library, and Food and Drug Administration drug library) and perform virtual screening (molecular docking and binding free energy evaluation) (Liu, 2016), which are characterized by high speed and high throughput. Compounds with high priority scores are selected for further screening by cellular experiments, animal experiments, or clinical trials, which may finally give rise to lead compounds.

CADD is a structure-based drug design method. Biological/chemical properties, no matter static properties (e.g., chemical composition) or dynamical changes (e.g., the conformational changes of DNA with different ethanol concentrations (Fang et al., 1999)), are fundamentally determined by molecular structure, which underscores the importance of studying structure-property relationship. Molecular mechanics (MM) (Humphrey et al., 1996; Phillips et al., 2010) and quantum mechanics (QM) (Csányi et al., 2004; Habasaki and Okada, 2006; Kwangho Nam and York, 2008; Mochizuki et al., 2007; Noel et al., 2010) are two approaches of MD aiming at the accurate computation of the moving trajectory of every atom in a biomolecule until the movement is stabilized, reaching the stable (low energy) conformation. Owing to the large number of atoms in a biomolecule, an MD simulation is usually computational costly. To expedite drug screening, less accurate but computationally efficient MM force field, such as Merck Molecular Force Fields (MMFF), are generally used (Bret et al., 2000; Cheng et al., 2000; Cieplak et al., 2009; Giese, 2005; Halgren, 2015; Tu and Laaksonen, 2001). Structure-based drug design mainly includes receptor-based methods (molecular docking, de novo design, MD simulation, homology modeling) (Cho et al., 2015; Nam et al.,

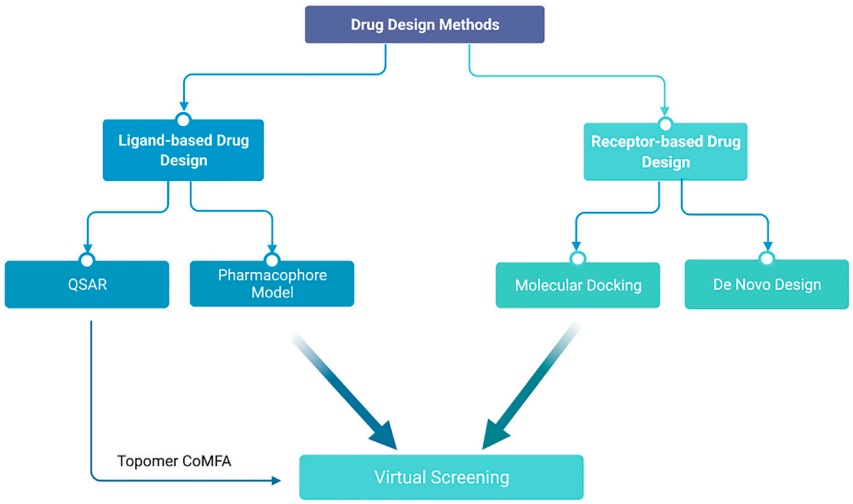

Figure 1. Classification of drug design methods

2003, Nam et al., 2011; Semper et al., 2021) and ligand-based methods (QSAR, pharmacophore, substructure search) (Anderson, 2003; Bohacek et al., 1996; Chang et al., 1992; Greer et al., 1994; Shim et al., 2014; Ma et al., 2012; Verlinde and Hol, 1994). They all involve the interaction between the receptor and ligand, but their focuses are different (Figure 1).

Receptor-based methods are mainly based on the three-dimensional structure of the receptor to find a matching ligand. To treat a disease by targeting some human receptor protein, a large number of small-molecule compounds need to be screened by molecular docking or de novo design to find the ones that fit well with the crystal structure of the protein (Butterfoss and Kuhlman, 2006; Degrado, 1997; Lichtenstein et al., 2012). If the crystal structure is unavailable, the structure can usually be estimated by homology modeling, namely, by constructing the human protein structure based on the corresponding structure of another species and the amino acid sequences of both (which only have small differences). Although receptor-based methods are dominant in drug screening, they still have great limitations. First, these methods rarely consider factors such as protein flexibility, the influence of water molecules, solvation effects, and the conformational limitations of the ligand. At present, most molecular docking procedures have certain defects such as the inability to correctly deal with the induced coincidence effect, the solvation effect, and the poor ranking ability of the scoring function. Second, although the efficiency of molecular docking is relatively high, its speed is far from enough for a million-level compound library. Finally, the crystal structure of some proteins has not been solved for any species, making ligand-based methods infeasible. For example, the membrane proteins are difficult to purify and crystallize owing to hydrophobicity.

Ligand-based methods can be used to circumvent the aforementioned problems because they are in principle independent of the receptor protein. They start from a known effective ligand to discover the substructures or structural characteristics that are genuinely responsible for the drug efficacy, which can then be used to guide the selection or design of the analogs of the ligand. In brief, the methods predict new ligands based on the features extracted from a known active ligand. These new ligands are candidate drugs and need to be tested further. Even if the crystal structure of the corresponding receptor is unknown for all the species, the ligands can still be tested by *in vitro* experiments (e.g., the measurement of binding free energy, pk(a), logP, logD, and other indicators (Bash et al., 2002; Culler et al., 1993; Laitinen et al., 2010; Xing and Glen, 2010)), making the ligand-based methods always useful. Ligand-based methods mainly include pharmacophore-based methods (Kim et al., 2017; Kurogi and Guner, 2001; Lee et al., 2013; Yang, 2010) and QSAR model prediction.

Pharmacophore

Pharmacophore refers to the common characteristics of all the small-molecule compounds actively interacting with the target protein. There must be specific sites in the receptor for small molecules to bind; thus,

the ligands binding to the same receptor must have similar chemical substructures. The International Union of Pure and Applied Chemistry (IUPAC) defines pharmacophore as "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response." The pharmacophore model uses not only the topology similarity but also the functional similarity of the molecule, allowing for the use of the concept of bioisosterism to make the model more reliable (Wolber and Sippl, 2008). In CADD, the pharmacophore model is mainly applied in three areas:

a) *Structure-activity relationship*, which is established through the discovery and definition of key pharmacodynamical characteristics of the drug molecules.

b) *Scaffold hopping*, which is established through the discovery of compounds with novel core structures by modifying the central core structure of a known active compound (Jang et al., 2016).

c) *Target fishing*, which is established through the prediction of targets of a given compound according to the compound's pharmacophore characteristics. This also allows for the identification of the potential off-targets (side effects) of the compound so that its candidacy as a lead compound may be eliminated early in time.

To obtain an accurate pharmacophore model, the correct three-dimensional (3D) structure of the compound must first be identified, which necessitates the careful inspection of valence, bond order, protonation state, tautomerism, and stereoisomerism. Although pharmacophore deserves further development, their practical application still faces many difficulties. A pharmacophore is receptor specific and may become useless when the drug target is different from the receptor, even if the difference is small. A pharmacophore is also sensitive to the change of its parental chemical structures. That is, a slightly different chemical may not match the pharmacophore. The matching difficulty is determined by the number of features and the tolerance of the pharmacophore.

Quantitative structure-activity relationship

QSAR is an empirical mathematical model in which regression (Doo Ho Cho and Bum Tae Kim, 2001) and classification (Choi et al., 2010; Choi et al., 2009; Kim et al., 2006a, 2006b; Kim et al., 2008; Lee et al., 2017; You et al., 2015) is performed on many structure-property data to reveal statistically significant correlations between chemical structures and biological properties (Figure 2). A QSAR model can thus predict a new chemical's biological/toxicological properties based solely on the chemical's structure, i.e., without resorting to the time-consuming molecular docking, which makes both the training and application of QSAR highly efficient. This does not necessarily mean that the receptor information cannot be incorporated. Because the structure of a protein is determined by its amino acid (AA) sequence, the structure and activity of the receptor can be taken into account by introducing the AA sequence information into the QSAR model, which may markedly increase the accuracy of model prediction. In summary, QSAR is promising in drug development because it can process a large amount of compounds with high speed and without losing much precision.

QSAR has received an increasing number of applications in recent years, including drug design, drug toxicity prediction (Wu and Wang, 2018), the study of enzymes' chemical-biological interaction mechanisms, and the prediction of compounds' biological activity. QSAR is generally based on the following three methods: molecular description, chemical similarity search, and ML. They are described in the following.

QSAR: methods based on molecular descriptors

The core of QSAR lies in the acquisition of molecular descriptors (also called chemical descriptors), that is, the extraction of information-rich numerical features from chemical structures (Joung et al., 2012; Kim et al., 2009; Nilakantan et al., 1987; Park et al., 2013; Randić, 1993). The past 50 years has witnessed different types of molecular descriptors, which can be divided into quantitative descriptors (molecular field, molecular shape) and qualitative descriptors (daylight fingerprints, MACCS keys, MDL, Public keys). According to the data type, molecular descriptors can be divided into Boolean (e.g., chiral or not), integer (e.g., the number of rings), real (e.g., molecular weight), vector (e.g., dipole moment), tensor (e.g., electronic susceptibility), scalar field (e.g., electrostatic potential), and vector field (e.g., electrostatic potential gradient). According to the physical meaning, molecular descriptors can be divided into composition descriptors,

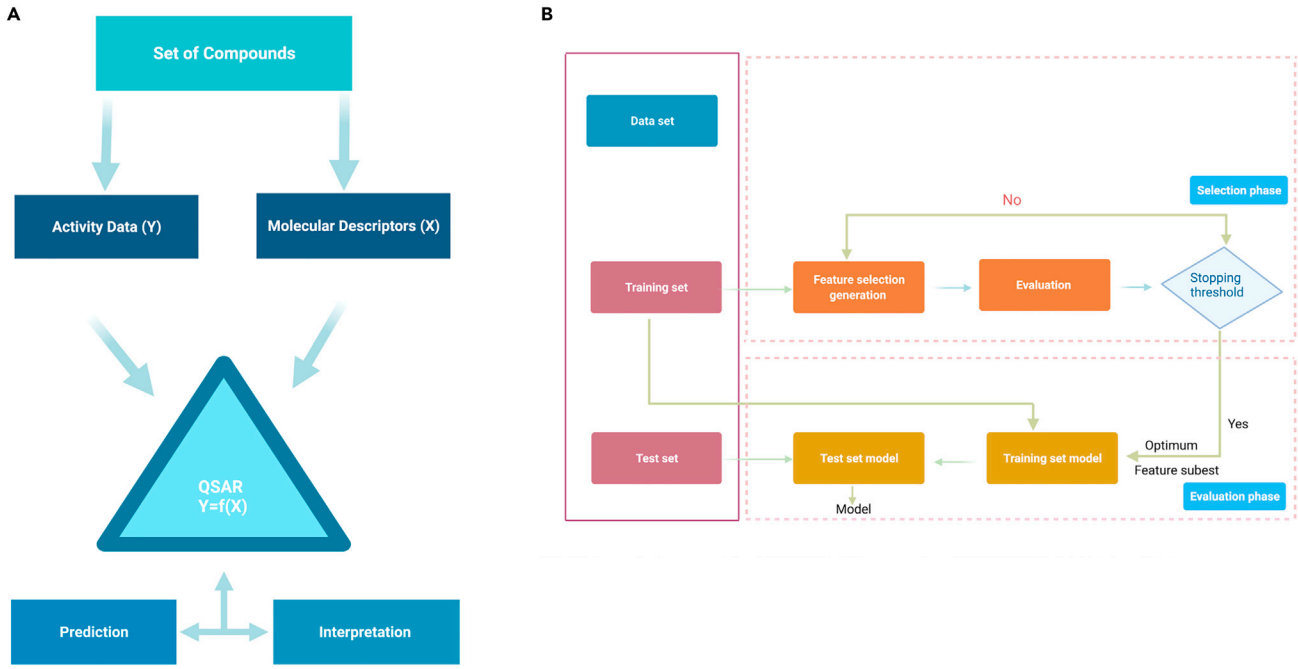

Figure 2. Quantitative structure-activity relationship

(A) The conception of QSAR.

(B) Understanding of QSAR from the perspective of machine learning.

molecular property descriptors, topological descriptors, and geometric descriptors. According to the structural dimension, molecular descriptors can range from 1D to 6D. Molecular descriptors can further be divided into experimental descriptors (logP, molar refractivity, dipole moment, polarizability, etc.) and theoretical descriptors (e.g., atom number, molecular weight, atom-type count, etc.). These different types of molecular descriptors are widely used in molecular data mining, compound diversity analysis, and compound activity prediction.

2D-QSAR

It is intuitive and concise to distinguish molecular descriptors according to the structural dimension (Azari and Iranmanesh, 2015; Morell et al., 2005; Ivanciuc and Braun, 2006; Karelson et al., 1996; Sandberg et al., 1998; Sheridan et al., 1996). One-dimensional descriptors are scalars that describe aggregated information, such as atom count, bond count, molecular weight, atomic properties, and fragment count (Shimamoto, 1999). Although simple, 1D descriptors are rather degenerative—disparate compounds may have the same value. Therefore, a 1D descriptor is usually used in addition to other high-dimensional descriptors, or combined with other 1D descriptors to form a vector.

Most molecular descriptors in the literature belong to 2D, which defines the connectivity of atoms in a molecule according to the properties of chemical bonds, including topological index, molecular profile, 2D autocorrelation, and chemical fingerprints (Hetényi et al., 2006). Two-dimensional descriptors generally deliver simple and useful structural information, which is rotation and translation invariant and is also invariant when the local structure is optimized. Another important invariance is graphic: the descriptor value does not change with the renumbering of the graphic nodes (vertices), which is very useful for structural differentiation. To facilitate the analysis of the large amount of 2D descriptors, Hong et al. reported the Mold^2 system, which can quickly generate 200 types of 2D descriptors for large-scale composite data sets (Hong et al., 2008). Other commonly used commercial software for the generation of 2D descriptors includes the dragon system, which can generate up to 5000 2D molecular descriptors (Mauri et al., 2006). More detailed information about 2D descriptors can be found in the studies by Durant et al., 2002; Duvenaud et al., 2015; Lo et al., 2018; Polanski, 2009; and Rogers and Hahn, 2010.

3D-QSAR

3D-QSAR facilitates the study of QSAR by introducing the 3D structural information of the drug molecule (Cho et al., 2012; Choi et al., 2006). Compared with 2D-QSAR, it has clearer physical meaning and richer information and can better reflect the nonbonding interaction characteristics between the drug and target molecules. Since 1980, 3D-QSAR has gradually replaced 2D-QSAR and become one of the main methods of mechanism-based rational drug design.

The 3D-QSAR chemical descriptors include autocorrelation descriptors, substituent constants, surface and volume descriptors, quantum chemical descriptors, and unique molecular scaffold (Todeschini and Consonni, 2009). These 3D chemical descriptors are useful to derive new drugs from an existing one through scaffold hopping. That is, replacing the existing drug's scaffold with others without significantly affecting the drug's binding activity. The chemical descriptors are obtained by extracting chemical features of the molecules from their conformations (Lipkowitz and Boyd, 2007), which can generally be obtained from crystal structures or from quantum chemistry computation by programs such as CORINA (Gasteiger et al., 1990). The criterion for the determination of molecular conformations is usually the lowest free energy. However, the actual conformation of the molecule corresponds to low energy but not necessarily the lowest one, and a molecule can switch between multiple conformations. Relevant experiments have confirmed that the *in vivo* active conformation of a drug molecule is generally in a low energy state but not the lowest. As such, it is not guaranteed that the predicted bioactivity can be verified by experiments. These complexities represent the key limitations of generating 3D chemical descriptors.

Commonly used 3D fingerprints include chemical characteristics based on pharmacological patterns, surface properties, molecular volume, and molecular interaction fields, of which the most famous one is based on molecular interaction field (MIF) (Hayakawa et al., 2020), which was realized by Goodford in the GRID program (Goodford, 1985). The MIF fingerprint is obtained by placing the ligand in a rectangular grid with fixed intervals to calculate the fingerprint characteristics of each grid point. In other words, the electrostatic force, 3D coordinates, hydrophilic (hydrophilic) and other dimensions are calculated independently for each grid point. The obtained MIF fingerprint can then be used in a specific 3D-QSAR model to predict the activity relationship between the molecules comprising the complex. More detailed information about 3D-QSAR models can be found in the studies by Baskin and Zhokhova, 2013; Cramer et al., 1988; Datar et al., 2006; Dixon et al., 2006; Gerhard Klebe and Mietzner, 1994; Gohlke and Klebe, 2002; Hopfinger, 1980; Low and Vinter, 2008; Simon et al., 1977; Varela et al., 2012; Veselovsky et al., 2001; and Walters and Hinds, 1994.

Higher dimensional QSAR

Three-dimensional QSAR is computational costly, and its performance is sensitive to the changes in the conformation/orientation of the ligand. To overcome these drawbacks, higher dimensional QSAR models have been developed. Four-dimensional QSAR has solved the conformation/orientation problem by simultaneously considering multiple structural conformations (Andrade et al., 2010). For example, Ash and Fourches calculated a 3D descriptor based on a 20-ns MD simulation trajectory of the atoms of Erk2 (Ash and Fourches, 2017). This 4D (3D space + 1D time) chemical descriptor can effectively distinguish the most active Erk2 inhibitors from those inactive and highly enriched Erk2 inhibitors. Five-dimensional QSAR takes into account factors such as receptor flexibility and inducible fit (Vedani and Dobler, 2002). Six-dimensional QSAR takes into account an additional factor, namely, the effect of solvation on the main receptor-ligand interaction (Vedani et al., 2005).

Search based on compound structure similarity

Chemical similarity search is one of the most popular techniques for drug discovery based on ligands (Willett et al., 1998). It aims to query similar compounds with known active molecules in terms of structure from a database. The basic assumption is that compounds that are similar in function have similar chemical structures. However, this assumption is not always accurate. For example, the "active cliff" effect (Hu and Bajorath, 2020), a situation in which a small modification to a functional group leads to a sudden change in activity, does not meet the aforementioned assumptions and may lead to failure. To evaluate the structural similarity between two molecules, the Tanimoto coefficient (Tc, known as the Jaccard index), a measure of the similarity between two compounds, was proposed (Lipkus, 1999). A higher Tc score indicates that two compounds are more similar, but Tc does not provide details regarding what kind of chemical group they share. More detailed information about 3D chemical similarity can be found in the studies by Bero et al., 2018; Cheeseright et al., 2008; Dossetter et al., 2013; Ferreira and Couto, 2010; Lo et al., 2016; and Rush et al., 2005.

Further considerations

The success of QSAR lies in the fact that the difference in the representation of the ligand structure truly represents the essential difference in the scope of the ligand. For example, the difference in bioactivity of two compounds is solely determined by their structural difference, no matter how complex the subsequent physicochemical and biological interactions are. The ultimate goal of ligand comparison is to maximize the space of positive and negative samples. If the developed QSAR models are expected to be useful for drug synthesis in addition to drug screening, properties need to be considered such as synthesizability, hydrophobicity, drug-likeness, Lipinski rule, and false-positive issues (Alam and Khan, 2017).

RESULTS

QSAR models based on ML

The ML-based QSAR model currently plays a notable role in drug design and screening, property prediction, category prediction, etc. Since the 1990s, ML techniques such as support vector machines (SVMs) and random forests (RFs) have been widely used to discover or design new medicines. For example, its application to a scoring function based on molecular docking has been reported (Lima et al., 2016). Data mining based on chemical graphs can derive a set of two-dimensional or three-dimensional chemical descriptors, which are packaged into chemical fingerprints in various ML models and prediction tasks. A key innovation in this field is the combination of big data and ML, including the mining of gene sequences and protein structure, single-cell sequences, multi-omics interaction data, and the interaction between proteins and genes, which enables the QSAR model to learn and predict a wider range of biological laws.

In the past, QSAR models focused only on the predictive ability of candidate compounds; thus, the commonly used ML models are discriminative prediction models. ML methods are mainly divided into discriminant prediction models and generative models (Jebara and Pentland, 2001). The essence of the discriminative model is to learn the mapping relationship between the characteristics of the characterizing sample and the prediction target from the data, and it is related to pattern recognition, which means that the model can only discriminate and predict. The generative model directly learns the distribution of data from the sample. In addition to being combined with the discriminant model, the generative model can also directly generate new samples based on the learned distribution and is thus used to generate candidate compounds. The methods to learn patterns hidden in data are mainly divided into supervised learning, semisupervised learning, unsupervised learning, and self-supervised learning. Because the QSAR model focuses only on the predictive ability of a model for candidate compounds, we introduce only the discriminant models commonly used under supervised learning. The key to this type of model is the selection of features, the selection of the algorithm itself, and the quality of the training set. Most aspects of feature construction have been included in the chemical descriptor section of the previous chapter. Examples of the extracted features can be found in the literature (e.g., Tomal et al., 2016).

Conformation-related indicators have been taken into account in the QSAR, and conformational calculations for isomers can be calculated using quantum chemistry, semiempirical methods, and molecular dynamics simulations. Because the relative stability of conformational isomerism is related to toxicity detection, the conformation of the compound may change as the solvent environment changes, which affects the three-dimensional correlation characteristics and then impacts the final screening result. Therefore, the methods of calculating the isomer conformations can be used to predict whether the relative stability of the conformation will affect the experimental results. In addition, the ML-based QSAR method can be applied to examine the energy changes between conformational changes, to determine how the conformation is affected by the solvent environment, and to predict the relationship between conformational changes and molecular energy (Lokuge et al., 2010; Singh et al., 2016).

The importance of these features varies in different structure-activity or attributes prediction tasks, but ML can learn from the data which features are more important. Therefore, we need only to exhaust all the features related to the nature of QSAR tasks as much as possible and apply the traditional ML algorithm to learn from the training set.

ML-based QSAR methods are effective under the usual condition that the system under study is so complex that it cannot be physically modeled. Indeed, many physical and chemical properties are difficult to calculate by theoretical methods such as density functional theory (DFT) or MD, but they can usually be computed by cheminformatics models, particularly the water solubility and logP (the logarithm of octanol:

water partition coefficient), which are directly related to drug activity. The properties that are indirectly related to drug activity, such as the melting point, solubility, and sublimation energy, can also be calculated by cheminformatics. Of course, the QSAR methods may not be suitable when the underlying mechanisms are sufficiently clear. For example, quantities such as dipole moment, polarizability, and vibration frequency can be directly computed by quantum chemistry.

The characteristics of a compound in the traditional QSAR model are produced in an ideal environment. In reality, however, the solvent environment is not ideal. For example, we know that acidic chemicals are easily absorbed in an acidic environment. Therefore, to achieve the desired effect for a drug in an actual environment, the hydrophilicity and lipophilicity of the drug need to be considered (Ditzinger et al., 2019). The main environmental factors of a solvent can be found in the study by Bruno et al., 2021.

With the aforementioned preparation, the next stage is to build a QSAR model based on ML, which mainly consists of the following steps:

(1) "Molecular coding," which uses a vector of various characteristics to represent compound molecules, where the chemical characteristics and properties can be learned from the chemical structure or experimental data. The molecular characteristics should be independent as much as possible to avoid excessive correlation. The biological activity data and solvent environmental coefficients should be as accurate and sufficient as possible, and they are best obtained by conventional experiments.

(2) An appropriate number of compounds must be chosen to construct the training/testing sets. Ideally, conformation optimization through MM/QM is carried out first to improve the data quality. Unsupervised learning is used to identify the most relevant attributes and reduce the dimensionality of the feature vector. Supervised ML model is applied to discover an empirical function (explicit or implicit) that can achieve the optimal mapping between the input feature vector and the biological activity.

(3) Internal/external verification of the model is performed to determine its applicability and predictability. After generating exhaustive feature sets, the algorithm determines which feature combinations can be used for new predictions by finding the largest difference between positive and negative samples in the training set; thus, similar data distributions between the training and testing sets should exist. The techniques to detect out of distribution (OOD) can be found in the study by Hendrycks and Gimpel, 2016. The key to constructing a ML model is to constantly add and try more feature combinations, change the divisions of the training set, and adjust the algorithm parameters until the best model is obtained.

ML models trained on specific target sets are not universal, but the latest ML methods are improving versatility. Advances in computing power and software performance have also been used to improve QSAR models. By continuously integrating new algorithms and descriptive features, the model can be continuously optimized. Although the general idea of the ML-based QSAR is the same, the input characteristics may be different for different targets. The latest progress is the ability to learn general representations, such as self-learning, multitask learning, AutoEncoder, Generative Adversarial Networks (GAN), and Bidirectional Encoder Representations from Transformers (BERT) (Caruana, 1997; Devlin et al., 2019; Goodfellow et al., 2014; Liou et al., 2014; Xu et al., 2020). The QSAR learning models still face several problems beyond their construction (Dearden et al., 2009), which are detailed in the study by Artem et al., 2014.

DL invigorates QSAR

With excellent performance in imaging, speech, machine translation, etc (Minar and Naher, 2018), DL has entered into many biological fields, including genes, proteins, metabolites, microbiomes, and population-wide genetic variation, synthetic biology, drug discovery, and diseases (Alkawaa et al., 2018; Golkov et al., 2020; Hill et al., 2018; Zeng et al., 2021). The promising DL methods include capsule networks (Inokuma et al., 2010; Xi et al., 2017), multitask learning (Antropova et al., 2017; Wang et al., 2015b; Zhu et al., 2016), GANs, self-encoding decoders (Marchi et al., 2015; Wang et al., 2018; Xu et al., 2014; Yao et al., 2017; Zhao et al., 2016), Variational AutoEncoders (VAEs) (Panych et al., 2015), Long Short Term Memory Networks (LSTMs) (Baytas et al., 2017; Gers and Schmidhuber, 2001; Graves et al., 2005; Yildirim, 2018), transfer learning (Fernandes et al., 2017; Pan and Yang, 2010; Paul et al., 2016; Zoph et al., 2016), deep

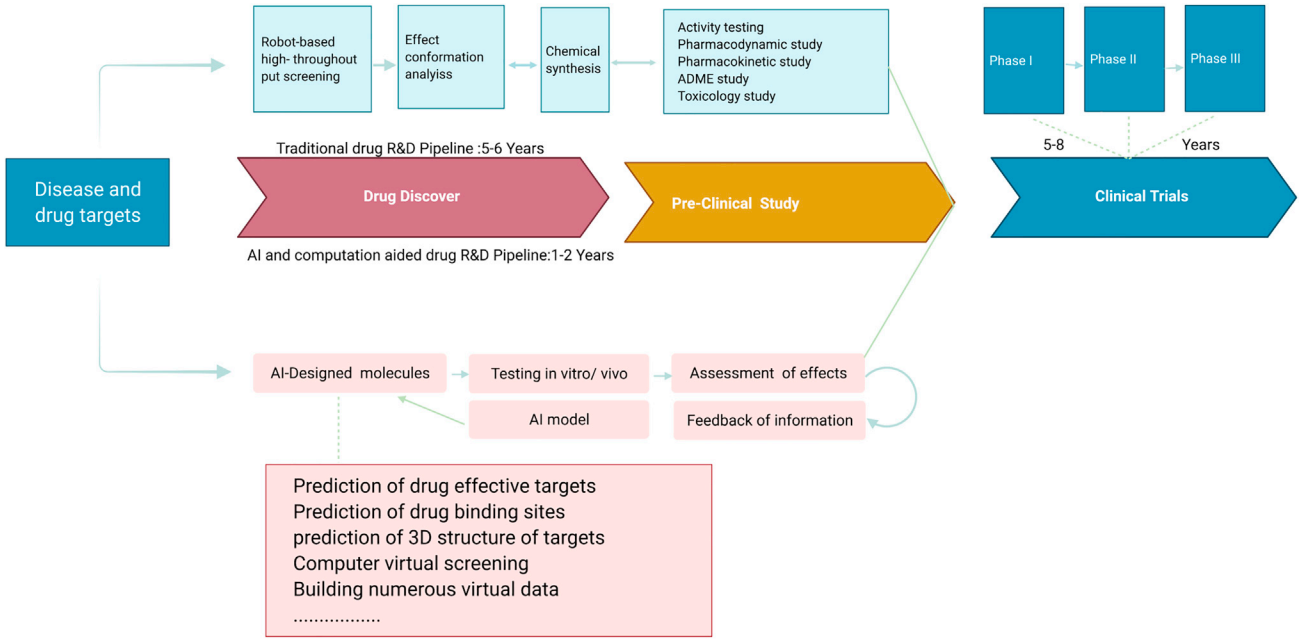

Figure 3. Main steps and scenarios of ML-based drug discovery

neural networks (DNNs) (Yoshioka et al., 2014), and CNNs (Horváth et al., 2017; Luo, 2015; Parashar et al., 2017; Wang, 2013; Xue et al., 2016). Because the biological information is generally rich, multitask learning can be performed to gain a better structural representation. For example, the microbial metabolic network and the loop network formed by the parasitic relationship are very scarce; they can be generated by a GAN. When using DL to predict the protein structure from a gene sequence, VAE can be used to encode compound structures and then be combined with the GAN for learning (Lin, 2009; Widera, 2010; Yu et al., 2015). With switch networks formed by atoms, strategy networks and Monte Carlo trees can be used to predict chemical reactions (Anderson and Long, 2003). CNN can be used to perform learning based on force fields and the protein-ligand complex (Clore and Gronenborn, 2010). A matching network, which is a variant of one-hot learning (Alaya et al., 2019), can be used to evaluate whether a new compound matches the receptor (Deka and Quddus, 2014; Tan et al., 2013). The cheminformatics-based LSTM methods are also frequently reported (Shu et al., 2018). The advantages, notwithstanding, DL entails a large amount of data and computation and is essentially a "black-box"; thus, it cannot replace all traditional ML methods, which are currently irreplaceable owing to better interpretability and lower data requirements.

Biological applications of DL

Since the late 1990s, DL has become a useful tool for drug discovery and can help researchers understand and construct the relationship between a chemical structure and its biological activity (Jing et al., 2018) (Figure 3). In the early days, structure-based methods such as cheminformatics are most successful in drug discovery (Feng et al., 2013; Hwang et al., 2020; Kim et al., 2010; Nam et al., 2014). They were used to conduct chemical structure searches, generate descriptors, construct fingerprints, and analyze chemical similarity. Expert experiences, combined with manually generated explanatory rules, were used to screen compounds. The later DL can capture the nonlinearity and complex relationships in the available data and generate more compact features than those manually generated by experts. Through multilayer networks, DL can directly process raw, unstructured data such as sequences and three-dimensional structures and has better nonlinear fitting capabilities. These multilayer networks are proven to have infinite VC dimensions (Vapnik et al., 1994) and thus can divide any data space. DL has been successfully applied to CADD (Chen, 2014; Quang et al., 2015). Although DL is more suitable for ligand-based drug discovery, there exist several interesting structure-based DL applications. In the AtomNet method (Wallach et al., 2015), for example, the input molecule is discretized into a three-dimensional grid and then sent to a convolutional neural network (CNN) and a fully connected layer to predict the binding affinity. This model needs not to preprocess but uses a learnable representation to identify the pair data of interacting atoms. AtomNet uses

the same atom space to describe both the receptor and ligand, which is more natural, requires no preprocessing, and allows for the characterization and regularization of the interaction between the receptor and ligand atoms. To apply AtomNet, the 3D protein-ligand complex needs some conversion and coding (Klebe, 2000). First, the complex is formalized as a cube with 20 angstrom side length, and the cube is placed at the geometric center of the ligand. Then, a 3D grid with 1-angstrom resolution is used to discretize the positions of heavy atoms. This method allows for the representation of an atom as a 4D vector (three coordinates and one feature value) and the whole input as a 4D tensor. As another example, the protein-ligand complexes in the PDBbind database were used to train and test the neural network (Evers et al., 2003). The complexes were protonated and charged using UCSF Chimera with Amber ff14SB for standard residues and AM1-BCC for nonstandard residues and ligands. This study can determine which atom deletion causes the largest decrease in the predicted value. It also found novel interactions between the receptor residues and the ligand: e.g., Tyr693 forms a hydrogen bond with the ligand; Met713 forms a hydrophobic interaction with the ligand.

Another important biological application of DL is the clarification of the molecular mechanisms of how protein structural changes and amino acid mutations cause the change of protein-protein interaction (PPI) networks. In a DL application, instead of using the traditional MM/QM (Vesely, 2001), the protein structure is first divided into fragments and partitions, and then the most effective fragments are predicted. DL can, on the one hand, seek interpretability from the successfully trained model, i.e., the interpretation of the feature contribution of a single sample; the combination of existing field experiences then helps us gain insights into complex relationships. On the other hand, DL can combine raw data to generate specific functional characterization and fragment representation; thus, it can facilitate the study of the mechanism of a protein structure.

Comparing DL with traditional ML

DL and traditional ML have their respective advantages and application suitability (Camacho et al., 2018; Chan et al., 2019; Pei et al., 2019; Yuan et al., 2019). The traditional ML relies more on the domain knowledge (theories and mathematical models), while DL directly extracts patterns, which are not always explainable, from the raw data. For tasks such as image analysis, DL is certainly better; it is particularly useful in *de novo* molecular design and reaction prediction (Button et al., 2017; Hartenfeller et al., 2012; Kang and Liu, 2021; Schneider and Schneider, 2018). For tasks with structured descriptors as input, DL seems to be at least as effective as traditional ML methods. Although DL can achieve better performance in biological activity prediction through multitasking, ML-based QSAR can also achieve better performance through continuous improvement of the feature engineering, which can be achieved by dividing the protein into fragments and functional regions to remove the influence of the orientation, flexibility, and solvent of a protein molecule. Although DL is better for automatic feature engineering, it entails a large amount of data. In contrast, ML can use a small amount of training data (Altae-Tran et al., 2016).

One common ground of DL and traditional ML is the learning algorithms, including supervised and unsupervised learning. Supervised learning is related to a specific task and thus can obtain task-related feature representations. Its input data must be of a similar structure. If a new task is related to the original one, one can generally use the representations obtained by supervised learning to join the new task to improve the effectiveness of the model. Unsupervised learning can obtain a data-distributed representation of its own existence, which is irrelevant to a specific task. It yields the patterns themselves, that is, the general characteristics of the data set. The use of unsupervised learning is promising to create new synthetic biology samples and then to determine the legibility of the new samples based on the structural knowledge. For example, one can use the general common-sense model learned by unsupervised learning to check whether a synthesized new compound is reasonable. Furthermore, unsupervised training can obtain some general common sense. If this common sense is a prerequisite for certain tasks, then the representations learned can be used to determine whether a sample belongs to this pattern. Multitask learning is a kind of supervised learning with some unsupervised features. It has great advantages when the data are scarce, which is typical in the study of the parasitic relationship between the host and microorganisms in a microbiome.

Combining DL with traditional ML

Although traditional ML and DL have their own advantages and disadvantages, they can be used in combination. For example, the DL-based face recognition is apparently better, but the learned features may be meaningless and the interpretability is thus lacking. This shortcoming can be compensated by the

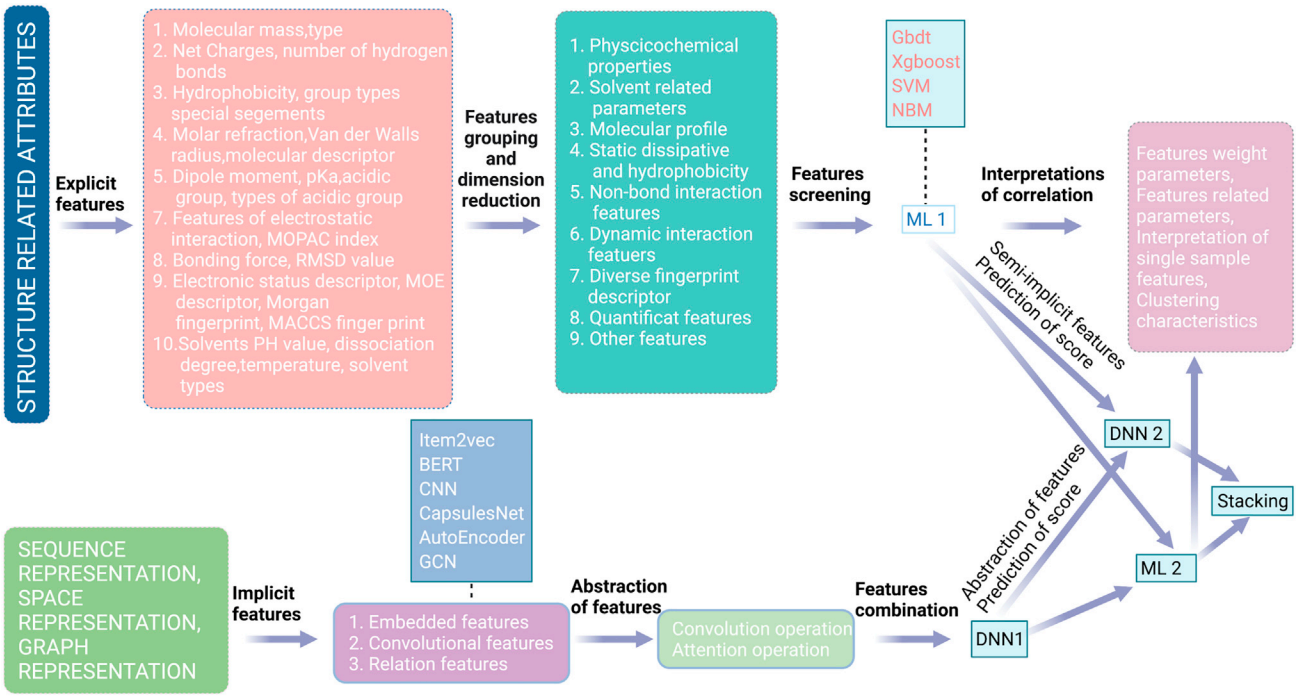

Figure 4. ML-DL combined QSAR model

ML-based recognition. Therefore, their combination can achieve a better performance. In the following, such a combined model is outlined (Figure 4).

Data-oriented feature extraction. Different models have different requirements of the input data. To implement traditional ML models, the input data should be well prepared to reduce feature dimension and promote feature grouping; the interactions between the features need to be constructed manually, which relies heavily on experience. To implement DL models, raw data are welcome because DL algorithms can automatically extract features. In image recognition, for example, DL can extract hierarchical features from the pixel-level raw data through its convolution kernel. Feature extraction has two types: explicit and implicit. The explicit features, such as chemical descriptors, chemical fingerprints, physical and chemical properties, chargeability, substituent groups, fragments, geometric properties, are usually generated by experts. Being overlap with each other or redundant, these features may need to be processed by applying algorithms such as restricted Boltzmann machines (Pinaya et al., 2016; Zhang et al., 2015) and AutoEncoder to generate new, more compact, distributed features. The implicit feature extraction is very useful in processing sequences such as amino acid sequences, the atomic sequence of a compound, the 3D structure of a protein, and atomic grid structures, where the features can be learned by directly applying sequence processing algorithms including CNN, LSTM, and automatic codecs. For other applications, algorithms such as word2vec (Altszyler et al., 2018; Wang et al., 2016), BERT (Sun et al., 2019), GANs (Yi et al., 2019), and AutoEncoder can be used for unsupervised training.

The dimension of features and the number of data samples must be well balanced. A large feature dimension necessitates the acquisition of more data, which may be costly or even impossible. To solve this problem, the usual method is to filter the features, group them, or apply AutoEncoder or VAE for dimensionality reduction. However, this method usually cannot reduce the amount of data because both unsupervised feature expressions and supervised learning are involved. We therefore propose the following ML-DL combined strategy. By using the traditional ML, various explanatory features related to the training target are first constructed and grouped based on common underlying information; dimensionality reduction is then performed by applying feature selection or AutoEncoder, and the compact and effective explicit features are obtained. By using DL, the other types of data (e.g., sequence and structure data) can be used for unsupervised training to obtain compact and implicitly distributed features. Finally, both the implicit features

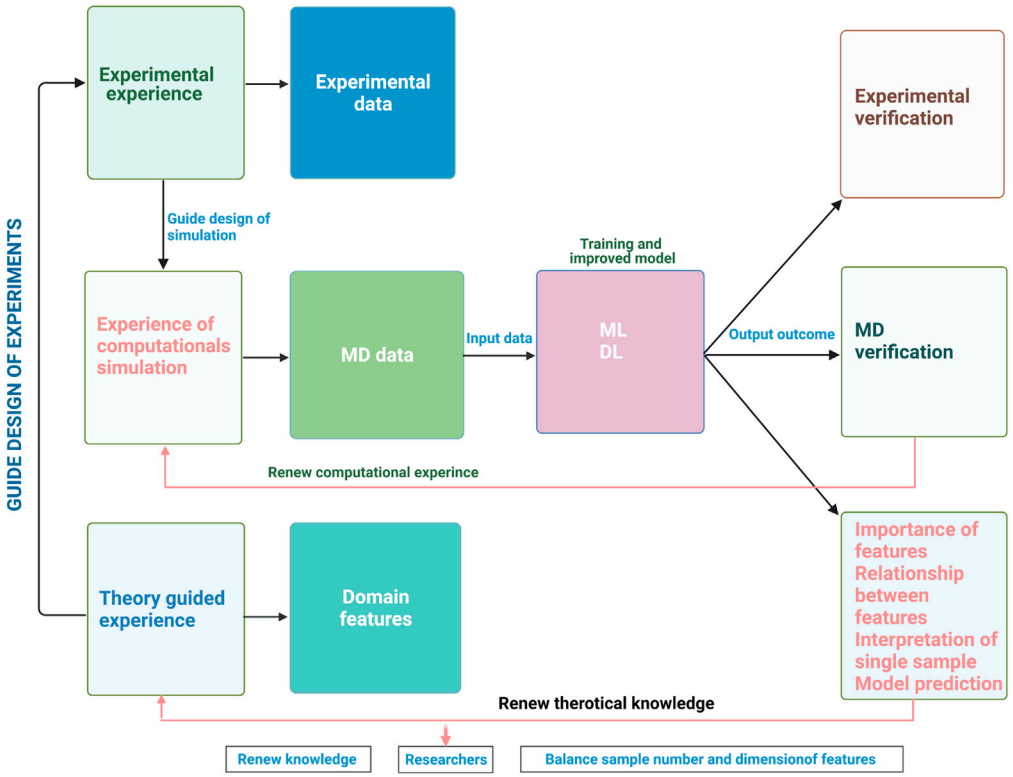

Figure 5. Iterative integration of wet experiments, MD simulation, traditional ML, and DL

(which reflect more information about the low-level data) and explicit features (which are more relevant to the task) are sent to the DNN to train the model.

Iterative training to generate more effective features. The explicit features of chemicals, usually in the form of molecular descriptor and fingerprint, are extracted from the structural chemical data and are grouped according to their correlation values. The redundancy of features is reduced by filtering out invalid features by using traditional ML algorithms such as xgboost and gbdt. These steps are iterated several times to determine the features to be input to a module called stacking (Palangi et al., 2014; Zhan et al., 2018; Zhang et al., 2017). The input samples to stacking include both chemical data and the features (sample = data + feature). The samples are divided into N parts, $N-1$ of which are used to train a classifier and the remaining one is used to test the prediction made by the trained classifier. In total, N such classifiers are used to obtain N new sets of features, which replace the old sets of features in the next round of iteration.

On the other hand, the implicit features of chemicals are extracted from the nonstructural chemical data such as the chemical sequence, molecular bond, 3D coordinates by using DL methods such as transformer, CNN, and GNN. The samples are then trained by a DNN, whose last layer output is saved and then, together with the above explicit features, is sent to a stacking module consisting of DNN and ML. Through training by the stacking module, a more effective model is obtained.

Iterative integration of different models. QSAR is based on both experience-driven and data-driven research models, including wet experiments, MD simulation, ML, and DL. Their cooperation makes QSAR increasingly powerful. The experiments usually do not reveal the reaction mechanisms at the molecular and atomic levels, which can be revealed to a certain extent through MD simulation. ML-based QSAR is data driven (good at processing 2D and 3D data) and is oriented by feature extraction. The limitations of traditional ML are that the data are usually low-dimensional and the features must be interpretable and manually constructed. To overcome these limitations, DL is developed to automatically extract features from both unstructured and structured data (which could be high-dimensional). The integration of the four models would make QSAR powerful, for which a framework is provided (Figure 5). In the framework,

the four models communicate with each other and adjust themselves iteratively to optimize the synergism of our field experience with ML/DL predictions.

There are two research studies that have implemented part of the framework in Figure 5. Yasushi Okuno's group at Kyoto University proposed an interdisciplinary framework called DEFMap, which combines experimental data, MD simulation, and DL to extract protein dynamic information from cryo-EM density maps (Matsumoto et al., 2021). Shuguang Yuan's group used part of the framework to prove that the frozen structure analyzed in the traditional artificial cell membrane environment is very different from the real cell membrane environment (Zhang et al., 2021), subverting the traditional understanding that the 3D structure of the membrane protein is the same as its physiological state in the artificial cell membrane or precipitant environment.

Conclusions

As a part of Big Data science, ML and DL are based on information theory and data fitting theories, where predictions depend mainly on a large amount of effective and reliable data, high-speed computation (for MM/QM simulations to process experiment data such as those from NMR, X-ray, and cryo-EM), and the number of compounds. These requirements cannot be met in many areas of computational biology; thus, ML/DL does not always work well. This problem can be partly solved in some fields by using powerful public tools such as AlphaFold (Senior et al., 2019). The wide application of ML in many fields depends on its human-like problem-solving capacity, including classification and clustering, explanation and verification, relevance and factor importance. But now the integration of iterative thinking is ever-increasingly important. Only through the cyclical upward process of prediction-verification-update can we form a systematic way of thinking and solve the problem from an overarching perspective.

In this article, we primarily reviewed ML-/DL-based QSAR methods. In addition to traditional descriptors, we emphasized the integration of MD with experimental solvent-related data. We also proposed a better framework that takes advantage of the respective merits of ML and DL to process different forms of data. The framework, through iteratively adjust itself, can potentially reveal the importance of some manual features, the correlation between features, and explain the contribution of different features in a single sample. The obtained embedded features or distributed vectors can help to understand the relationship between the constituent elements and the importance of local structures. The framework would ultimately achieve high efficiency and low cost in the fields of biomedicine and materials, especially in terms of QSAR.

ACKNOWLEDGMENTS

This work was partly supported by the National Natural Science Foundation of China (61773196, 32070681), Guangdong Provincial Special Projects on COVID-19 (2020KZDZX1182), Guangdong Provincial Key Laboratory Funds (2019B030301001, 2017B030301018), Shenzhen Research Funds (JCYJ20170817104740861), Shenzhen Peacock Plan (KQTD2016053117035204), and by the Center for Computational Science and Engineering of Southern University of Science and Technology. The kind help of Prof. No's group is acknowledged.

AUTHOR CONTRIBUTIONS

J.M., J.A, K.T.N., X.Z., and G.W. conceived the idea, wrote the article, and prepared the figures. All the authors discussed the results and commented on the manuscript.

DECLARATION OF INTERESTS

The authors declare they have no conflict of interest.

REFERENCES

Adrian, M., Dubochet, J., Lepault, J., and Mcdowall, A.W. (1984). Cryo-electron microscopy of viruses. Nature *308*, 32–36.

Alam, S., and Khan, F. (2017). 3D-QSAR studies on maslinic acid analogs for anticancer activity against breast cancer cell line MCF-7. Sci. Rep. *7*, 6019.

Alaya, M.Z., Bussy, S., Gaïffas, S., and Guilloux, A. (2019). Binarsity: a penalization for one-hot encoded features. J. Mach. Learn. Res. *20*, 1–34.

Alkawaa, F.M., Chaudhary, K., and Garmire, L.X. (2018). Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. J. Proteome Res. *17*, 337–347.

Altae-Tran, H., Ramsundar, B., Pappu, A.S., and Pande, V. (2016). Low data drug discovery with one-shot learning. ACS Cent. Sci. *3*, 283.

Altszyler, E., Sigman, M., and Slezak, D.F. (2018). Corpus specificity in LSA and Word2vec: the role of out-of-domain documents. RepL4NLP (Association for Computational Linguistics), 1–10.

Alves, V.M., Muratov, E.N., Capuzzi, S.J., Politi, R., Low, Y., Braga, R.C., Zakharov, A.V., Sedykh, A., Mokshyna, E., Farag, S., et al. (2016). Alarms about structural alerts. Green. Chem. *18*, 4348–4360.

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

Anderson, J.B., and Long, L.N. (2003). Direct Monte Carlo simulation of chemical reaction systems: prediction of ultrafast detonations. J. Chem. Phys. *118*, 3102–3110.

Andrade, C.H., Pasqualoto, K.F.M., Ferreira, E.I., and Hopfinger, A.J. (2010). 4D-QSAR: perspectives in drug design. Molecules *15*, 3281–3294.

Antropova, N., Huynh, B., and Giger, M. (2017). Multi-task learning in the computerized diagnosis of breast cancer on DCE-MRIs. arXiv, arXiv:1701.03882.

Ash, J., and Fourches, D. (2017). Characterizing the chemical space of ERK2 Kinase inhibitors using descriptors computed from molecular dynamics trajectories. J. Chem. Inf. Model. *57*, 1286–1299.

Artem, C., Muratov, E.N., Denis, F., Alexandre, V., Baskin, I.I., Mark, C., John, D., Paola, G., Martin, Y.C., Roberto, T., et al. (2014). QSAR modeling: where have you been? Where are you going to? J. Med. Chem. *57*, 4977.

Ash, S., Cline, M.A., Homer, R.W., Tad Hurst, A., and Smith, G.B. (1997). SYBYL line notation (SLN): a versatile language for chemical structure representation. J. Chem. Inf. Comput. Sci. *37*, 71–79.

Atzori, L., Iera, A., and Morabito, G. (2010). The internet of things: a survey. Comput. Netw. *54*, 2787–2805.

Azari, M., and Iranmanesh, A. (2015). Edge-Wiener descriptors in chemical graph theory: a survey. Curr. Org. Chem. *19*, 219–239.

Bailey, T.L., and Elkan, C. (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Mach. Learn. *21*, 51–80.

Bash, P.A., Field, M.J., and Karplus, M. (2002). Free energy perturbation method for chemical reactions in the condensed phase: a dynamic approach based on a combined quantum and molecular mechanics potential. J. Am. Chem. Soc. *109*, 8092–8094.

Baskin, I.I., and Zhokhova, Nelly I. (2013). The continuous molecular fields approach to building 3D-QSAR models. J. Comput. Aided Mol. Des. *28*, 427–442.

Baytas, I.M., Xiao, C., Zhang, X., Wang, F., Jain, A.K., and Zhou, J. (2017). Patient Subtyping via Time-Aware LSTM Networks (SIGKDD), pp. 65–74.

Bazoon, M., Stacey, D.A., Cui, C., and Harauz, G. (2002). A hierarchical artificial neural network system for the classification of cervical cells. ICNN *94*, 3525–3529.

Bero, S., Muda, A., Choo, Y.-H., Muda, N., and Pratama, S. (2018). Weighted Tanimoto coefficient for 3D molecule structure similarity measurement. arXiv, arXiv:1806.05237.

Bohacek, R.S., Mcmartin, C., and Guida, W.C. (1996). The art and practice of structure-based drug design: a molecular modeling perspective. Med. Res. Rev. *16*, 3.

Bret, C., Field, M.J., and Hemmingsen, L. (2000). A chemical potential equalization model for treating polarization in molecular mechanical force fields. Mol. Phys. *98*, 751–763.

Bruno, C.D., Harmatz, J.S., Duan, S.X., Zhang, Q., Chow, C.R., and Greenblatt, D.J. (2021). Effect of lipophilicity on drug distribution and elimination: influence of obesity. Br. J. Clin. Pharmacol. *87*, 3197–3205.

Butterfoss, G.L., and Kuhlman, B. (2006). Computer-based design of novel protein structures. Annu. Rev. Biophys. Biomol. Struct. *35*, 49.

Button, A.L., Hiss, J.A., Schneider, P., and Schneider, G. (2017). Scoring of de novo designed chemical entities by macromolecular target prediction. Mol. Inf. *36*, 1600110.

Cai, X., Zheng, W., and Li, Z. (2019). High-throughput screening strategies for the development of anti-virulence inhibitors against staphylococcus aureus. Curr. Med. Chem. *26*, 2297–2312.

Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., and Collins, J.J. (2018). Next-generation machine learning for biological networks. Cell *173*, 1581–1592.

Capener, C.E., Shrivastava, I.H., Ranatunga, K.M., Forrest, L.R., Smith, G.R., and Sansom, M.S.P. (2000). Homology modeling and molecular dynamics simulation studies of an inward rectifier potassium channel. Biophys. J. *78*, 2929–2942.

Carneiro, G., Chan, A.B., Moreno, P.J., and Vasconcelos, N. (2007). Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. Pattern Anal. Mach. Intell. *29*, 394–410.

Caruana, R. (1997). Multitask Learn. Mach. Learn. *28*, 41–75.

Cavasotto, C.N., and Phatak, S.S. (2009). Homology modeling in drug discovery: current trends and applications. Drug Discov. Today *14*, 676–683.

Chan, H.C.S., Shan, H., Dahoun, T., Vogel, H., and Yuan, S. (2019). Advancing drug discovery via artificial intelligence. Trends Pharmacol. Sci. *40*, 592–604.

Chang, S.-K., Jang, M., Han, S.Y., Lee, J.H., Kang, M.H., and No, K.T. (1992). Molecular recognition of butylamines by Calixarens-based ester ligands. Chem. Lett. *21*, 1937–1940.

Cheeseright, T.J., Mackey, M.D., Melville, J.L., and Vinter, J.G. (2008). FieldScreen: virtual screening using molecular fields. Application to the DUD data set. J. Chem. Inf. Model. *48*, 2108–2117.

Chen, Y. (2014). Machine Learning for Large-Scale Genomics: Algorithms, Models and Applications (UC Irvine).

Cheng, A., Best, S.A., Jr, K.M.M., and Reynolds, C.H. (2000). GB/SA water model for the Merck molecular force field (MMFF). J. Mol. Graph. Model. *18*, 273–282.

Cheng, T., Li, Q., Zhou, Z., Wang, Y., and Bryant, S.H. (2012). Structure-based virtual screening for drug discovery: a problem-centric review. AAPS J. *14*, 133–141.

Cho, S.G., No, K.T., Goh, E.M., Kim, J.K., Shin, J.H., Joo, Y.D., and Seong, S.Y. (2005). Optimization of neural networks architecture for impact sensitivity. Bull. Korean Chem. Soc. *26*, 399–408.

Cho, N.C., Cha, J.H., Kim, H., Kwak, J., Kim, D., Seo, S.H., Shin, J.S., Kim, T., Park, K.D., Lee, J., et al. (2015). Discovery of 2-aryloxy-4-amino-quinazoline derivatives as novel protease-activated receptor 2 (PAR2) antagonists. Bioorg. Med. Chem. *23*, 7717–7727.

Cho, Y.S., No, K.T., and Cho, K.H. (2012). yaInChI: modified InChI string scheme for line notation of chemical structures. SAR QSAR Environ. Res. *23*, 237–255.

Choi, S.Y., Shin, J.H., Ryu, C.K., Nam, K.Y., No, K.T., and Park Choo, H.Y. (2006). The development of 3D-QSAR study and recursive partitioning of heterocyclic quinone derivatives with antifungal activity. Bioorg. Med. Chem. *14*, 1608–1617.

Choi, I., Kim, S.Y., Kim, H., Kang, N.S., Bae, M.A., Yoo, S.-E., Jung, J., and No, K.T. (2009). Classification models for CYP450 3A4 inhibitors and non-inhibitors. Eur. J. Med. Chem. *44*, 2354–2360.

Choi, I., Kim, H., Jung, J., Nam, K.Y., Yoo, S.E., Kang, N.S., and No, K.T. (2010). Bayesian model for the classification of GPCR agonists and antagonists. Bull. Korean Chem. Soc. *31*, 2163–2169.

Cieplak, P., Dupradeau, F.Y., Duan, Y., and Wang, J. (2009). Polarization effects in molecular mechanical force fields. J. Phys. Condens. Matter *21*, 333102.

Clancey, W.J., and Shortliffe, E.H. (1985). Readings in medical artificial intelligence. J. Am. Med. Assoc. *253*, 3011–3012.

Clore, G.M., and Gronenborn, A.M. (2010). Structures of larger proteins, protein-ligand and protein-DNA complexes by multidimensional heteronuclear NMR. Protein Sci. *3*, 372–390.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. arXiv, arXiv:1705.02364.

Cramer, R.D. (2012). The inevitable QSAR renaissance. J. Comput. Aided Mol. Des. *26*, 35–38.

Cramer, R.D., Patterson, D.E., and Bunce, J.D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J. Am. Chem. Soc. *110*, 5959–5967.

Csányi, G., Albaret, T., Payne, M.C., and De, V.A. (2004). Learn on the fly": a hybrid classical and

quantum-mechanical molecular dynamics simulation. Phys. Rev. Lett. *93*, 175503.

Culler, D., Karp, R., Patterson, D., Sahay, A., Schauser, K.E., Santos, E., Subramonian, R., and Von Eicken, T. (1993). LogP: towards a realistic model of parallel computation. ACM SIGPLAN Not *28*, 1–12.

Datar, P.A., Khedkar, S.A., Malde, A.K., and Coutinho, E.C. (2006). Comparative residue interaction analysis (CoRIA): a 3D-QSAR approach to explore the binding contributions of active site residues with ligands. J. Comput. Aided Mol. Des. *20*, 343–360.

Dearden, J., Cronin, M.T.D., and Kaiser, K. (2009). How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). SAR QSAR Environ. Res. *20*, 241–266.

Degrado, W.F. (1997). Proteins from scratch. Science *278*, 80–81.

Deka, L., and Quddus, M. (2014). Network-level accident-mapping: distance based pattern matching using artificial neural network. Accid. Anal. Prev. *65*, 105–113.

Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. NAACL (Association for Computational Linguistics), arXiv:1810.04805.

Dijk, E.L.V., Auger, H., Yan, J., and Thermes, C. (2014). Ten years of next-generation sequencing technology. Trends Genet. *30*, 418–426.

Ditzinger, F., Price, D.J., Ilie, A.R., Köhl, N.J., Jankovic, S., Tsakiridou, G., Aleandri, S., Kalantzi, L., Holm, R., Nair, A., et al. (2019). Lipophilicity and hydrophobicity considerations in bio-enabling oral formulations approaches – a PEARRL review. J. Pharm. Pharmacol. *71*, 464–482.

Dixon, S.L., Smondyrev, A.M., Knoll, E.H., Rao, S.N., Shaw, D.E., and Friesner, R.A. (2006). PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. J. Comput. Aided Mol. Des. *20*, 647–671.

Doo Ho Cho, S.K.L., and Bum Tae Kim, K.T.N. (2001). Quantitative structure-activity relationship (QSAR) study of new fluorovinyloxyacetamides. Bull. Korean Chem. Soc. *22*, 388–394.

Dossetter, A.G., Griffen, E.J., and Leach, A.G. (2013). Matched molecular pair analysis in drug discovery. Drug Discov. Today *18*, 724–731.

Dubochet, J., Adrian, M., Chang, J.J., Homo, J.C., Lepault, J., Mcdowall, A.W., and Schultz, P. (1988). Cryo-electron microscopy of vitrified specimens. Q. Rev. Biophys. *21*, 129–228.

Durant, J.L., Leland, B.A., Henry, D.R., and Nourse, J.G. (2002). Reoptimization of MDL keys for use in drug discovery. J. Chem. Inf. Comput. Sci. *42*, 1273–1280.

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R.P. (2015). Convolutional networks on graphs for learning molecular fingerprints. arXiv, arXiv:1509.09292.

Edwards, C.D., Chris, L., Peter, E., and Hessel, E.M. (2016). Development of a novel quantitative structure-activity relationship model to accurately predict pulmonary absorption and replace routine use of the isolated perfused respiring rat lung model. Pharm. Res. *33*, 2604–2616.

Ertel, W. (2017). Reinforcement learning. In Introduction to Artificial Intelligence, W. Ertel, ed. (Springer), pp. 289–311.

Evers, A., Gohlke, H., and Klebe, G. (2003). Ligand-supported homology modelling of protein binding-sites using knowledge-based potentials. J. Mol. Biol. *334*, 327–345.

Ewing, T.J.A., Makino, S., Skillman, A.G., and Kuntz, I.D. (2001). Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. J. Comput. Aided Mol. Des. *15*, 411–428.

Fang, Y., Hoh, J.H., and Spisz, T.S. (1999). Ethanol-induced structural transitions of DNA on mica. Nucleic Acids Res. *27*, 1943–1949.

Feng, E., Shin, W.J., Zhu, X., Li, J., Ye, D., Wang, J., Zheng, M., Zuo, J.P., No, K.T., Liu, X., et al. (2013). Structure-based design and synthesis of C-1- and C-4-modified analogs of zanamivir as neuraminidase inhibitors. J. Med. Chem. *56*, 671–684.

Feng, X., Yang, L.T., Wang, L., and Vinel, A. (2012). Internet of things. Int. J. Commun. Syst. *25*, 1101–1102.

Fernandes, K., Cardoso, J.S., and Fernandes, J. (2017). Transfer Learning with Partial Observability Applied to Cervical Cancer Screening. Pattern Recognition and Image Analysis Lecture Notes in Computer Science (Springer International Publishing)), pp. 243–250.

Ferreira, J.D., and Couto, F.M. (2010). Semantic similarity for automatic classification of chemical compounds. Plos Comput. Biol. *6*, e1000937.

Figueiredo, M.A.T., and Jain, A.K. (2002). Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Mach. Intell. *24*, 381–396.

Freitag, D. (2000). Machine learning for information extraction in informal domains. Mach. Learn. *39*, 169–202.

Gao, Y.D., and Huang, J.F. (2011). An extension strategy of Discovery Studio 2.0 for non-bonded interaction energy automatic calculation at the residue level. Zool. Res. *32*, 262–266.

Gasteiger, J., Rudolph, C., and Sadowski, J. (1990). Automatic generation of 3D-atomic coordinates for organic molecules. Tetrahedron Comput. Methodol. *3*, 537–547.

Gerhard Klebe, U.A., and Mietzner, Thomas (1994). Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J. Med. Chem. *37*, 4130–4146.

Gers, F.A., and Schmidhuber, E. (2001). LSTM recurrent networks learn simple context-free and context-sensitive languages. IEEE Trans. Neural Netw. *12*, 1333–1340.

Giese, T.J. (2005). Development and Validation of New-Generation Molecular Mechanical Force Fields and Semiempirical Hamiltonians, Ph.D. Thesis (University of Minnesota).

Gohlke, H., and Klebe, G. (2002). DrugScore Meets CoMFA: adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. J. Med. Chem. *45*, 4153–4170.

Golkov, V., Skwark, M.J., Mirchev, A., Dikov, G., Geanes, A.R., Mendenhall, J., Meiler, J., and Cremers, D. (2020). 3D deep learning for biological function prediction from physical fields. 2020 International Conference on 3D Vision (3DV).

Gombar, V.K., Polli, J.W., Humphreys, J.E., Wring, S.A., and Serabjit-Singh, C.S. (2004). Predicting P-glycoprotein substrates by a quantitative structure-activity relationship model. J. Pharm. Sci. *93*, 957–968.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., and Bengio, Y. (2014). Generative adversarial nets. arXiv, arXiv:1406.2661.

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

Graves, A., Ndez, S., and Schmidhuber, J. (2005). Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. ICANN 2005 (Springer), pp. 799–804.

Greer, J., Erickson, J.W., Baldwin, J.J., and Varney, M.D. (1994). Application of the three-dimensional structures of protein target molecules in structure-based drug design. J. Med. Chem. *37*, 1035–1054.

Gupta, P.B., Onder, T.T., Jiang, G., Tao, K., Kuperwasser, C., Weinberg, R.A., and Lander, E.S. (2009). Identification of selective inhibitors of cancer stem cells by high-throughput screening. Cell *138*, 645–659.

Habasaki, J., and Okada, I. (2006). Molecular dynamics simulation of alkali silicates based on the quantum mechanical potential surfaces. Mol. Simul. *9*, 319–326.

Halgren, T.A. (2015). Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J. Comput. Chem. *17*, 490–519.

Hansch, C., Maloney, P.P., Fujita, T., and Muir, R.M. (1962). Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. Nature *194*, 178–180.

Hartenfeller, M., Zettl, H., Walter, M., Rupp, M., Reisen, F., Proschak, E., Weggen, S., Stark, H., and Schneider, G. (2012). DOGS: reaction-driven de novo design of bioactive compounds. Plos Comput. Biol. *8*, e1002380.

Hartmman, A.P., Jornada, D.H., and Melo, E.B.D. (2013). A new, fully validated and interpreted quantitative structure-activity relationship model of p -aminosalicylic acid derivatives as neuraminidase inhibitors. Chem. Pap. *67*, 556–567.

Hashmi, A.S.K. (2007). Gold-catalyzed organic reactions. Chem. Rev. *107*, 3180–3211.

Hayakawa, D., Sawada, N., Watanabe, Y., and Gouda, H. (2020). A molecular interaction field describing nonconventional intermolecular interactions and its application to protein–ligand interaction prediction. J. Mol. Graph. Model. *96*, 107515.

Hendrycks, D., and Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv, arXiv:1610.02136.

Hetényi, C., Paragi, G., Maran, U., Timár, Z., Karelson, M., and Penke, B. (2006). Combination of a modified scoring function with two-dimensional descriptors for calculation of binding affinities of bulky, flexible ligands to proteins. J. Am. Chem. Soc. *128*, 1233–1239.

Hill, S.T., Kuintzle, R., Teegarden, A., Danaee, P., and Hendrix, D.A. (2018). A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. Nucleic Acids Res. *46*, 8105–8113.

Hong, H., Xie, Q., Ge, W., Qian, F., Fang, H., Shi, L., Su, Z., Perkins, R., and Tong, W. (2008). Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J. Chem. Inf. Model. *48*, 1337–1344.

Hopfinger, A.J. (1980). A QSAR investigation of dihydrofolate reductase inhibition by Baker Triazines based upon molecular shape analysis. J. Am. Chem. Soc. *102*, 7196–7206.

Hopfinger, A.J., Wang, Shen, Tokarski, John S., Jin, Baiqiang, Albuquerque, Magaly, Madhav, Prakash J., and Duraiswami, C. (1997). Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. J. Am. Chem. Soc. *119*, 10509–10524.

Horváth, A., Hillmer, M., Lou, Q., Hu, X.S., and Niemier, M. (2017). Cellular Neural Network Friendly Convolutional Neural Networks: CNNs with CNNs. DATE (IEEE), pp. 145–150.

Hou, T., and Xu, X. (2002). ADME evaluation in drug discovery. J. Mol. Model. *8*, 337–349.

Hu, H., and Bajorath, J. (2020). Activity cliffs produced by single-atom modification of active compounds: systematic identification and rationalization based on X-ray structures. Eur. J. Med. Chem. *207*, 112846.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. J. Mol. Graph. *14*, 33–38.

Hwang, S., Shin, H.K., Shin, S.E., Seo, M., Jeon, H.N., Yim, D.E., Kim, D.H., and No, K.T. (2020). PreMetabo: an in silico phase I and II drug metabolism prediction platform. Drug Metab. Pharmacokinet. *35*, 361–367.

Igual, L., and Seguí, S. (2017). Supervised learning. In Introduction to Data Science, L. Igual and S. Seguí, eds. (Springer), pp. 67–96.

In, Y.-Y., Lee, S.-K., Kim, P.-J., and No, K.-T. (2012). Prediction of acute toxicity to fathead minnow by local model based QSAR and global QSAR approaches. Bull. Korean Chem. Soc. *33*, 613–619.

Inokuma, Y., Yoshioka, S., and Fujita, M. (2010). A molecular capsule network: guest encapsulation and control of Diels-Alder reactivity. Angew. Chem. *49*, 8912–8914.

Ivanciuc, O., and Braun, W. (2006). Robust quantitative modeling of peptide binding affinities for MHC molecules using physical-chemical descriptors. Protein Pept. Lett. *14*, 903–916.

Jain, A. (2017). Computer aided drug design. J. Phys. Conf. Ser. *884*, 012072.

Jang, J.W., Cho, N.C., Min, S.J., Cho, Y.S., Park, K.D., Seo, S.H., No, K.T., and Pae, A.N. (2016). Novel Scaffold identification of mGlu1 receptor negative allosteric modulators using a hierarchical virtual screening approach. Chem. Biol. Drug Des. *87*, 239–256.

Jaworska, J., Nikolovajeliazkova, N., and Aldenberg, T. (2005). QSAR applicability domain estimation by projection of the training set in descriptor space: a review. Altern. Lab. Anim. *33*, 445–459.

Jebara, T., and Pentland, A.P. (2001). Discriminative, Generative and Imitative Learning. Ph.D. thesis (MIT).

Jing, Y., Bian, Y., Hu, Z., Wang, L., and Xie, X.Q.S. (2018). Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. AAPS J. *20*, 58.

Joung, J.-Y., Kim, H.-J., Kim, H.-M., Ahn, S.-K., Nam, K.-Y., and No, K.-T. (2012). Prediction models of P-glycoprotein substrates using simple 2D and 3D descriptors by a recursive partitioning approach. Bull. Korean Chem. Soc. *33*, 1123–1127.

Kaelbling, L.P., Littman, M.L., and Moore, A.P. (1996). Reinforcement learning: a survey. J. Artif. Intell. Res. *4*, 237–285.

Kang, P.-L., and Liu, Z.-P. (2021). Reaction prediction via atomistic simulation: from quantum mechanics to machine learning. iScience *24*, 102013.

Karelson, M., Lobanov, V.S., and Katritzky, A.R. (1996). Quantum-chemical descriptors in QSAR/QSPR studies. Chem. Rev. *96*, 1027–1044.

Karelson, M., Lobanov, V.S., and Katritzky, A.R. (2010). Quantum-chemical descriptors in QSAR/QSPR studies. Cheminform *96*, 1027–1044.

Kassel, D.B. (2004). Applications of high-throughput ADME in drug discovery. Curr. Opin. Chem. Biol. *8*, 339–345.

Kim, J.H. (2019). Next-generation sequencing technology and personal genome data analysis. In Genome Data Analysis (Springer Singapore), pp. 17–31.

Kim, H.-J., Cho, Y.S., Koh, H.Y., Kong, J.Y., No, K.T., and Pae, A.N. (2006a). Classification of dopamine antagonists using functional feature hypothesis and topological descriptors. Bioorg. Med. Chem. *14*, 1454–1461.

Kim, H.-J., Choo, H., Cho, Y.S., Koh, H.Y., No, K.T., and Pae, A.N. (2006b). Classification of dopamine, serotonin, and dual antagonists by decision trees. Bioorg. Med. Chem. *14*, 2763–2770.

Kim, H.J., Park, W.K., Yong, S.C., Kyoung, T.N., Hun, Y.K., Choo, H., and Ae, N.P. (2008). Classification of piperazinylalkylisoxazole library by RP. Bull. Korean Chem. Soc. *29*, 111–116.

Kim, D.N., Cho, K., Oh, W.S., Lee, C.J., and Lee, S.K. (2009). EaMEAD: activation energy prediction of CYP450 mediated metabolism with effective atomic descriptor. J. Chem. Inf. Model. *49*, 1643–1654.

Kim, N.D., Park, E.S., Kim, Y.H., Moon, S.K., Lee, S.S., Ahn, S.K., Yu, D.Y., No, K.T., and Kim, K.H. (2010). Structure-based virtual screening of novel tubulin inhibitors and their characterization as anti-mitotic agents. Bioorg. Med. Chem. *18*, 7092–7100.

Kim, K.Y., Shin, S.E., and No, K.T. (2015). Assessment of quantitative structure-activity relationship of toxicity prediction models for Korean chemical substance control legislation. Environ. Health Toxicol. *30* (*Suppl*), s2015007.

Kim, S.-H., Choi, J., Lee, K., and No, K.T. (2017). Comparison of three-dimensional ligand-based pharmacophores among 11 phosphodiesterases (PDE 1 to PDE 11) pharmacophores. Bull. Kor. Chem. Soc. *38*, 1033–1037.

Kitchen, D.B., Hélène, D., Furr, J.R., and Jürgen, B. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug Discov. *3*, 935–949.

Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. J. Mol. Biol. *295*, 337–356.

Krieger, E., Nabuurs, S.B., and Vriend, G. (2003). Homology modeling. Methods Biochem. Anal. *44*, 509–523.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Commun. ACM *60*, 84–90.

Kukar, M., Kononenko, I., Grošelj, C., Kralj, K., and Fettich, J. (1999). Analysing and improving the diagnosis of ischaemic heart disease with machine learning. Artif. Intell. Med. *16*, 25–50.

Kurogi, Y., and Guner, O.F. (2001). Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. Curr. Med. Chem. *8*, 1035–1055.

Kwangho Nam, J.G., and York, Darrin M. (2008). Quantum mechanical/molecular mechanical simulation study of the mechanism of Hairpin ribozyme catalysis. J. Am. Chem. Soc. *130*, 4680–4691.

Kwon, Y.E., Park, J.Y., No, K.T., Shin, J.H., Lee, S.K., Eun, J.S., Yang, J.H., Shin, T.Y., Kim, D.K., Chae, B.S., et al. (2007). Synthesis, in vitro assay, and molecular modeling of new piperidine derivatives having dual inhibitory potency against acetylcholinesterase and Aβ1–42 aggregation for Alzheimer's disease therapeutics. Bioorg. Med. Chem. *15*, 6596–6607.

Laitinen, T., Kankare, J.A., and Peräkylä, M. (2010). Free energy simulations and MM-PBSA analyses on the affinity and specificity of steroid binding to

antiestradiol antibody. Proteins Struct. Funct. Bioinf. *55*, 34–43.

Lampi, M.A., Gurska, J., Huang, X.D., Dixon, D.G., and Greenberg, B.M. (2010). A predictive quantitative structure-activity relationship model for the photoinduced toxicity of polycyclic aromatic hydrocarbons to Daphnia magna with the use of factors for photosensitization and photomodification. Environ. Toxicol. Chem. *26*, 406–415.

Larios, E., Zhang, Y., Yan, K., Di, Z., Ledévédec, S., Groffen, F., and Verbeek, F.J. (2012). Automation in Cytomics: A Modern RDBMS Based Platform for Image Analysis and Management in High-Throughput Screening Experiments.

Le, Q.V. (2013). Building High-Level Features Using Large Scale Unsupervised Learning.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436.

Lee, S.-K., Cho, S.-G., Park, J.-S., Kim, K.-Y., and No, K.-T. (2012). MS-HEMs: an on-line management system for high-energy molecules at ADD and BMDRC in Korea. Bull. Kor. Chem. Soc. *33*, 855–861.

Lee, S., Kang, Y.M., Park, H., Dong, M.S., Shin, J.M., and No, K.T. (2013). Human nephrotoxicity prediction models for three types of kidney injury based on data sets of pharmacological compounds and their metabolites. Chem. Res. Toxicol. *26*, 1652–1659.

Lee, K., You, H., Choi, J., and No, K.T. (2017). Development of pharmacophore-based classification model for activators of constitutive androstane receptor. Drug Metab. Pharmacokinet. *32*, 172–178.

Li, A.P. (2001). Screening for human ADME/Tox drug properties in drug discovery. Drug Discov. Today *6*, 357–366.

Li, K.C., Marcovici, P., Phelps, A., Potter, C., Tillack, A., Tomich, J., and Tridandapani, S. (2013). Digitization of medicine: how radiology can take advantage of the digital revolution. Acad. Radiol. *20*, 1479–1494.

Lichtenstein, B.R., Farid, T.A., Kodali, G., Solomon, L.A., Anderson, J.L.R., Sheehan, M.M., Ennist, N.M., Fry, B.A., Chobot, S.E., Bialas, C., et al. (2012). Engineering oxidoreductases: maquette proteins designed from scratch. Biochem. Soc. Trans. *40*, 561.

Lill, M.A., and Danielson, M.L. (2011). Computer-aided drug design platform using PyMOL. J. Comput. Aided Mol. Des. *25*, 13–19.

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

Lin, M.S. (2009). A physics-based energy function for ab initio protein structure prediction and refinement. Dissertations & Theses - Gradworks.

Liou, C.-Y., Cheng, W.-C., Liou, J.-W., and Liou, D.-R. (2014). Autoencoder for words. Neurocomputing *139*, 84–96.

Lipkowitz, K.B., and Boyd, D.B. (2007). Approaches to three-dimensional quantitative structure-activity relationships.

Lipkus, A.H. (1999). A proof of the triangle inequality for the Tanimoto distance. J. Math. Chem. *26*, 263–265.

Liu, A. (2016). The Discovery of Small Molecular Inhibitors of PD-1/pd-L1 Pathway. Doctor (Jilin University).

Lo, Y.-C., Senese, S., Damoiseaux, R., and Torres, J.Z. (2016). 3D chemical similarity networks for structure-based target prediction and Scaffold Hopping. ACS Chem. Biol. *11*, 2244–2253.

Lo, Y.C., Rensi, S.E., Wen, T., and Altman, R.B. (2018). Machine learning in chemoinformatics and drug discovery. Drug Discov. Today *102*, 71.

Lokuge, S., Hewavitarne, H., Wimalaratne, P., and Ranawana, R. (2010). Machine learning based Qsar for discovering potential drug candidate from endemic plants of Sri Lanka- case study: Hiv-1 Rt. VCON *10*, 12–17.

Low, C.M.R., and Vinter, J.G. (2008). Rationalizing the activities of diverse cholecystokinin 2 receptor antagonists using molecular field points. J. Med. Chem. *51*, 565–573.

Lu, I.-L., Huang, C.-F., Peng, Y.-H., Lin, Y.-T., Hsieh, H.-P., Chen, C.-T., Lien, T.-W., Lee, H.-J., Mahindroo, N., Prakash, E., et al. (2006). Structure-based drug design of a novel family of PPARgamma partial agonists: virtual screening, X-ray crystallography, and in vitro/in vivo biological activities. J. Med. Chem. *49*, 2703–2712.

Luo, M. (2015). The Research of Brain Tumor Segmentation Based on MRI Multi-Modality Images and 3D-CNNs Features (Southern Medical University).

Lutz, M.W., Menius, J.A., Choi, T.D., Laskody, R.G., Domanico, P.L., Goetz, A.S., and Saussy, D.L. (1996). Experimental design for high-throughput screening. Drug Discov. Today *1*, 277–286.

Ma, S.L., Joung, J.Y., Lee, S., Cho, K.H., and No, K.T. (2012). PXR ligand classification model with SFED-weighted WHIM and CoMMA descriptors. SAR QSAR Environ. Res. *23*, 485–504.

Marchi, E., Vesperini, F., Eyben, F., Squartini, S., and Schuller, B. (2015). A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks. ICASSP *2015*, 1996–2000.

Matsumoto, S., Ishida, S., Araki, M., Kato, T., Terayama, K., and Okuno, Y. (2021). Extraction of protein dynamics information from cryo-EM maps using deep learning. Nat. Mach. Intell. *3*, 153–160.

Mauri, A., Consonni, V., Pavan, M., Todeschini, R., and Chemometrics, M. (2006). DRAGON software: an easy approach to molecular descriptor calculations. MATCH Commun. Math. Comput. Chem. *56*, 237–248.

Mcinnes, C. (2007). Virtual screening strategies in drug discovery. Curr. Opin. Chem. Biol. *11*, 494–502.

Minar, M.R., and Naher, J. (2018). Recent advances in deep learning: an overview. arXiv, arXiv:1807.08169.

Mochizuki, Y., Komeiji, Y., Ishikawa, T., Nakano, T., and Yamataka, H. (2007). A fully quantum mechanical simulation study on the lowest n–π* state of hydrated formaldehyde. Chem. Phys. Lett. *437*, 66–72.

Møller, M.F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw. *6*, 525–533.

Morell, Christophe, Grand, André, and Torolabbé, A. (2005). New dual descriptor for chemical reactivity. J. Phys. Chem. A. *109*, 205–212.

Morris, G.M., and Lim-Wilby, M. (2008). Molecular docking. Methods Mol. Biol. *443*, 365–382.

Nam, K.Y., Chang, B.H., Han, C.K., Ahn, S.K., and No, K.T. (2003). Investigation of the protonated state of HIV-1 protease active site. Bull. Korean Chem. Soc. *24*, 817–823.

Nam, K.-Y., Kang, J.H., No, K.T., and Ahn, S.K. (2014). Identification of Polo-like kinase 1 inhibitors using structure-based molecular design. Bull. Korean Chem. Soc. *35*, 1929–1930.

Nam, Ky-Youb, Oh, W.S., Kim, C., Song, M.Y., Joung, J.Y., Kim, S.Y., Park, J.S., Gang, S.M., Cho, Y.U., No, K.T., et al. (2011). Computational drug discovery approach based on nuclear factor-κB pathway dynamics. Bull. Kor. Chem. Soc. *32*, 1–6.

Nilakantan, R., Bauman, N., and Dixon, J.S. (1987). Topologial torsion: a new molecular descriptor for sar applications. Comparison with other descriptors. J. Chem. Inf. Model. *27*, 82–85.

Noel, Y., D'Arco, P., Demichelis, R., Zicovich-Wilson, C.M., and Dovesi, R. (2010). On the use of symmetry in the ab initio quantum mechanical simulation of nanotubes and related materials. J. Comput. Chem. *31*, 855–862.

Ohashi, W., and Tanaka, H. (2010). Benefits of pharmacogenomics in drug development—earlier launch of drugs and less adverse events. J. Med. Syst. *34*, 701–707.

Palangi, H., Deng, L., and Ward, R.K. (2014). Recurrent Deep-Stacking Networks for Sequence Classification. ChinaSIP (IEEE), pp. 510–514.

Pan, S.J., and Yang, Q. (2010). A survey on transfer learning. IEEE Trans. Knowl. Data Eng. *22*, 1345–1359.

Pan, S., Zhang, H., Rush, J., Eng, J., Zhang, N., Patterson, D., Comb, M.J., and Aebersold, R. (2005). High throughput proteome screening for biomarker detection. Mol. Cell. Proteomics *4*, 182.

Panych, L.P., Oesterle, C., Zientara, G.P., and Hennig, J. (2015). Implementation of a fast gradient-echo SVD encoding technique for dynamic imaging. Magn. Reson. Med. *35*, 554–562.

Parashar, A., Rhu, M., Mukkara, A., Puglielli, A., Venkatesan, R., Khailany, B., Emer, J., Keckler, S.W., and Dally, W.J. (2017). SCNN: an accelerator for compressed-sparse convolutional

neural networks. ACM SIGARCH Computer Architecture News *45*, 27–40, arXiv:1708.04485.

Park, H., Kim, K.K., Kim, C., Shin, J.-M., and No, K.T. (2013). Descriptor-based profile analysis of kinase inhibitors to predict inhibitory activity and to grasp kinase selectivity. Bull. Korean Chem. Soc. *34*, 2680–2684.

Pasquier, C., and Hamodrakas, S.J. (2009). An hierarchical artificial neural network system for the classification of transmembrane proteins. Protein Eng. *12*, 631–634.

Paul, R., Hawkins, S.H., Balagurunathan, Y., Schabath, M.B., Gillies, R.J., Hall, L.O., and Goldgof, D.B. (2016). Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. Tomography *2*, 388–395.

Pei, J., Zheng, Z., and Merz, K.M. (2019). Random forest refinement of the KECSA2 knowledge-based scoring function for protein decoy detection. J. Chem. Inf. Model. *59*, 1919–1929.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L., Schulten, K., et al. (2010). Scalable molecular dynamics with NAMD. J. Comput. Chem. *26*, 1781–1802.

Pinaya, W.H.L., Gadelha, A., Doyle, O.M., Noto, C., Zugman, A., Cordeiro, Q., Jackowski, A.P., Bressan, R.A., and Sato, J.R. (2016). Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. Sci. Rep. *6*, 38897.

Piotr, Tu, Zhuowen, Belongie, and Serge. (2006). Supervised learning of edges and object boundaries. CVPR *06*, 1964–1971.

Polanski, J. (2009). Receptor dependent multidimensional QSAR for modeling drug - receptor interactions. Curr. Med. Chem. *16*, 3243–3257.

Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics *31*, 761–763.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv, arXiv:1511.06434.

Randić, M. (1993). Novel molecular descriptor for structure—property studies. Chem. Phys. Lett. *211*, 478–483.

Rapaport, D.C. (2004). The Art of Molecular Dynamics Simulation, 2 Edition (Cambridge University Press).

Rapaport, D.C., Blumberg, R.L., Mckay, S.R., and Christian, W. (2002). The art of molecular dynamics simulation. Comput. Sci. Eng. *1*, 70–71.

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. J. Chem. Inf. Model. *50*, 742–754.

Rush, T.S., Grant, J.A., Mosyak, L., and Nicholls, A. (2005). A shape-based 3-D Scaffold hopping method and its application to a bacterial protein–protein interaction. J. Med. Chem. *48*, 1489–1495.

Sabljić, A., Güsten, H., Verhaar, H., and Hermens, J. (1995). QSAR modelling of soil sorption. Improvements and systematics of log K OC vs. log K OW correlations. Chemosphere *31*, 4489–4514.

Sahami, M. (1997). Supervised and Unsupervised Discretization of Continuous Features.

Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., and Wold, S. (1998). New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J. Med. Chem. *41*, 2481–2491.

Sasakawa, T., Hu, J., and Hirasawa, K. (2010). A brainlike learning system with supervised, unsupervised, and reinforcement learning. Electr. Eng. Jpn. *162*, 32–39.

Schneider, P., and Schneider, G. (2018). Polypharmacological drug—target inference for chemogenomics. Mol. Inf. *37*, e1800050.

Secco, J., Farina, M., Demarchi, D., Corinto, F., and Gilli, M. (2016). Memristor Cellular Automata for Image Pattern Recognition and Clinical Applications. ISCAS 2016 (IEEE), pp. 1378–1381.

Semper, C., Watanabe, N., and Savchenko, A. (2021). Structural characterization of nonstructural protein 1 from SARS-CoV-2. iScience *24*, 101903.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., et al. (2019). Protein structure prediction using multiple deep neural networks in the 13th critical assessment of protein structure prediction (CASP13). Proteins *87*, 1141–1148.

Sheridan, R.P., Miller, M.D., Underwood, D.J., and Kearsley, S.K. (1996). Chemical similarity using geometric atom pair descriptors. J. Chem. Inf. Comput. Sci. *36*, 128–136.

Shim, H.J., Yang, H.R., Kim, H.I., Kang, S.A., No, K.T., Jung, Y.H., and Lee, S.T. (2014). Discovery of (E)-5-(benzylideneamino)-1H-benzo[d]imidazol-2(3H)-one derivatives as inhibitors PTK-6. Bioorg. Med. Chem. Lett. *24*, 4659–4663.

Shimamoto, N. (1999). One-dimensional diffusion of proteins along DNA. J. Biol. Chem. *274*, 15293–15296.

Shin, W.J., Nam, K.Y., Kim, N.D., Kim, S.H., No, K.T., and Seong, B.L. (2016). Identification of a small benzamide inhibitor of influenza virus using a cell-based screening. Chemotherapy *61*, 159–166.

Shin, H.K., Kim, K.Y., Park, J.W., and No, K.T. (2017). Use of metal/metal oxide spherical cluster and hydroxyl metal coordination complex for descriptor calculation in development of nanoparticle cytotoxicity classification model. SAR QSAR Environ. Res. *28*, 875–888.

Shin, H.K., Seo, M., Shin, S.E., Kim, K.-Y., Park, J.-W., and No, K.T. (2018). Meta-analysis of Daphnia magnananotoxicity experiments in accordance with test guidelines. Environ. Sci. Nano *5*, 765–775.

Shu, L.O., Ng, E.Y.K., Ru, S.T., and Acharya, U.R. (2018). Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. Comput. Biol. Med., S0010482518301446.

Simon, Z., Badilescu, I., and Racovitan, T. (1977). Mapping of dihydrofolate-reductase receptor site by correlation with minimal topological (steric) differences. J. Theor. Biol. *66*, 485–495.

Singh, N., Shah, P., Dwivedi, H., Mishra, S., Tripathi, R., Sahasrabuddhe, A.A., and Siddiqi, M.I. (2016). Integrated machine learning, molecular docking and 3D-QSAR based approach for identification of potential inhibitors of trypanosomal N-myristoyltransferase. Mol. Biosyst. *12*, 3711–3723.

Sondak, V.K.S.E. (1990). New directions for medical artificial intelligence. Comput. Math. Appl. *20*, 313–319.

Song, S.W., Kim, S.D., Oh, D.Y., Lee, Y., Lee, A.C., Jeong, Y., Bae, H.J., Lee, D., Lee, S., Kim, J., et al. (2019). High-throughput screening: one-step generation of a drug-releasing hydrogel microarray-on-a-chip for large-scale sequential drug combination screening. Adv. Sci. *6*, 1801380.

Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., Zhang, J., Chan, L., and Cao, R. (2019). Survey of machine learning techniques in drug discovery. Curr. Drug Metab. *20*, 185–193.

Strasser, P., Fan, Q., Martin Devenney, A., Weinberg, W.H., and Nørskov, J.K. (2003). High throughput experimental and theoretical predictive screening of materials – a comparative study of search strategies for new fuel cell anode catalysts. J. Phys. Chem. B *107*, 11013–11021.

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., and Jiang, P. (2019). BERT4Rec: sequential recommendation with bidirectional encoder representations from transformer. Proceedings of the 28th ACM international conference on information and knowledge management, 1441–1450.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., and Feuston, B.P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. *43*, 1947.

Tan, J.W., Deng, S.J., Ye, F.W., and Zeng, D.P. (2013). Variability analysis of T network impedance matching. Appl. Mech. Mater. *427-429*, 620–623.

Thangapandian, S., John, S., Son, M., Arulalapperumal, V., and Lee, K.W. (2013). Development of predictive quantitative structure-activity relationship model and its application in the discovery of human leukotriene A4 hydrolase inhibitors. Fut. Med. Chem. *5*, 27–40.

Todeschini, R., and Consonni, V. (2009). Molecular Descriptors for Chåemoinformatics (Wiley-VCH).

Tomal, J.H., Welch, W.J., and Zamar, R.H. (2016). Exploiting multiple descriptor sets in QSAR studies. J. Chem. Inf. Model. *56*, 501–509.

Trott, O., and Olson, A.J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. *31*, 455–461.

Tu, Y., and Laaksonen, A. (2001). Atomic charges in molecular mechanical force fields: a theoretical insight. Phys. Rev. E: Stat. Nonlinear, Soft Matter Phys. *64*, 026703.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 384–394.

Vapnik, V., Levin, E., and Cun, Y. (1994). Measuring the VC-dimension of a learning machine. Neural Comput. *6*, 851–876.

Varela, R., Walters, W.P., Goldman, B.B., and Jain, A.N. (2012). Iterative refinement of a binding pocket model: active computational steering of lead optimization. J. Med. Chem. *55*, 8926–8942.

Vedani, A., and Dobler, M. (2002). 5D-QSAR: the key for simulating induced fit? J. Med. Chem. *45*, 2139–2149.

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

Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W., and Taylor, R.D. (2010). Improved protein-ligand docking using GOLD. Proteins Struct. Funct. Bioinf. *52*, 609–623.

Verlinde, C.L., and Hol, W.G. (1994). Structure-based drug design: progress, results and challenges. Structure *2*, 577–587.

Veselovsky, A.V., Tikhonova, O.V., Skvortsov, V.S., Medvedev, A.E., and Ivanov, A.S. (2001). An approach for visualization of the active site of enzymes with unknown three-dimensional structures. SAR QSAR Environ. Res. *12*, 345–358.

Vesely, F.J. (2001). Quantum mechanical simulation. In Computational Physics: An Introduction, F.J. Vesely, ed. (Springer US), pp. 195–214.

Wallach, I., Dzamba, M., and Heifets, A. (2015). AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. Math. Z. *47*, 34–46.

Walters, D.E., and Hinds, R.M. (1994). Genetically evolved receptor models: a computational approach to construction of receptor models. J. Med. Chem. *37*, 2527–2536.

Wang, S. (2013). The Study about PET/CT Image Registration Method Based on CNN and Mutual Information (Qingdao University).

Wang, Q., He, J., Wu, D., Wang, J., Yan, J., and Li, H. (2015a). Interaction of α-cyperone with human serum albumin: determination of the binding site by using Discovery Studio and via spectroscopic methods. J. Lumin. *164*, 81–85.

Wang, X., Zhang, T., Chaim, T.M., Zanetti, M.V., and Davatzikos, C. (2015b). Classification of MRI under the presence of disease heterogeneity using multi-task learning: application to bipolar disorder. Med. Image Comput. Comput. Assist. Interv. *9349*, 125–132.

Wang, J., Zhang, J., An, Y., Lin, H., Yang, Z., Zhang, Y., and Sun, Y. (2016). Biomedical event trigger detection by dependency-based word embedding. BMC Med. Genomics *9*, 45.

Wang, C., Elazab, A., Jia, F., Wu, J., and Hu, Q. (2018). Automated chest screening based on a hybrid model of transfer learning and convolutional sparse denoising autoencoder. Biomed. Eng. Online *17*, 63.

Wang, Y., Ye, Z., Wan, P., and Zhao, J. (2019). A survey of dynamic spectrum allocation based on reinforcement learning algorithms in cognitive radio networks. Artif. Intell. Rev. *51*, 493–506.

Wenz, C. (1982). Development and drugs: more not less. Nature *297*, 173–174.

Widera, P. (2010). Evolutionary Symbolic Discovery for Bioinformatics, Sysems and Synthetic Biology. GECCO '10 (Association for Computing Machinery), pp. 1991–1998.

Willett, P., Barnard, J.M., and Downs, G.M. (1998). Chemical similarity searching. J. Chem. Inf. Comput. Sci. *38*, 983–996.

Wiskott, L., and Sejnowski, T.J. (2002). Slow feature analysis: unsupervised learning of invariances. Neural Comput. *14*, 715.

Wolber, Gerhard, and Sippl, W. (2008). Pharmacophore identification and pseudo-receptor modeling. In The Practice of Medicinal Chemistry, 4th Ed, C.G. Wermuth., D. Aldous., P. Raboisson., and D. Rognan., eds. (Academic Press), pp. 489–510.

Wold, S. (2010). Validation of QSAR's. Mol. Inf. *10*, 191–193.

Wu, Y., and Wang, G. (2018). Machine learning based toxicity prediction: from chemical structural description to transcriptome analysis. Int. J. Mol. Sci. *19*, 2358.

Xi, E., Bing, S., and Jin, Y. (2017). Capsule network performance on complex data. arXiv, arXiv:1712.03480.

Xing, L., and Glen, R.C. (2010). Novel methods for the prediction of logP, pK(a), and logD. J. Chem. Inf. Comput. Sci. *33*, 231.

Xu, J., Lei, X., Hang, R., and Wu, J. (2014). Stacked Sparse Autoencoder (SSAE) based framework for nuclei patch classification on breast cancer histopathology. ISBI 2014 (IEEE), pp. 999–1002.

Xu, D., Gopale, M., Zhang, J., Brown, K., Begoli, E., and Bethard, S. (2020). Unified medical language system resources improve sieve-based generation and Bidirectional Encoder Representations from Transformers (BERT)–based ranking for concept normalization. J. Am. Med. Inform. Assoc. *27*, 1510–1519.

Xue, D.X., Zhang, R., Feng, H., and Wang, Y.L. (2016). CNN-SVM for microvascular morphological type recognition with data augmentation. J. Med. Biol. Eng. *36*, 755–764.

Yang, S.Y. (2010). Pharmacophore modeling and applications in drug discovery: challenges and recent advances. Drug Discov. Today *15*, 444–450.

Yang, L., Yang, X., Pu, X., Qian, Z., and Wang, K. (2004). The absorption, distriution, metabolism, excretion, toxicity (ADME/Tox.) platform construction of novel drugs research. J. Peking Univ. Health Sci. *36*, 5–8.

Yang, Y.F., Wang, L., Yan, S.J., Zhi, L.I., Wang, Y.Y., and Zhang, W.S. (2011). Discovery Studio software in the analysis of the blood-brain barrier penetrations of active components of traditional Chinese medicines. Chin. Pharmacol. Bull. *27*, 739–740.

Yao, J., Zhou, C., and Motani, M. (2017). Spatio-Temporal Autoencoder for Feature Learning in Patient Data with Missing Observations, *2017* (BIBM), pp. 886–890.

Yi, X., Walia, E., and Babyn, P. (2019). Generative adversarial network in medical imaging: a review. Med. Image Anal. *58*, 101552.

Yildirim, Ö. (2018). A novel wavelet sequences based on deep bidirectional LSTM network model for ECG signal classification. Comput. Biol. Med. *96*, 189.

Yoshioka, T., Chen, X., and Gales, M.J.F. (2014). Impact of single-microphone dereverberation on DNN-based meeting transcription systems. Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 5527–5531.

You, H., Lee, K., Lee, S., Hwang, S.B., Kim, K.Y., Cho, K.H., and No, K.T. (2015). Computational classification models for predicting the interaction of compounds with hepatic organic ion importers. Drug Metab. Pharmacokinet. *30*, 347–351.

Yu, K., Liu, C., Kim, B.G., and Lee, D.Y. (2015). Synthetic fusion protein design and applications. Biotechnol. Adv. *33*, 155–164.

Yuan, S., Dahoun, T., Brugarolas, M., Pick, H., Filipek, S., and Vogel, H. (2019). Computational modeling of the olfactory receptor Olfr73 suggests a molecular basis for low potency of olfactory receptor-activating compounds. Commun. Biol. *2*, 141.

Zeng, M., Li, M., Fei, Z., Wu, F., Li, Y., Pan, Y., and Wang, J. (2021). A deep learning framework for identifying essential proteins by integrating multiple types of biological information. IEEE ACM Trans. Comput. Bi. *18*, 296–305.

Zhan, C., Zhang, L., Zhong, Z., Didi-Ooi, S., Lin, Y., Zhang, Y., Huang, S., and Wang, C. (2018). Deep learning approach in automatic iceberg - ship detection with sar remote sensing data. arXiv, arXiv:1812.07367.

Zhang, C.X., Nan-Nan, J.I., and Wang, G.W. (2015). Restricted Boltzmann machines. Chin. J. Eng. Math. *2015*, 159–173.

Zhang, Y., Dijkman, P.M., Zou, R., Zandl-Lang, M., Sanchez, R.M., Eckhardt-Strelau, L., Köfeler, H., Vogel, H., Yuan, S., Kudryashev, M., et al. (2021). Asymmetric opening of the homopentameric 5-HT3A serotonin receptor in lipid bilayers. Nat. Commun. *12*, 1074.

Zhang, Y., Huang, Q., Ma, X., Yang, Z., and Jiang, J. (2017). Using multi-features and ensemble learning method for imbalanced malware classification. In 2016 IEEE Trustcom/BigDataSE/ISPA, pp. 965–973.

Zhao, A. (2003). The situation in bioinformatics research and development. China Biotechnol. *23*, 101–103.

Zhao, G., Wang, X., Niu, Y., Tan, L., and Zhang, S.X. (2016). Segmenting brain tissues from Chinese visible human dataset by deep-learned features with stacked autoencoder. Biomed. Res. Int. *2016*, 1–12.

Zheng, Y., Qiao, X., Yang, X., Chen, G., and Li, X. (2013). Quantitative structure-activity relationship model for bioconcentration factors of halogenated organic compounds. Asian J. Ecotoxicol. *8*, 772–777.

Zhu, X., and Goldberg, A.B. (2009). Introduction to semi-supervised learning. Synth. Lect. Artif. Intell. Mach. Learn. *3*, 1–130.

Zhu, X., Suk, H.I., Lee, S.W., and Shen, D. (2016). Subspace regularized sparse multi-task learning for multi-class neurodegenerative disease identification. IEEE Trans. Biomed. Eng. *63*, 607–618.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. arXiv, arXiv:1604.02201.