HHS Public Access

Author manuscript

Drug Discov Today. Author manuscript; available in PMC 2019 August 01.

Published in final edited form as:

Drug Discov Today. 2018 August ; 23(8): 1538–1546. doi:10.1016/j.drudis.2018.05.010.

Machine learning in chemoinformatics and drug discovery

Yu-Chen Lo, **Stefano E. Rensi**, **Wen Torng**, and **Russ B. Altman**[*]
Department of Bioengineering, Stanford University, Stanford, CA, USA

Abstract

Chemoinformatics is an established discipline focusing on extracting, processing and extrapolating meaningful data from chemical structures. With the rapid explosion of chemical 'big' data from HTS and combinatorial synthesis, machine learning has become an indispensable tool for drug designers to mine chemical information from large compound databases to design drugs with important biological properties. To process the chemical data, we first reviewed multiple processing layers in the chemoinformatics pipeline followed by the introduction of commonly used machine learning models in drug discovery and QSAR analysis. Here, we present basic principles and recent case studies to demonstrate the utility of machine learning techniques in chemoinformatics analyses; and we discuss limitations and future directions to guide further development in this evolving field.

Keywords

Machine learning; chemoinformatics; data mining; drug discovery

Introduction

Machine learning is currently one of the most important and rapidly evolving topics in computer-aided drug discovery [1]. In contrast to physical models that rely on explicit physical equations like quantum chemistry or molecular dynamics simulations, machine learning approaches use pattern recognition algorithms to discern mathematical relationships between empirical observations of small molecules and extrapolate them to predict chemical, biological and physical properties of novel compounds. Also, in comparison to physical models, machine learning techniques are more efficient and can easily be scaled to big datasets without the need for extensive computational resources. One of the primary application areas for machine learning in drug discovery is helping researchers understand and exploit relationships between chemical structures and their biological activities or SAR [2]. For instance, given a hit compound from a drug screening campaign, we might wish to know how its chemical structure can be optimized to improve its binding affinity, biological

---

[*] Corresponding author: Altman, R.B. (rbaltman@stanford.edu).

Teaser: Machine learning is a rapidly evolving area in chemoinformatics and drug discovery where intelligent models are constructed based on chemical features of known drugs to predict properties of novel compounds.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

responses or physiochemical properties. Fifty years ago, this type of problem could only be addressed through numerous costly, time-consuming, labor-intensive cycles of medicinal chemistry synthesis and analysis. Today, modern machine learning techniques can be used to model QSAR, or quantitative structure–property relationships (QSPR), and develop artificial intelligence programs that accurately predict *in silico* how chemical modifications might influence biological behavior [3]. Many physiochemical properties of drugs, such as toxicity, metabolism, drug–drug interactions and carcinogenesis, have been effectively modeled by QSAR techniques [3]. Early QSAR models, such as Hansch and Free–Wilson analysis, used simple multivariate regression models to correlate potency ($logIC_{50}$) with substructure motifs and chemical properties like solubility (logP), hydrophobicity, substituent pattern and electronic factors [4]. Although groundbreaking and successful, these approaches were ultimately limited by unavailability of experimental data and the linearity assumption made in modeling. Therefore, advanced chemoinformatics and machine learning techniques capable of modeling nonlinear datasets, as well as big data of increasing depth and complexity, are needed.

Overview of chemoinformatics

Chemoinformatics is a broad field that encompasses computer science and chemistry with the goal of utilizing computer information technology to solve problems in the field of chemistry such as chemical information retrieval and extraction, compound database searching and molecular graph mining [5,6]. Other areas of chemoinformatics related to drug discovery also include computer-aided drug synthesis (a very broad field with >50 years' history), chemical space exploration, pharmacophore and scaffold analysis, library design, among others [7,8]. Converting a compound structure into chemical information applicable for machine learning tasks requires multilayer computational processing from chemical graph retrieval, descriptor generation, fingerprint construction to similarity analysis, in which each layer is built upon the successful development of previous layers and often has a substantial impact on the quality of the chemical data for machine learning (Figure 1).

Chemical graph theory

To understand how the structures of chemicals influence their biological activities, it is imperative to review the foundations of chemical graph theory [9]. A chemical graph, also known as a 'molecular graph' or 'structural graph', is a mathematical construct comprising an ordered pair G = (V,E), where V is a set of vertices (atoms) connected by a set of edges (bonds) E. Chemical graph theory maintains that, because chemical structures are fully specified by their graph representations, they contain the information necessary to model and provide insight into a wide range of biological phenomena. Several variations of chemical graphs have been proposed [10]. Weighted chemical graphs assign values to edges and vertices to indicate bond lengths and other atomic properties [11]. Chemical pseudographs or reduced graphs use multiple edges and self-loops to capture detailed bond valence information [7]. Regardless of flavor, chemical graphs represent atomic connectivity using a bond adjacency matrix, or topological distance matrix, which supports the computation of several topological indices useful for chemoinformatics modeling [12].

Garcia-Domenech *et al.* demonstrated the application of chemical graphs for chemometric analysis. In their study, they proposed an equation that combined pseudograph vertex degree derived from the adjacency matrix with two key parameters from the complete graph to model the electronegativity of 30 elements from the main group of the periodic table [10]. More recently, Fourches and Tropsha developed the advanced dataset graph analysis (ADDAGRA) approach. In this work, they combined multiple graph indices from bond connectivity matrices to compare and quantify chemical diversity for large compound sets using chemical space networks in high-dimensional space. The study showed that the ADDAGRA approach could uncover shared chemical space between chemical databases to improve SAR analysis [13].

Chemical descriptors

Chemical descriptors are numerical features extracted from chemical structures for molecular data mining, compound diversity analysis and compound activity prediction [14–16]. Chemical descriptors can be one-dimensional (0D or 1D), 2D, 3D or 4D (Table 1) [17]. One-dimensional descriptors are scalars that describe aggregate information such as atom counts, bond counts, molecular weight, sums of atomic properties or fragment counts [18]. Although simple to compute, 1D descriptors suffer from degeneracy problems where distinct compounds are mapped to identical descriptor values for a given descriptor. Thus, 1D descriptors are usually used in concert with higher-dimensional descriptors or expressed as a vector of multiple 1D descriptors. 2D chemical descriptors are the most frequent descriptor type reported in the literature, and include topological indices, molecular profiles and 2D autocorrelation descriptors [18]. An important feature of 2D descriptors, which makes them useful for structure differentiation, is graph invariance where descriptor values are unaffected by the renumbering of graph nodes (vertices). To facilitate analysis of the large space of 2D descriptors, Hong *et al.* reported the Mol^2 system that rapidly generates up to 200 types of 2D descriptors for large compound datasets [19]. Other commercial software packages commonly used in the descriptor generation include the DRAGON system, which can generate up to 5000 types of descriptors as part of several QSAR studies [20,21].

3D chemical descriptors extract chemical features from 3D coordinate representations and are considered the most sensitive to structural variations [22–25]. Well-known 3D descriptors include autocorrelation descriptors, substituent constants, surface:volume descriptors and quantum–chemical descriptors [18]. 3D chemical descriptors are useful for identifying ‘scaffold hops’ – distinct chemical scaffolds with similar binding activities [26]. A key limitation of 3D chemical descriptors in QSAR analysis is the computational complexity of conformer generation and structure alignments; which are absent of any guarantees that predicted conformations correspond to relevant bioactive conformations. 4D chemical descriptors are an extension of 3D chemical descriptors that simultaneously consider multiple structural conformations [27]. Ash and Fourches applied molecular dynamics simulation on ERK2 kinase to compute 3D descriptors over a grid box based on the 20 ns trajectory, and showed that such 4D chemical descriptors can effectively differentiate the most active ERK2 inhibitors from the inactive ones with superior enrichment rates [28].

Chemical fingerprints

Chemical fingerprints are high-dimensional vectors, commonly used in chemometric analysis and similarity-based virtual screening applications, the elements of which are chemical descriptor values [29]. MACCS substructure fingerprints are 2D binary fingerprints (0 and 1), with each of 166 bits indicating the presence or absence of particular substructure keys [30]. Daylight fingerprints and extended connectivity fingerprints (ECFP) extract chemical patterns of up to a specified length or diameter from a chemical graph. In comparison to the predefined substructure keys of MACCS, these fingerprints can dynamically index features using hash functions and often yield higher specificity when searching complex structures [31]. The latest development in 2D fingerprints are continuous kernel and neural embedded fingerprints – internal representations learned by support vector machines (SVMs) and neural networks. Duvenaud *et al.* extended the convolution concept to molecules represented as 2D molecular graphs for extracting molecular representation [32]. The architecture generalizes the fingerprint computation such that the representation can be learned via back-propagation in a data-driven manner, and improves predictions of solubility, drug efficacy and organic photovoltaic efficiency.

3D fingerprints commonly used in 3D-QSAR studies include chemical features based on pharmacophoric patterns, surface properties, molecular volumes or molecular interaction fields [24,33]. One of the best known 3D fingerprints is molecular interaction field (MIF), as implemented in the GRID program by Goodford [34]. The MIF-based fingerprint places the ligand in a rectangular grid with a fixed interval and calculates the electronic, steric and hydrophobic contribution independently at each grid point. The resulting MIF-based fingerprints can then be used in comparative molecular field analysis (CoMFA) by deriving relationships between 3D grid points and compound activities [35]. The dependency upon the relative orientation of the molecules within the grid box is a major limitation of 3D-QSAR techniques such as CoMFA analysis. To remove the dependency of ligand orientation in 3D-QSAR analysis, Baskin and Zhokhova recently introduced the continuous molecular field (CMF) approach which replaced grid points with continuous function to represent molecular fields and showed that its simplest form provides either comparable or enhanced predictive performance in comparison with state-of-the-art CoMFA methods [36].

Chemical similarity analysis

Chemical similarity search is a fundamental technique for ligand-based drug discovery [37]. Its objective is to identify and return database compounds with structures and bioactivities similar to query compounds [38]. The chemical similarity principle, which states that compounds with similar structures will probably have similar bioactivities, is an underlying assumption of similarity-based virtual screening [39]. However, this assumption might not always be valid. For example, ‘activity cliffs’ where minor modification of functional groups causes an abrupt change in activity violate this principle and can cause failure of QSAR models [40,41]. The structural similarity of two molecules is most commonly evaluated by computing the Tanimoto coefficient (Tc) of their chemical fingerprints. The Tc, also known as the Jaccard index, is a measure of similarity between sets that compute a similarity score as the fraction of bits shared by two feature vectors. High Tc values indicate two compounds are similar but do not provide information dimensions of similarity, such as which specific chemical groups they share.

Chemical similarity can also be evaluated based on 3D structural features of compounds. The 3D Tanimoto index is a common 3D similarity metric that computes the fraction of shared molecular volumes between two comparing ligands [42]. Examples of volume-based similarity implementation include the rapid overlap of chemical structures (ROCS) program – the most popular shape similarity approach in drug discovery based on Gaussian representation of molecular shape [43]. An alternative 3D similarity metric is the pharmacophoric similarity, which considers only the volume overlap between crucial functional groups. Lo *et al.* developed the ShapeAlign program that combines 2D and 3D metrics based on the Obabel PF2 fingerprint, shapes and pharmacophoric points for unsupervised 3D chemical similarity clustering [44,45]. The validation study using 20 known drug classes retrieved from the directory of useful decoys (DUD) showed that the combined metrics outperformed either 2D or 3D metrics and successfully detected shared 3D features between several structural distinct HIV reverse transcriptase (HIVRT) inhibitors. A similar concept related to pharmacophoric similarity is molecular filed similarity as implemented in the FieldAlign tool (by the CRESSET company), which uses energetic probes to identify similar ligands that might not have explicit structural overlap [46]. Recently, Ferreira and Couto developed a new similarity measure called chemical sematic similarity to classify chemical compounds based on their semantic characterization such as drug annotation in the ChEMBL database [47]. The study showed that comparing compounds by their functional roles improved predictions of several drug properties by complementing existing compound classification systems.

Analog analysis seeks to characterize chemical transformations, which are defined over pairs of molecules. Recently, the matched molecular pairs (MMP) formalism has emerged as a way to define a specific type of transformation or relationship, non-ring single-bond substitutions and facilitate the development of methods for indexing and searching analog relationships [48]. The fragment-indexing algorithm developed by Hussain and Rea [48] is currently the most widely used MMP search method but does not support a similarity search. Rensi and Altman developed a method for computing the similarity of chemical transformations using Tanimoto kernel embedded fingerprints and extended a fuzzy search capability to the MMP framework [49]. They demonstrated the capability to query MMP relationships at multiple levels of contextual abstraction with stable results over a range of dataset sizes of over four orders of magnitude from 103 high-impact pharmacological targets.

Machine learning models in QSAR

Machine learning techniques can be broadly classified as supervised or unsupervised learning [50] (Table 2). For the supervised learning, labels are assigned to the training data and, once trained, the model can predict labels for given data inputs. Supervised machine learning models include regression analysis, k-nearest neighbor (kNN), Bayesian probabilistic learning, SVMs, random forests and neural networks. Unsupervised machine learning techniques learn underlying patterns of molecular features directly from unlabeled data. A special case of supervised learning is semi-supervised learning or tranductive learning, in which a small amount of labeled data is mixed with labeled data in the training process to improve the learning accuracy for modeling a small and unbalanced dataset [51]. Unsupervised methods include dimensionality reduction techniques such as principal components analysis (PCA), independent components analysis (ICA) and several supervised methods that can also support unsupervised learning, such as SVMs, probabilistic graphical models and neural networks [52– 55]. Clustering algorithms represent another family of unsupervised algorithms, where the dataset is first divided by predefined distance metrics in high-dimensional space and the labels are later assigned based on the number of observed categories. Modern machine learning techniques offer a powerful suite of techniques to explore nonlinear SAR relationships with high accuracy and precision.

Naive Bayes

Naive Bayes classifiers are probabilistic models based on Bayes' rule [56–58]. They estimate the probability that a given item of data is correctly assigned to a certain label based on the prior probability distribution (priors) representing the relative proportions of labels in training sets. If multiple labels are presented, then the probability associated with each label is conditionally independent. A well-known example of this approach is the PASS program for predicting drug activities [59]. In the PASS program, the priors are first established for a set of biological active compounds based on the proportions of chemical substructures in the active and inactive class. Then, a variant of Naive Bayes is used to estimate the drug activity based on the query structures using the prior probability distribution. Chen *et al.* demonstrated their efficiency in large-scale virtual screens for important pharmacological properties such as cytochrome P450 inhibition, human plasma protein binding and bioavailability in animal models (*rattus norvegicus*) [60].

Regression analysis

Regression analysis can refer to linear regression modeling for continuous data or logistic regression analysis modeling for categorical data [61]. Given a set of training data points, the goal of linear regression analysis is to find a linear function of a set of predictor variables, such that the fitted line minimizes the distances to the data points along the dimensions of a set of outcome variables. Early QSAR techniques like Hansch and Free–Wilson analysis make extensive use of multivariate linear regression. However, correlations between features and high-dimensional feature spaces present challenges for the application of linear regression models in QSAR. Several techniques such as regularization, dimensionality reduction and genetic algorithms are available to combat the twin curses of dimensionality and collinearity, which result in model overfitting and coefficient coupling which confound accuracy and interpretability [62]. L1 regularization methods and evolutionary algorithms shrink the number of variables explicitly by selecting small subsets that are most relevant to the outcome being predicted by the QSAR model [63]. By contrast, L2 regularization methods like Gaussian processes and ridge regression reduce the 'effective' number of variables (VC-dimensionality) without changing their actual number [64,65]. Recently, Algamal *et al.* demonstrated the utility of an adaptive least absolute shrinkage and selection operator (LASSO) variable selection approach for predicting the anticancer potency of imidazo-pyridine derivatives [66]. In another study, Helguera *et al.*

used evolutionary variable selection to model the activity and selectivity of monoamine oxidase inhibitors [67]. By contrast, dimensionality reduction techniques such as principal components analysis (PCA) transform large sets of correlated variables into smaller sets of uncorrelated features [68]. In a seminal study on QSAR classification, Gao *et al.* used PCA to decorrelate features for prediction of estrogen receptor binding [69]. More recently, Rensi and Altman demonstrated performance improvements over LASSO regression for predicting activity against a broad set of pharmacological protein targets using kernel principal components analysis, and nonlinear variant of PCA [70]. Another popular regression method is partial least squares (PLS), which couples dimensionality reduction with multivariate regression to transform predictors into uncorrelated variables that are maximally correlated with the activity or property of interest. Erikkson *et al.* recommend PLS as a first-line approach to QSAR modeling for its superior efficiency and accuracy relative to explicitly combing unsupervised dimensionality with multivariate regression, and PLS is used extensively in 3D-QSAR [5]. However, tight coupling of dimensionality reduction and model fitting can limit utility in unsupervised or semi-supervised problems where knowledge of the outcome variable is missing or incomplete. Although linear regression analysis has been successfully applied in many drug optimization problems, underlying linearity and vector space assumptions which are not valid for most QSAR problems are a significant limitation. Thus, careful selection of the features and modeled system, although crucial, is sometimes insufficient to ensure the success of linear regression models.

k-Nearest neighbors

In kNN, the data containing labeled and unlabeled nodes are represented in a high-dimensional feature space and the labels from the closest nodes are transferred to the query using a majority-voting rule [71,72]. Here, the value k specifies the number of closest neighbors participating in the voting system. kNN in ligand-based virtual screening can be thought of as an extension of chemical similarity search to supervised learning, where a chemical similarity metric such as Tc is used as a measure of distance between compounds, and bioactivities are predicted from the top search results.

However, there is no principled way of choosing the number of nearest neighbors to use, and values of k that are too high or low can yield unfavorable false-positive or false-negative rates. This was addressed by the similarity ensemble approach (SEA), which compares chemical similarity values to a randomized background score similar to that used in a BLAST sequence similarity search [73]. Lo *et al.* proposed another approach for large-scale compound drug-target profiling called chemical similarity network analysis pull-down (CSNAP) [74]. Instead of defining nearest neighbor values, the CSNAP approach used a threshold network to cluster compounds based on a predefined Tc cut-off. After an initial clustering step, query compounds were assigned the most probable drug targets by ranking the shared targets among the first-order neighbors. Recently, Huang *et al.* developed the most-similar ligand-based target inference (MOST) approach which utilizes explicit bioactivity of the most-similar ligands to predict targets of the query compound [75]. They showed that the MOST approach could alleviate false-positive predictions associated with the common nearest neighbor similarity search.

Random forest

Random forest is an ensemble learning method where multiple decision trees are built based on the training data and a majority-voting scheme similar to kNN is used to make classification or regression predictions for new inputs [57]. Svetnik *et al.* demonstrated the utility of random forest models in QSAR classification and regression for a number of important pharmacological transporters, targets and properties such as P-glycoprotein (PGP), cyclooxegenase-2 (COX2) and blood–brain barrier permeability [76]. They achieved accuracy comparable to SVMs and neural networks with superior interpretability.

Support vector machines

SVMs solve the classification problem by using nonlinear kernel functions to map data into high-dimensional space by finding an optimally separating hyperplane [77]. The hyperplane is fit to maximize the margin between support vectors, points nearest to the decision boundary and is expressed as a linear combination of data points. Liu *et al.* used SVMs in a QSAR study of transcription factors activator protein (AP)-1 and nuclear factor (NF)-κB by ethyl 2-[(3-methyl-2,5-dioxo(3-pyrrolinyl))amino]-4-(trifluoromethyl)pyrimidine-5-carboxylate derivatives [78]. More recently, Nekoei *et al.* combined a genetic variable selection variable approach with SVMs to identify a number of structural features of aminopyrimidine-5-carbaldehyde oxime derivatives that are responsible for strong vascular endothelial growth factor (VEGF)-2 inhibition activity [79].

Neural networks and deep learning

Artificial neural networks (ANNs) are a family of machine learning algorithms, inspired by the operations of neurons in the brain [80]. Each neuron in an ANN receives numerous input signals (analogous to dendrites), performs a weighted sum of the inputs, generates an activation response through a nonlinear activation function (analogous to cell body) and passes the output signals to subsequent connected neurons (analogous to axons). Multilayer ANNs can be constructed by organizing neurons into different layers and connecting neurons in consecutive layers. The combination of nonlinear units enables ANNs to learn highly complex functions of the inputs. ANNs have been widely applied to all branches of chemoinformatics, including modeling QSAR/QSPR properties of small molecules as well as performing pharmacokinetic and pharmacodynamic analysis [81–83]. We refer the readers to the work by Baskin *et al.* for a latest comprehensive review of ANN-based methods in chemoinformatics [84].

Deep learning networks are a recent extension of ANNs, which utilize deep and specialized architectures to learn useful features from raw data [85]. The recent success of deep learning provides an opportunity to develop tools for automatically extracting task-specific representations of chemical structures. Deep convolutional neural networks (CNNs) comprise a subclass of deep learning networks [86,87]. In CNNs, local filters scan through the input space to search for recurring local spatial patterns that are useful for the classification performance. Owing to unique local spatial properties of images, CNNs have achieved great success in the computer vision community [87,88], and have recently been applied to the biomedical field. For example, Torng and Altman viewed protein structures as

‘3D images’ with four different atom-type channels, and used 3D-CNNs to analyze amino acid microenvironment similarities and to predict effects of mutations in proteins [89].

Graph convolutional networks (GCNs) are variants of CNNs that have been commonly applied to 2D molecular graph analysis. GCNs employ similar concepts of local spatial filters, but operate on graphs, to learn features from graph neighborhoods. Following the first application of GCNs in QSAR analysis by Baskin *et al.* [90], different graph convolutional architectures for learning small molecule representations have also been proposed, each defining local graph neighborhoods and convolution operations in different ways. For example, Duvenaud *et al.* used different ‘degree filters’ to learn features for nodes with different degrees [32]. Kearns *et al.* employed ‘Weave modules’, that integrate information from all atoms and atom pairs to learn molecular features [91]. More recently, Hechtlinger *et al.* used random walks to define the local neighborhood of each node in a graph [92].

Recurrent neural networks (RNNs) are another major family of deep neural networks that have been widely used in natural language processing [93]. Long-short-term-memory (LSTM) networks are a subclass of RNNs that use gated units and memory cells to capture long- and short-term temporal dependencies within input sequences [94]. LSTM networks have been applied to *de novo* drug design, where the LSTM model is trained to learn ‘grammatical structures’ within SMILES strings and to output novel molecules following the learned rules [95]. Variational autoencoders (VAEs) [96], generative adversarial networks (GANs) [97] and deep reinforcement learning [98] have also been applied to learning latent representations of molecules [99] and to generating new compounds with desired molecular properties [100,101].

QSAR modeling

The general protocol for constructing QSAR models for drug discovery has been systematized and consists of several modular steps involving the chemoinformatics and machine learning techniques previously discussed. The first step is ‘molecular encoding’ where the chemical features and properties are derived from chemical structures or lookup of experimental results. Second, a feature selection step is performed where unsupervised learning techniques are used to identify the most relevant properties and reduce the dimensionality of the feature vector. Finally, in the learning phase, a supervised machine learning model is applied to discover an empirical function (either explicitly or implicitly) that can achieve an optimal mapping between the input feature vectors and the biological responses. Building an accurate QSAR model also requires careful consideration and selection of the SAR datasets used for training and model validation [102]. This includes strict separation of training and test sets for initial model creation and the test sets for final model performance evaluation. The performances of the QSAR models are commonly evaluated by standard metrics such as sensitivity, specificity, precision and recall. For unbalanced datasets, area-under-curve (AUC) derived from receiver-operating-characteristics (ROC) curves can be used. Although 3D-QSAR methods like CoMFA consider structural conformation, the approach necessitates substantial computation resource and is subject to uncertainty generated from conformation prediction, ligand orientation and structural alignment. Thus, the 2D-QSAR model can be competitive and sometimes even superior to 3D-QSAR approaches [42,103].

Concluding remarks and future directions

Machine learning techniques have been widely applied in the field of chemoinformatics to discover and design new drugs with superior biological activities. Mathematical mining of chemical graphs enables the derivation of a constellation of 2D or 3D chemical descriptors, which are packaged as chemical fingerprints in a diverse array of machine learning models and predictive tasks. A key area of innovation in the field is the marriage of big data and machine learning to predict wider ranges of biological phenomena. Traditional drug design methods based on simple ligand–protein interactions are no longer sufficient for meeting clinical drug safety criteria. High drug attrition rates from severe side effects often involve biological pathways and systematic responses at higher levels. Consequently, incorporating multiple data types and sources, also known as ‘data fusion’ techniques, that aggregate structural, genetic and pharmacological data from the molecular to organism level, will be crucial for the discovery of safer and more-effective drugs [104]. Likewise, novel machine learning models capable of processing big data at high volume, velocity and veracity with great versatility are also needed. Recent evolution in deep learning networks has proven to be a promising architecture for efficient learning from massive datasets for modern drug discovery campaigns [105]. Other aspects of machine learning techniques such as increased data interpretability to prove mechanistic hypothesis as well as methods preventing overfitting are also important topics that warrant further development in the field of machine-learning-based drug discovery.

Acknowledgments

We thank all members of the Helix group at Stanford University for their helpful feedback and suggestions. The project was supported by Stanford Dean’s Postdoctoral Fellowship, Genentech, Pfizer and the following funding sources: NIH GM102365 and FDA U01FD004979.

References

1. Varnek A, Baskin I. Machine learning methods for property prediction in chemoinformatics: Quo Vadis? J Chem Inf Model. 2012; 52:1413–1437. [PubMed: 22582859]
2. Ali SM, et al. Butitaxel analogues: synthesis and structure-activity relationships. J Med Chem. 1997; 40:236–241. [PubMed: 9003522]
3. Cherkasov A, et al. QSAR modeling: where have you been? Where are you going to? J Med Chem. 2014; 57:4977–5010. [PubMed: 24351051]
4. Kubinyi H. Free Wilson analysis. Theory, applications and its relationship to Hansch analysis. Quantitative Structure–Activity Relationships. 1988; 7:121–133.
5. Gasteiger J, editorHandbook of Chemoinformatics: from Data to Knowledge. Wiley-VCH; 2003.
6. Varnek A, Baskin II. Chemoinformatics as a theoretical chemistry discipline. Mol Inform. 2011; 30:20–32. [PubMed: 27467875]
7. Bajorath JR, editorChemoinformatics and Computational Chemical Biology. Humana Press; 2011.
8. Kapetanovic IM. Computer-aided drug discovery and development (CADDD): *insilico- chemico*-biological approach. Chem Biol Interact. 2008; 171:165–176. [PubMed: 17229415]
9. Bonchev D, Rouvray DH, editorsChemical Graph Theory: Introduction and Fundamentals. Abacus Press; 1991.

10. Garcia-Domenech R, et al. Some new trends in chemical graph theory. Chem Rev. 2008; 108:1127–1169. [PubMed: 18302420]
11. Trinajstić N, editorChemical Graph Theory. CRC Press; 1983.
12. Cormen TH, Cormen TH, editorsIntroduction to Algorithms. MIT Press; 2001.
13. Fourches D, Tropsha A. Using graph indices for the analysis and comparison of chemical datasets. Mol Inform. 2013; 32:827–842. [PubMed: 27480235]
14. Khan AU. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discov Today. 2016; 21:1291–1302. [PubMed: 27326911]
15. Testa B, Seiler P. Steric and lipophobic components of the hydrophobic fragmental constant. Arzneimittelforschung. 1981; 31:1053–1058. [PubMed: 7196751]
16. Hansch C, , et al., editorsExploring QSAR. American Chemical Society; 1995.
17. Consonni V, Todeschini R, editorsHandbook of Molecular Descriptors. Wiley-VCH; 2000.
18. Bajorath J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. J Chem Inf Comput Sci. 2001; 41:233–245. [PubMed: 11277704]
19. Hong H, et al. Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J Chem Inf Model. 2008; 48:1337–1344. [PubMed: 18564836]
20. Sawada R, et al. Benchmarking a wide range of chemical descriptors for drug–target interaction prediction using a chemogenomic approach. Mol Inform. 2014; 33:719–731. [PubMed: 27485418]
21. Chavan S, et al. Towards global QSAR model building for acute toxicity: Munro database case study. Int J Mol Sci. 2014; 15:18162–18174. [PubMed: 25302621]
22. Saiz-Urra L, et al. Quantitative structure-activity relationship studies of HIV-1 integrase inhibition. 1. GETAWAY descriptors. Eur J Med Chem. 2007; 42:64–70. [PubMed: 17030481]
23. Karelson M, et al. Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev. 1996; 96:1027–1044. [PubMed: 11848779]
24. Kubinyi H, , et al., editors3D QSAR in Drug Design. Kluwer Academic; 1998.
25. Sliwoski G, et al. Autocorrelation descriptor improvements for QSAR: 2DA_Sign and 3DA_Sign. J Comput Aided Mol Des. 2016; 30:209–217. [PubMed: 26721261]
26. Hu Y, et al. Recent advances in scaffold hopping. J Med Chem. 2017; 60:1238–1246. [PubMed: 28001064]
27. Andrade CH, et al. 4D-QSAR: perspectives in drug design. Molecules. 2010; 15:3281–3294. [PubMed: 20657478]
28. Ash J, Fourches D. Characterizing the chemical space of ERK2 kinase inhibitors using descriptors computed from molecular dynamics trajectories. J Chem Inf Model. 2017; 57:1286–1299. [PubMed: 28471171]
29. Raymond JW, Willett P. Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. J Comput Aided Mol Des. 2002; 16:59–71. [PubMed: 12197666]
30. Yeo WK, et al. Extraction and validation of substructure profiles for enriching compound libraries. J Comput Aided Mol Des. 2012; 26:1127–1141. [PubMed: 22983491]
31. Heikamp K, Bajorath J. Large-scale similarity search profiling of ChEMBL compound data sets. J Chem Inf Model. 2011; 51:1831–1839. [PubMed: 21728295]
32. Duvenaud DK, , et al. Convolutional networks on graphs for learning molecular fingerprints. In: Jordan MI, , et al., editorsAdvances in Neural Information Processing Systems. MIT Press; 2015. 2224–2232.
33. Verma J, et al. 3D-QSAR in drug design--a review. Curr Top Med Chem. 2010; 10:95–115. [PubMed: 19929826]
34. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem. 1985; 28:849–857. [PubMed: 3892003]
35. Cramer RD, et al. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc. 1988; 110:5959–5967. [PubMed: 22148765]
36. Baskin II, Zhokhova NI. The continuous molecular fields approach to building 3D-QSAR models. J Comput Aided Mol Des. 2013; 27:427–442. [PubMed: 23719959]

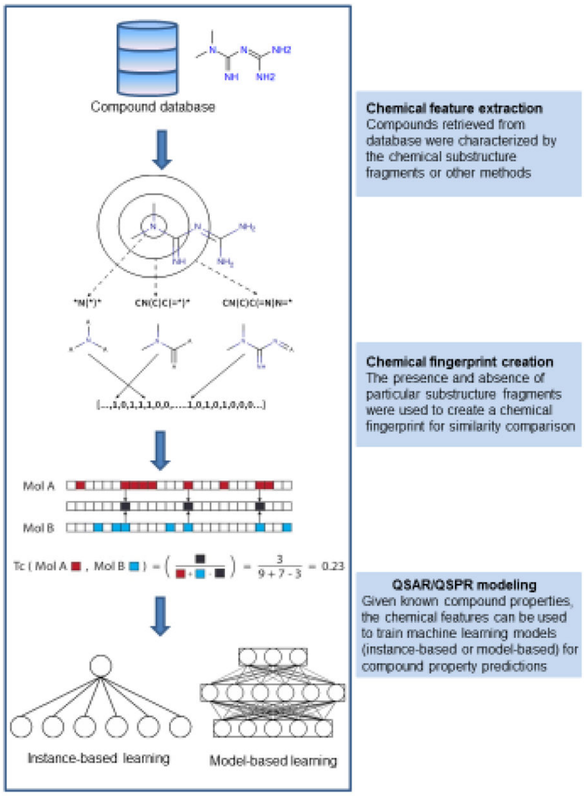
37. Sheridan RP, Kearsley SK. Why do we need so many chemical similarity search methods? Drug Discov Today. 2002; 7:903–911. [PubMed: 12546933]
38. Maldonado AG, et al. Molecular similarity and diversity in chemoinformatics: from theory to applications. Mol Divers. 2006; 10:39–79. [PubMed: 16404528]
39. Bajorath J. Molecular similarity concepts for informatics applications. Methods Mol Biol. 2017; 1526:231–245. [PubMed: 27896745]
40. Hu Y, et al. Advancing the activity cliff concept. F1000Res. 2013; 2:199. [PubMed: 24555097]
41. Stumpfe D, et al. Advancing the activity cliff concept, part II. F1000Res. 2014; 3:75. [PubMed: 24741442]
42. Hu G, et al. Performance evaluation of 2D fingerprint and 3D shape similarity methods in virtual screening. J Chem Inf Model. 2012; 52:1103–1113. [PubMed: 22551340]
43. Rush TS 3rd, et al. A shape-based 3-D scaffold hopping method and its application to a bacterial protein–protein interaction. J Med Chem. 2005; 48:1489–1495. [PubMed: 15743191]
44. Lo YC, et al. 3D chemical similarity networks for structure-based target prediction and scaffold hopping. ACS Chem Biol. 2016; 11:2244–2253. [PubMed: 27285961]
45. Lo YC, et al. Computational cell cycle profiling of cancer cells for prioritizing FDA-approved drugs with repurposing potential. Sci Rep. 2017; 7:11261. [PubMed: 28900159]
46. Cheeseright TJ, et al. FieldScreen: virtual screening using molecular fields. Application to the DUD data set. J Chem Inf Model. 2008; 48:2108–2117. [PubMed: 18991371]
47. Ferreira JD, Couto FM. Semantic similarity for automatic classification of chemical compounds. PLoS Comput Biol. 2010.
48. Hussain J, Rea C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model. 2010; 50:339–348. [PubMed: 20121045]
49. Rensi S, Altman RB. Flexible analog search with kernel PCA embedded molecule vectors. Comput Struct Biotechnol J. 2017; 15:320–327. [PubMed: 28458783]
50. Nasrabadi NM. Pattern recognition and machine learning. J Electronic Imaging. 2007; 16:049901.
51. Kondratovich E, et al. Transductive support vector machines: promising approach to model small and unbalanced datasets. Mol Inform. 2013; 32:261–266. [PubMed: 27481521]
52. Hyvarinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw. 2000; 13:411–430. [PubMed: 10946390]
53. Chuprina A, et al. Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. J Chem Inf Model. 2010; 50:470–479. [PubMed: 20297844]
54. MacCuish JD, MacCuish NE. Chemoinformatics applications of cluster analysis. Comput Mol Sci. 2014; 4:34–48.
55. Akella LB, DeCaprio D. Cheminformatics approaches to analyze diversity in compound screening libraries. Curr Opin Chem Biol. 2010; 14:325–330. [PubMed: 20457001]
56. Bender A, et al. “Bayes affinity fingerprints” improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept? J Chem Inf Model. 2006; 46:2445–2456. [PubMed: 17125186]
57. Schierz AC. Virtual screening of bioassay data. J Cheminform. 2009; 1:21. [PubMed: 20150999]
58. Hert J, et al. New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. J Chem Inf Model. 2006; 46:462–470. [PubMed: 16562973]
59. Poroikov VV, et al. Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds. J Chem Inf Comput Sci. 2000; 40:1349–1355. [PubMed: 11128093]
60. Chen B, et al. Comparison of random forest and Pipeline Pilot Naive Bayes in prospective QSAR predictions. J Chem Inf Model. 2012; 52:792–803. [PubMed: 22360769]
61. Marill KA. Advanced statistics: linear regression, part II: multiple linear regression. Acad Emerg Med. 2004; 11:94–102. [PubMed: 14709437]
62. Kubinyi H. Evolutionary variable selection in regression and PLS analyses. J Chemometrics. 1996; 10:119–133.

63. Frank lEFriedman JH. A statistical view of some chemometrics regression tools. Technometrics. 1993; 35:109–135.
64. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970; 12:55–67.
65. Seeger M. Gaussian processes for machine learning. Int J Neural Syst. 2004; 14:69–106. [PubMed: 15112367]
66. Algamal ZY, et al. High-dimensional QSAR prediction of anticancer potency of imidazo [4,5-b] pyridine derivatives using adjusted adaptive LASSO. J Chemometrics. 2015; 29:547–556.
67. Helguera AM, et al. Combining QSAR classification models for predictive modeling of human monoamine oxidase inhibitors. Eur J Med Chem. 2013; 59:75–90. [PubMed: 23207409]
68. Owen JR, et al. Visualization of molecular fingerprints. J Chem Inf Model. 2011; 51:1552–1563. [PubMed: 21696145]
69. Gao H, et al. Binary quantitative structure–activity relationship (QSAR) analysis of estrogen receptor ligands. J Chem Inf Comput Sci. 1999; 39:164–168. [PubMed: 10094611]
70. Rensi SE, Altman RB. Shallow representation learning via kernel PCA improves QSAR modelability. J Chem Inf Model. 2017; 57:1859–1867. [PubMed: 28727421]
71. Khanfar MA, Taha MO. Elaborate ligand-based modeling coupled with multiple linear regression and k nearest neighbor QSAR analyses unveiled new nanomolar mTOR inhibitors. J Chem Inf Model. 2013; 53:2587–2612. [PubMed: 24050502]
72. Sahigara F, et al. Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. J Cheminformatics. 2013; 5:27.
73. Keiser MJ, et al. Relating protein pharmacology by ligand chemistry. Nat Biotechnol. 2007; 25:197–206. [PubMed: 17287757]
74. Lo YC, et al. Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. PLoS Comput Biol. 2015; 11:e1004153. [PubMed: 25826798]
75. Huang T, et al. MOST: most-similar ligand based approach to target prediction. BMC Bioinformatics. 2017; 18:165. [PubMed: 28284192]
76. Svetnik V, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci. 2003; 43:1947–1958. [PubMed: 14632445]
77. Noble WS. What is a support vector machine? Nat Biotechnol. 2006; 24:1565–1567. [PubMed: 17160063]
78. Liu H, et al. QSAR study of ethyl 2-[(3-methyl-2, 5-dioxo (3-pyrrolinyl)) amino]-4-(trifluoromethyl) pyrimidine-5-carboxylate: an inhibitor of AP-1 and NF-κB mediated gene expression based on support vector machines. J Chem Inf Comput Sci. 2003; 43:1288–1296. [PubMed: 12870922]
79. Nekoei M, et al. QSAR study of VEGFR-2 inhibitors by using genetic algorithm-multiple linear regressions (GA-MLR) and genetic algorithm-support vector machine (GA-SVM): a comparative approach. Med Chem Res. 2015; 24:3037–3046.
80. Zurada JM, editorIntroduction to Artificial Neural Systems. West St. Paul: 1992.
81. Myint K-Z, et al. Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. Mol Pharm. 2012; 9:2912–2923. [PubMed: 22937990]
82. Devillers J. Prediction of mammalian toxicity of organophosphorus pesticides from QSTR modeling. SAR QSAR Environ Res. 2004; 15:501–510. [PubMed: 15669705]
83. Gobburu JV, Chen EP. Artificial neural networks as a novel approach to integrated pharmacokinetic-pharmacodynamic analysis. J Pharm Sci. 1996; 85:505–510. [PubMed: 8742942]
84. Baskin II, et al. A renaissance of neural networks in drug discovery. Expert Opin Drug Discov. 2016; 11:785–795. [PubMed: 27295548]
85. LeCun Y, et al. Deep learning. Nature. 2015; 521:436–444. [PubMed: 26017442]
86. LeCun Y, , et al. Handwritten digit recognition with a back-propagation network. In: Jordan MI, , et al., editorsAdvances in Neural Information Processing Systems. MIT Press; 1990. 396–404.
87. Krizhevsky A, , et al. Imagenet classification with deep convolutional neural networks. In: Jordan MI, , et al., editorsAdvances in Neural Information Processing Systems. MIT Press; 2012. 1097–1105.

88. Szegedy C, , et al. Going deeper with convolutions. IEEE; 2015.
89. Torng W, Altman RB. 3D deep convolutional neural networks for amino acid environment similarity analysis. BMC Bioinformatics. 2017; 18:302. [PubMed: 28615003]
90. Baskin II, et al. A neural device for searching direct correlations between structures and properties of chemical compounds. J Chem Inf Comput Sci. 1997; 37:715–721.
91. Kearnes S, et al. Molecular graph convolutions: moving beyond fingerprints. J Comput Aided Mol Des. 2016; 30:595–608. [PubMed: 27558503]
92. Hechtlinger Y, , et al. A generalization of convolutional neural networks to graph-structured data. 2017. Available at: https://arxiv.org/pdf/1704.08165.pdf
93. Bahdanau D, et al. Neural machine translation by jointly learning to align and translate. 2014 arXiv:1409.0473.
94. Graves A, , et al. Speech recognition with deep recurrent neural networks. Acoustics, Speech and Signal Processing, 2013 IEEE International Conference; 2013. 6645–6649.
95. Segler MH, , et al. Generating focused molecule libraries for drug discovery with recurrent neural networks. 2017. Available at: https://arxiv.org/pdf/1701.01329.pdf
96. Kingma DP, Welling M. Auto-encoding variational bayes. 2013 arXiv:1312.6114.
97. Goodfellow I, , et al. Generative adversarial nets. In: Jordan MI, , et al., editorsAdvances in Neural Information Processing Systems. MIT Press; 2014. 2672–2680.
98. Mnih V, et al. Human-level control through deep reinforcement learning. Nature. 2015; 518:529. [PubMed: 25719670]
99. Kusner MJ, et al. Grammar variational autoencoder. 2017 arXiv:1703.01925.
100. Kadurin A, et al. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties. in silico Mol Pharm. 2017; 14:3098–3104. [PubMed: 28703000]
101. Olivecrona M, et al. Molecular *de-novo* design through deep reinforcement learning. J Cheminformatics. 2017; 9:48.
102. Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform. 2010; 29:476–488. [PubMed: 27463326]
103. Nettles JH, et al. Bridging chemical and biological space: “target fishing” using 2D and 3D molecular descriptors. J Med Chem. 2006; 49:6802–6810. [PubMed: 17154510]
104. Searls DB. Data integration: challenges for drug discovery. Nat Rev Drug Discov. 2005; 4:45–58. [PubMed: 15688072]
105. Chen H, , et al. The rise of deep learning in drug discovery. Drug Discov Today. 2018.

Highlights

- Chemical graph theory and descriptors in drug discovery
- Chemical fingerprint and similarity analysis
- Machine learning models for virtual screening
- Future challenges and direction in machine-learning-based drug discovery



Chemical feature extraction
Compounds retrieved from database were characterized by the chemical substructure fragments or other methods

Chemical fingerprint creation
The presence and absence of particular substructure fragments were used to create a chemical fingerprint for similarity comparison

QSAR/QSPR modeling
Given known compound properties, the chemical features can be used to train machine learning models (instance-based or model-based) for compound property predictions

Figure 1.

Computational workflow for chemoinformatics analysis using machine learning. The first step of chemoinformatics analysis is feature extraction, through which the compound is characterized by substructure fragments or other chemical descriptors (first box). The chemical features of the compound are represented by chemical fingerprints and applied for compound similarity comparison based on the presence and absence of shared chemical features. The chemical fingerprint can be used for predicting other chemical and physiochemical properties in QSAR/QSPR analysis using diverse machine learning models including making inference from the training data by comparison (instance-based learning) or from the trained statistical model (model-based learning) (second box).

Table 1

Common chemical descriptors for QSAR/QSPR analysis

Chemical descriptors	**Based on**	**Examples**
Theoretical descriptors		
0D	Molecular formula	Molecular weights, atom counts, bond counts
1D	Chemical graph	Fragment counts, functional group counts
2D	Structural topology	Weiner index, Balaban index, Randic index, BCUTS
3D	Structural geometry	WHIM, autocorrelation, 3D-MORSE, GETAWAY
4D	Chemical conformation	Volsurf, GRID, Raptor
Experimental descriptors		
Hydrophobic parameters	Hydrophobicity	Partition coefficents (logP), hydrohobic substituent constant (π)
Electronic parameters	Electronic properties	Acid dissociation constant, Hammett constant
Steric parameters	Steric properties	Taft steric constant, Charton’s constant

Table 2

Summary of machine learning methods

Methods+A3:C13	Descriptions	Refs
Supervised learning		
Multiple regression analysis	A statistical process to find relationships between dependent variables and one or more independent variables	[61]
k-nearest neighbor	An instance-based learning where an object is classified by the majority rule among its k nearest neighbor, where k is an integer	[72]
Naive bayes	A probablistic approach that uses probability prior and Bayes rule to predict membership by assuming feature independency	[58]
Random forest	A classification technique based on the essemble of multiple decision trees and majority voting rules	[76]
Neural network and deep learning	A model-based learning method that learns from input data based on layers of connected neurons consisting of input layers, multiple hidden layers (for deep learning) and output layers	[85]
Support vector machine	A statistical method that maps data into high-dimensional space to identify a lower dimensional hyperplane that maximizes the data separation using a nonlinear kernel. This is achieved by maximizing the margins between hyperplanes known as support vectors	[77]
Unsupervised learning		
k-means clustering	A classsification method that classifies data into k groups by minimizing within-group distances to the centroid	[54]
Hierarchical clustering	A classification method that builds a hierachy of clusters by agglomerative clustering e.g., merging smaller clusters or divisive clustering e.g., splitting a large cluster to smaller ones	[54]
Principal component analysis	A statistical method that uses orthogonal procedure to transform a set of correlated features to new independent variables called principal components	[55]
Independent component analysis	A statistical method that separates a multivariable output into statistical independent additive components	[52]