C/e: AN C/e QA: NA

EXPERT OPINION ON DRUG DISCOVERY, 2016
http://dx.doi.org/10.1080/17460441.2016.1201262

Taylor & Francis
Taylor & Francis Group

REVIEW

A renaissance of neural networks in drug discovery

Igor I. Baskin[a], David Winkler[b,c,d,e] and Igor V. Tetko [f,g]

[a]Faculty of Physics, M.V. Lomonosov Moscow State University, Moscow, Russia; [b]CSIRO Manufacturing, Melbourne, VIC, Australia; [c]Monash Institute for Pharmaceutical Sciences, Monash University, Parkville, VIC, Australia; [d]Latrobe Institute for Molecular Science, Bundoora, VIC, Australia; [e]School of Chemical and Physical Sciences, Flinders University, Bedford Park, SA, Australia; [f]Helmholtz Zentrum München – German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Neuherberg, Germany; [g]BigChem GmbH, Neuherberg, Germany

ABSTRACT

Introduction: Neural networks are becoming a very popular method for solving machine learning and artificial intelligence problems. The variety of neural network types and their application to drug discovery requires expert knowledge to choose the most appropriate approach.

Areas covered: In this review, the authors discuss traditional and newly emerging neural network approaches to drug discovery. Their focus is on backpropagation neural networks and their variants, self-organizing maps and associated methods, and a relatively new technique, deep learning. The most important technical issues are discussed including overfitting and its prevention through regularization, ensemble and multitask modeling, model interpretation, and estimation of applicability domain. Different aspects of using neural networks in drug discovery are considered: building structure-activity models with respect to various targets; predicting drug selectivity, toxicity profiles, ADMET and physicochemical properties; characteristics of drug-delivery systems and virtual screening.

Expert opinion: Neural networks continue to grow in importance for drug discovery. Recent developments in deep learning suggests further improvements may be gained in the analysis of large chemical data sets. It's anticipated that neural networks will be more widely used in drug discovery in the future, and applied in non-traditional areas such as drug delivery systems, biologically compatible materials, and regenerative medicine.

ARTICLE HISTORY
Received 30 March 2016
Accepted 9 June 2016
Published online xx xxx xxxx

KEYWORDS
Deep learning; neural network ensembles; neural networks; overfitting; structure-activity relationships

1. Introduction

No other machine-learning method has such a long and rich history full of great hope and deep frustration as artificial neural networks (ANNs). McCulloch and Pitts [1], in the 1940s, attempted to create a mathematical model of the human brain. Following the important development of the perceptron, the first algorithm for pattern recognition by a two-layer ANN was proposed by Rosenblatt [2]. However, as this was unable to simulate the basic exclusive-or operation, a period of stagnation of neural network research ensued. Neural network research revived following the invention (and several independent reinventions) of the backpropagation algorithm [3], offering an efficient solution to the exclusive-or problem. Neural networks became very popular in the mid-1980s due to the concept of parallel distributed processing (connectionism) popularized by Rumelhart and McClelland, the development of neocognitron (the first convolutional ANN) by Fukushima [4], self-organizing maps by Kohonen [5], and energy-based recurrent ANNs by Hopfield [6]. This optimism was followed by the second period where ANNs were in competition with some newly emerged, very efficient, and mathematically well-grounded methods. Very recently, ANNs received another stimulus due to the development of the deep-learning concept by Hinton and colleagues [7–9]. These methods may outperform alternative state-of-the-art machine-learning methods in drug data modeling benchmarking competitions. In addition, deep learning has achieved human-competitive and higher performance on several important image and speech recognition benchmarks and has the potential to revolutionize machine learning and artificial intelligence.

The first application of ANNs to drug discovery dates back to the early 1970s when Hiller et al. [10] published a study using the Rosenblatt perceptron to classify substituted 1,3-dioxanes as physiologically active or inactive. In this work, elements of the chemical structures were projected onto the perceptron retina; the perceptron was trained using a set of compounds with known activities, and the trained neural network demonstrated good recognition ability on both the training and the test sets of compounds. The next stage of development occurred in 1990 with the first publications of Aoyama et al. dealing with the use of ANNs in Quantitate Structure–Activity Relationship (QSAR) studies [11]. For the last 25 years, this approach to modeling structure–activity relationships has matured into a well-established scientific field with numerous theoretical approaches and successful practical applications (see review articles [12–16]). The field now encompasses the use of ANNs for predicting not only different types of biological activity but also physicochemical, Absorption Distribution Metabolism Excretion and Toxicity

CONTACT Igor V. Tetko itetko@vcclab.org Helmholtz Zentrum München – German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstaedter Landstrasse 1, b. 60w, D-85764, Neuherberg, Germany.

Article highlights

- Backpropagation neural networks are universal approximators for structure-activity relationships
- Different regularization techniques efficiently prevent overfitting and enhance predictive performance
- Bayesian regularized neural networks are a reliable and effective tool with numerous applications in medicinal chemistry and materials design
- Associative neural networks use ensemble modeling to increase and predictive ability of structure-activity models and assess the reliability of prediction
- Deep learning involves formation of different levels of data representation
- Deep neural networks could particularly be useful for analyzing huge amounts of chemical and biological information for drug discovery, although they are computationally demanding.

This box summarizes key points contained in the article.

(ADMET), biodegradability and spectroscopic properties, and reactivity. The aim of this article is to review some important concepts and ideas accumulated in this field and to provide a guide to where the field is heading in the future.

2. Backpropagation neural networks

Multilayer feed-forward neural networks, also known as multi-layer perceptrons, comprise the most widely used architecture for ANNs (see Figure 1). They consist of units implementing the McCulloth–Pitts' model of neurons [1], which produce their output by computing the weighted sum of their inputs followed by a nonlinear transform (see Figure 1).

$$y = f(Z) = f\left(-t + \sum\nolimits_i w_i x_i\right),$$

where x_i is ith input of the unit, w_i is the corresponding adjustable weight mimicking the synaptic strength of biological neuron, z is the overall input to the unit, whereas $f(z)$ is a nonlinear transform function that could be associated with the activation of neural cells occurring whenever the overall input (which corresponds to cell membrane potential) exceeds some threshold value t. The latter function is usually taken as a step threshold function (e.g. in perceptrons [2]), a sigmoid function (either logistic function or hyperbolic tangent $f(z) = th(z) = (\exp(z) - exp(-z))/(exp(z) + exp(-z))$) in most of the modern applications, and a linear rectifier function in recent deep-learning studies.

In multilayer feed-forward ANNs, all units are organized into several layers; the units in each (i + 1)*th* layer receiving signals only from the *i*th layer. So, information flow proceeds in one direction from the first (input) layer, via one or several intermediate (hidden) layers, to the final (output) layer (see Figure 1). Multilayer feed-forward ANNs essentially generate models that consist of linear combinations of nonlinear kernel functions, so can be considered as universal mapping devices capable of approximating any continuous function given sufficient data. When these types of ANNs are used to predict properties of chemical compounds for drug discovery, units in the input layer accept the molecular descriptors, signals propagate via the nonlinear transfer functions in the hidden layers to the output layer, which predicts the corresponding property values. It has been shown mathematically that the relationship between any chemical property on its structure can be approximated using a multilayer feed-forward ANN and fragment descriptors [17,18]. When ANNs are applied to drug discovery, the modeled properties are often physicochemical and ADMET properties of organic compounds; toxicity end points; binding constants; or IC_{50} values with respect to various macromolecular biological targets, types, and profiles of biological activity, etc. (e.g. see comprehensive tables in the review article [13]).

To make correct predictions, an ANN must be trained using experimentally measured properties of a set of compounds. In training the model, the backpropagation ANN modifies the weights w so as to minimize the difference between predicted and experimental property values. Such coefficients are usually modified iteratively using the partial derivatives of the average prediction error with respect to the weights. Such derivatives can be efficiently computed by propagating errors in the opposite direction, from the output to the input layer, using the chain differentiation rule [3]. Once computed, they can be used to modify weights by taking a small step in the direction opposite to the gradient vector or conjugated to it, as in the 'delta-rule' algorithm [3,19]. Several more elaborate algorithms, such as resilient propagation [20] and the

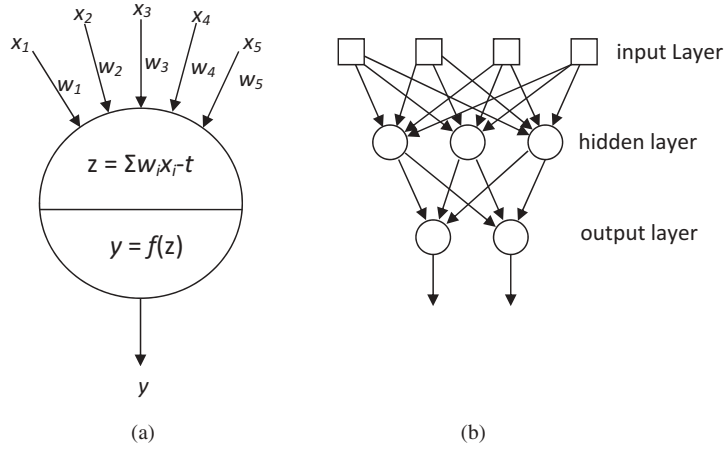

Figure 1. (a) McCulloth-Pitts' model of neurons; (b) multi-layer feed-forward ANN. Input data are propagated from the input layer to the output one. Input units are shown as squares in order not to confuse them with hidden and output units in which actual computation takes place.

Levenberg–Marquardt algorithm [21], have been shown to accelerate training. Nonetheless, for very large ANNs, the 'delta-rule' algorithm is still commonly used to train backpropagation neural networks.

In constructing QSAR models for drug discovery using multilayer ANNs, overtraining can occur [22]. This causes the ANNs to learn to predict the properties of the training set very well while failing to make useful predictions for compounds not used in training. Overfitting can be avoided by the use of a validation set of compounds, also not used in training, that monitor the predictive performance of the neural network model and stop training when it starts to deteriorate [14,22]. A third set of compounds called a test set, not used in training or validation, is required to give an unbiased assessment of the predictive performance of the neural network model. The overtraining problem can also be tackled using various regularization techniques, such as L1-, L2-, and max-norm [23] regularization with weights decay; Bayesian regularization [24–26]; or the dropout technique [27] suggested recently for deep learning (see below). In these cases, it is not necessary to use data in a validation set (or even, theoretically, in a test set).

Several other useful methods for applying backpropagation ANNs to drug discovery have been proposed. One concerns the use of ANNs with several output units corresponding to closely related properties (e.g. anticancer activity) to build QSAR models for all of them [28], the so-called multitask learning concept [29]. One study has demonstrated that the simultaneous prediction of 11 types of tissue–air partition coefficients using a single ANN with 11 output units is much more accurate in comparison with predictions made by 11 separate ANNs with a single output unit [28] due to the inductive transfer between the data concerning related end points. This opens up the possibility of building usefully predictive QSAR models for small data sets (e.g. for human end points) whenever more abundant data on closely related data (e.g. for rat end points) are available.

Another methodology useful for drug discovery concerns the concept of learned symmetry [30]. For example, if molecules form a congeneric set with common symmetrical skeleton with equivalent attachment points, then models should predict the same activity for molecules with equivalent substitution patterns. Such models were built by applying ANNs to training sets expanded by adding copies of molecules with equivalent substitution patterns. Their improved performance was demonstrated for 1,4-dihydropyridine calcium channel blockers of type and for hallucinogenic phenylalkylamines [30].

Neural networks can sometimes be used to interpret QSAR models. Analysis of neural network weights can be used to identify the most significant descriptors contributing to the model [31]. The distribution of partial derivatives of ANN outputs with respect to inputs was proposed as an index of descriptor relevance in another study [32]. Such analyses allow not allow accurate property predictions and model interpretations as do traditional statistical methods, but also revealing information on the nonlinearity of QSAR relationships, important for drug discovery.

Another method that is particularly useful for drug discovery is the autoencoder backpropagation ANN that employs a small hidden layer to reproduce input signals on the output units. If such ANN is trained on a set of compounds belonging to the same class, then by computing the reconstruction error (i.e. the difference between the values of the input and output units) for any test compound, one can detect whether it belongs to the same class. Hence, autoencoder ANNs solve the one-class classification (novelty detection) problem [33]. A virtual screening system based on autoencoder ANNs with molecular fingerprints as descriptors was developed and tested on a series of the inhibitors of glycogen synthase kinase [34]. It outperformed alternative approaches based on pharmacophore hypotheses and molecular docking in a retrospective study.

3. Bayesian-regularized neural networks

As described in the previous section, neural networks with a single hidden layer are 'universal approximators,' able to model any continuous function to arbitrary accuracy given sufficient training data. Feed-forward neural networks, like all other types of regression, can suffer from overtraining, overfitting, confusion about the optimal architecture for the network, becoming trapped in poor local optima on complex response surfaces, and inherent instability. Instability is common in regression because, as Tikhonov first stated, 'regression is an ill-posed problem in statistics' [35]. Instability is manifest by models becoming very sensitive to small changes in some model parameters and general lack of training robustness. Ill-posed problems can be converted into well-posed problems by regularization, a process where the complexity of a model is balanced against its ability to reproduce the training data faithfully.

The idea is conceptually simple and a balance is found between the ability of the model to fit the training data and the complexity of the model. Regression aims to minimize the cost function:

$$\min_f \sum_{i=1}^{n} V(f(\hat{x}_i), \hat{y}_i) + \lambda R(f),$$

where the λ parameter alters the balance between bias (model is too simple to capture any underlying relationships between, for example, molecular structure and drug activity) and variance (where the model is too complex and fits the data underlying relationship and the noise in the data [solid curve in Figure 2]).

Bayesian methods can be used to automatically find the optimal value of the regularization constant(s) (□ in the above example). The theory is relatively complex and has been described fully in prior publications [24]. The bottom line is that Bayesian regularization generates neural network models with few, if any, of the problems of unregularized backpropagation or feed-forward neural networks. Applying related methods that use a sparse Bayesian prior can generate very good quantitative structure–activity relationship models for pharmaceutically relevant properties that are robust, sparse, and often interpretable. These methods achieve excellent feature selection, an important issue for developing models that are optimally predictive and easier to understand in terms of

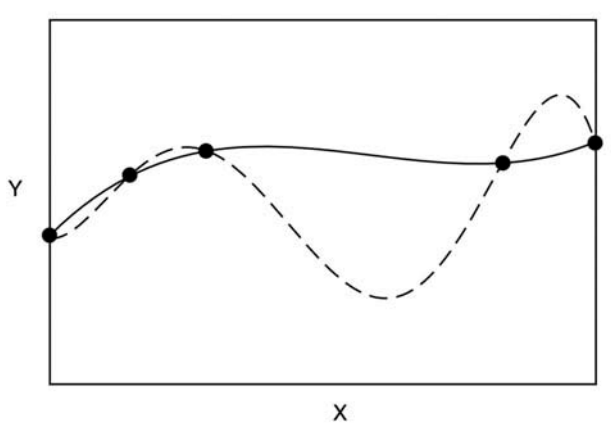

Figure 2. The solid and sashed functions both incur zero loss on the given data points. Regularization will induce a model to prefer the solid function, which may generalize better to other data points sampled from the underlying unknown distribution.

the underlying structure–activity relationships [36,37]. Bayesian-regularized neural networks generate robust models with optimal complexity, avoiding under- or overfitting, and also are relatively insensitive to the number of nodes used in the hidden layer.

Bayesian-regularized neural networks have been applied to a relatively wide variety of molecular design and property prediction problems [38]. The seminal works by Mackay [39], Bishop [40], and Figueiredo [41] laid out the theory of neural networks and Bayesian regularization. Burden and Winkler first applied them to QSAR and later to Quantitate Structure-Property Relationship (QSPR) problems. Researchers from AstraZeneca have employed Bayesian-regularized neural networks to model important physicochemical properties of drugs such as aqueous solubility [42] and log D the pH-dependent distribution of drugs between lipophilic and aqueous phases [43]. They have been used to successfully and quantitatively model acute toxicity of chemicals to *Tetrahymena pyriformis* [44], have been employed to predict the binding of peptide epitopes to MHC class II [45], the activities of inhalation anesthetics [46], etc. Recent studies using the Kaggle benchmark data sets have shown that Bayesian neural networks perform on average as well as the new deep-learning methods [47].

Bayesian-regularized neural networks have been used to develop a very large and widely applicable model of aqueous solubility of small organic molecules [48]. Modeling of drug activities has been an important application of Bayesian neural networks. Orre et al. used these modeling methods to find adverse drug combinations [49], Winkler and Burden used these methods to make quantitative predictions of drug partitioning through the blood–brain barrier [50], and Polley et al. reported robust and predictive models of intestinal absorption of drugs [51] and the potency and selectivity of farnesyl transferase inhibitors used for cancer therapy [52]. Caballero et al. added genetic selection to a Bayesian neural network (Bayesian-Regularized Genetic Neural Networks [BRGNN]) to model the selective inhibition of the calcium-activated potassium channel by clotrimazole analogs [53]. Fernandez et al. used BRGNN methods to model a diverse range of biological and physicochemical properties of small molecules [54]. They also modeled cyclin-dependent kinase inhibition by 1H-pyrazolo[3,4-d]pyrimidine derivatives using Bayesian-regularized neural network ensembles [53] and subsequently applied BRGNN techniques to create quantitative QSAR models for several drug–target interaction data sets.

Bayesian-regularized neural networks have also been used to make seminal contributions to the prediction of possible adverse biological effects of nanomaterials [55,56] and to the design of cell-targeting nanoparticles for personalized medicine [57]. In a related vein, Bayesian-regularized neural networks have been used to predict the very complex mesophases that occur in amphiphilic drug delivery systems [56], a computational problem that is essentially intractable by methods such as molecular dynamics simulations.

4. Ensemble and consensus models

The error of each machine-learning method involves two major factors: bias and variance. As unregularized neural networks are 'ill-posed' methods, small perturbations in data or descriptors may result in large changes in the predicted values [58,59]. Thus, models developed using the same or similar data set but with different initialization of neural network weights can provide different predictions for new data. The ensemble average, calculated over multiple predictors, can decrease the variance and is another way to improve the model generalization compared to that of individual networks. Clearly, the more similar the individual networks, the smaller the advantage of ensemble averages. To increase the variance of individual networks, differences in their training data sets could be maximized. However, it is essential that the training sets will contain the same information as the given data set. This can be achieved with the so-called bagging approach [59], which creates new training data sets by sampling with replacement from the initial set. Another way of increasing variance is to use different descriptors for each model. This can be achieved by subsampling descriptors from the initial set (this is used in bagging) or by using different sets of descriptors. Models developed by averaging models derived from different sets of descriptors are frequently called consensus models. Ensembles have been used in chemistry and drug discovery since the 1990s [22,58], while consensus models have become more popular since the 2000s [60,61]. Ensemble and consensus methods were recently used successfully for prediction of diverse properties, such as inhibitors of CYP450 [62], analysis of non-nucleoside HIV reverse-transcriptase inhibitors [63], potential endocrine disruptors [64], and others. Ensembles of neural network models frequently provided higher prediction accuracy compared to other methods in these studies.

Ensemble or consensus model prediction variation can be used to estimate the applicability domain of models. The basic hypothesis is that predictions of individual models will diverge for data points far from the training set points. Thus, high variance of prediction (STandard Deviation [STD]) can be used to detect molecules for which predictions are less reliable.

Benchmarking of different definitions of the applicability domain identified STD as the best measure of prediction reliability of molecules from regression studies [65]. A similar measure was also one the best ones for classification studies [66].

The training of individual neural network or their ensembles is a rather time-consuming problem and can be impractical if new data become available and models must be repeatedly retrained. The Associative Neural Network (ASNN) method based on a model of thalamocortical organization of the brain addresses this problem [67]. Each neural network in the ASNN ensemble can be considered a representation of one cortex column in the brain. The predicted values of each model, ordered by magnitude, can be considered a spatiotemporal representation of the training set by the ASNN. Thus, training samples are stored in the 'memory of the ASNN' as spatiotemporal patterns, together with predicted and real values. For each new sample, the ASNN retrieves the most similar stored patterns and uses prediction errors of these patterns to correct the prediction of the new data point. This 'local correction' efficiently increases prediction accuracy of the ensemble by decreasing the bias of the ensemble method. Moreover, the new patterns can be easily added to the 'memory' of the ASNN without a need to retrain the whole network, thus allowing the neural network to instantaneously learn new data. This feature tunes the global models to a local subset of data. For example, the ALOGPS 2.1 program was initially developed to predict octanol/water partition coefficients using organic molecules only [68]. The addition of a small training set of Pt complexes with measured logP values allowed this program to successfully predict Pt complexes in a blind test set [69]. It is interesting that the accuracy of this model was higher than models developed with Pt complexes. In a similar way, the logP algorithm was tuned to predict logD by providing *in house* data measured in pharma companies [70].

The high prediction power of the algorithm was demonstrated in several studies, where the ASNN-based models provided one of the highest prediction accuracies for prediction of physicochemical properties [71–73] as well as contributed the top ranked models in recent challenges organized by US EPA ToxCast and NIH Tox21 programs [74,75].

5. Self-organizing maps and related approaches

Kohonen's Self-Organizing Maps (SOM) is a biology-inspired topology-preserving nonlinear dimensionality reduction method that can map molecules from multidimensional descriptor space onto a 2D grid of neurons [5]. In this case, each molecule activates a single 'winner' neuron with the closest distance between its code vector and the molecule in descriptor space. The training algorithm of SOM guarantees that close molecules activate topologically close neurons in the competitive layer. Projection of molecules to the location of the corresponding winning neurons produces a map, in which neighborhood relations between molecules are mostly preserved. As structurally similar molecules tend to have similar activities, then molecules belonging to the same activity class are mapped either to the same neuron or to several neighboring neurons. The neurons can be colored according to the activity class of molecules mostly mapped to them. Such colored layer of neurons can be used for predicting activities of new molecules projected onto it and hence for conducting virtual screening. This mapping procedure underlies the use of SOM for drug discovery [76].

Not only individual molecules, but also local atom or bond descriptors, molecular fields, and mixture components can be mapped to neurons in the competitive layer to produce novel descriptors useful for drug discovery. 3D-QSAR methods CoMSA [77] and volume learning algorithm (VLA) [78] are based on mapping molecular fields. Recent publication on classification of mixtures of Chinese herbal medicines based on SOM is an example of this approach [79]. Quantitative predictions can be performed by hybrid ANNs containing SOM as the input layer for multilayer ANN. The classical example for this are the counter-propagation ANN, while the most recent example – the network for 'deep learning' of chemical data, in which the SOM layer of neurons is followed by layers of backpropagation ANN [80]. The latter network was used for predicting antibacterial activity of peptides.

Modifications of the generative topographic mapping, a probabilistic analog of SOM based on Bayesian learning, have recently been used in the field of drug discovery for visualizing chemical space [81], building activity landscapes [82], classification [83] and regression [82] QSAR models, comparing chemical libraries [81], predicting activity profiles [84], and performing inverse-QSAR studies [84].

6. Other types of neural networks

There are several dozens of other general-purpose types of neural networks, some of which have been used in structure–activity modeling and drug discovery [13,14]. They include Cascade-Correlation network with dynamically growing number of neurons; Radial Basis Functions Neural Network along with the Probabilistic Neural Network; and General Regression Neural Network closely related to it, a family of ANNs based on adaptive resonance theory (ART-1, ART-2, ARTMAP, etc.). One should also mention specialized ANNs designed to work directly with molecular graphs without the use of a precomputed set of molecular descriptors: a 'neural device for searching direct correlations between structures and properties of chemical compounds' with convolution architecture (see later) [85], recursive neural networks [86], graph machines [87], etc. Despite some success stories, currently these types of NNs are however rarely used for drug discovery.

A recurrent neural network (RNN) is a class of ANN where connections between units form a directed cycle. This creates a type of neural network that can model dynamic temporal behavior. Unlike feed-forward neural networks, RNNs can use their internal memory to process arbitrary sequences of inputs. They are particularly suitable for predicting time-varying parameters, although Bayesian neural networks and other have also been shown to do this successfully [88]. The theory of RNNs and their application to unsupervised pattern recognition have been described by Orre et al. [89]. This type of neural network has not been used often for drug discovery or modeling of related medical activities or properties. Goh et al. first

applied RNNs to predicting drug dissolution profiles, and important problem in the pharmaceutical industry [90]. More recently, Bonet and coworkers used RNNs to predict HIV drug resistance [91].

7. Deep learning

The term 'deep learning' refers to training multilayer ANNs with more than one of hidden layer and a large number (thousands) of hidden layer nodes (see major publications [8,9]). Before the advent of first 'deep' ANNs in the middle of 2000s, almost all standard machine-learning methods could be considered as 'shallow': they could formally be described by means of at most two layers of processing units [8]. Although multilayer ANNs with any number of hidden layers could formally be constructed, their training using backpropagation-based optimization algorithms usually fails whenever the number of hidden layers exceeds three or four [8]. This can be explained by the increased risk of overfitting with larger numbers of weights. It is also due to another peculiarity of the backpropagation algorithm, in which the values of error derivatives, which are propagated from the output layer back to the input one, vanish rapidly with the distance from the output layer. This is due to the multiplication of several small partial derivatives as required by the chain differentiation rule. As a result, only a couple of layers closest to the output one can actually be trained, whereas all weight parameters in the remaining hidden layers stay almost unchanged during the training. Since all adjustable weights of multilayer ANNs are usually initialized with small random numbers, during the training, the network tries to approximate the 'functional dependence' of the output values on the random numbers formed on the hidden units near the input layer and, not surprisingly, fails.

An efficient solution to this problem was found in 2006 by Hinton and Salakhutdinov [7] who suggested splitting the learning process into two stages: (1) representation learning [92] and (2) training the network using the learned representation. In the first successful implementation of this methodology, a cascade of Restricted Boltzmann Machines (RBMs) was used to learn a hierarchy of internal data representations [7]. Then, the weight parameters learned by RBMs were used to initialize the weights of the deep multilayer ANNs that were subsequently readjusted during the training using the standard backpropagation algorithm. In this way, multilayer ANNs with virtually any number of hidden layers can be trained efficiently.

After this pioneering study, the methodology of deep learning was augmented in several important ways. First, the sigmoidal transfer function was replaced by the linear rectifier function, usually producing stronger models [93]. Second, a new, powerful regularization technique, weight dropout [27], was introduced. To implement weight dropout nodes is randomly switched off during the training. The regularizing effect of the dropout technique in conjunction with the use of a rectifier transfer function means that it becomes possible to train very large ANNs with a huge number of hidden layer nodes and their interconnections without overtraining or overfitting [9]. Furthermore, it appears that with sufficiently big large data sets, it is not necessary to pretrain ANNs using cascades of RBMs or other autoencoders to learn data representation and the weights except between the final hidden and the output layers can just be set randomly once.

Another important technique that was successfully integrated with deep learning is convolutional architecture [94]. Convolutional ANNs have roots in the neocognitron [4] architecture specially designed to mimic information processing in visual cortex. Distinct from the standard multiple-layer ANNs working with fixed-size data vectors, convolutional ANNs are designed to work with data in the form of multiple arrays with variable size, such as 2D pixel matrices for images, while providing necessary invariance to irrelevant data transformations, such as shifts or distortions of images. Convolutional ANNs consist of two kinds of layers: convolutional layers and pooling layers. Each unit in a convolutional layer takes signals from a small patch of units from the previous layer through a set of weights shared by all units in the layer. Each unit in a pooling layer computes the maximum of signals coming from a patch of units in the previous layer. Stacks of several convolution and pooling units allow extraction of complex relevant features from images. In deep ANNs, convolution and pooling layers are typically placed at the input side of the network.

An important factor in the recent success of ANNs with deep learning is the use of fast graphics processing units (GPU) that significantly accelerate the training due to parallelization. Currently, a deep-learning ANN composed of millions of units with hundreds of millions adjustable weights organized in several dozen layers can be trained with huge data sets of hundreds of millions examples. Such networks have already achieved human or higher performance in solving tasks such as image and speech recognition.

Deep learning is not just a new term to designate the state-of-the-art in the domain of ANNs. It cannot be reduced to a simple application of additional techniques, such as dropout and rectifier units, or simple augmentation of the number of hidden layers in multilayer ANNs. Neither it cannot be reduced to a mere application of deep-learning software to solve old problems using traditional approaches. Deep learning is a new philosophy of predictive modeling. The success of the application of standard 'shallow' machine-learning methods is greatly influenced by how well the features representing data have been chosen using experience and domain knowledge. With very well-designed features, even the simplest linear or nearest-neighbors machine-learning methods can be applied to build predictive models. The great promise of the deep learning is to be able to extract necessary features with required invariance properties automatically from raw data via representation learning [9,92]. Deep-learning ANNs form multiple levels of representation in their hidden layers, with each subsequent layer forming representation of a higher, more complex and abstract, level than the previous one. With multiple (up to several dozen) hidden layers of nonlinear units, such ANN can learn extremely complex functions of its inputs with all necessary invariance properties, that is not always possible using standard machine-learning methods and manually tailored features. Due to the process of representation learning,

deep learning can easily profit from related data sets with multiple labels via multitask and transfer learning [29,95], as well as from data without labels via semi-supervised and transductive learning [96]. So, deep learning can be considered as an important step towards what is called artificial intelligence [8]. However, on the negative side, they cannot as easily perform sparse feature selection, important for optimizing predictions of new data and for simple interpretation of models. Methods such as Multiple Linear Regression with Expectation Maximization [37] can achieve efficient sparse feature selection so can be complementary to deep-learning methods. Additionally, on the positive side, although they perform as well on average as state-of-the-art shallow neural network methods like Bayesian-regularized neural networks, they may be faster to train and large cluster or GPU hardware, handle large data sets, and may be easier to code algorithmically.

Representation learning provided by deep multilayer ANNs will play an increasingly important role in computational drug discovery [97–99]. However, the question of molecular descriptors used to capture the important properties of molecules is still a relatively poorly answered one. Despite the large number and variety of molecular descriptors, none can be guaranteed to have universal applicability and provide optimal solutions to all problems arising in drug discovery. Deep-learning ANNs may alleviate this issue somewhat by generating novel and useful complex representations that may be more suited to solving specific tasks in this domain, albeit at the expense of generating models whose interpretation is even more difficult. However, the discovery of more suitable and chemically interpretable molecular descriptors is still an important, poorly solved problem in QSAR. One can also expect that the ability to integrate a large amount of related data using deep multilayer ANNs with multiple outputs will be very useful for drug discovery as it allows reuse of previously accumulated data and knowledge to meet new challenges in drug discovery.

Although first publications on the use of deep learning in the field of drug discovery appeared very recently [100–102], some of the key ideas underlying the concept of deep learning have already been used for building QSAR models. In 1997, the first multilayer ANN with convolutional layers containing shared weights ('receptors') and pooling layers ('collectors'), capable of extracting molecular features from raw data, was reported [85]. Like deep learning, convolutional ANNs were inspired by the neocognitron [4] architecture for image recognition. The analysis of pixels in images was replaced by analysis of atoms and bonds in molecules. The resulting 'neural device for searching direct correlations between structures and properties of organic compounds' allowed construction of QSAR models using raw molecular data without preliminary computation of molecular descriptors [85]. Another idea applied to QSAR modeling and discussed above is the use of ANNs with several outputs to predict several properties using the multitask learning framework [28].

Public attention was drawn to the application of deep learning to drug discovery in 2012 after publication in *The New York Times* of the results of a Kaggle competition sponsored by Merck [103]. The competition was won by a deep-learning ANN with a 15% improvement in accuracy over Merck's standard method. In 2014, an arcXiv article [104] written by the winning team showed that multitask (multiple outputs) deep ANNs outperformed alternative methods. Subsequently, in a more comprehensive study published [100], it was demonstrated that ANNs with several hidden layers largely provided better prospective predictions than Random Forests on a set of large, diverse QSAR data sets taken from Merck's drug discovery efforts. They also showed that the dropout regularization techniques and rectifier transfer function significantly improved prediction performance of QSAR models. For best deep ANN performance, they concluded that the ANNs should be not only deep but also wide, i.e., contain a lot of units in each of the layers. This contradicts the traditional belief that ANNs should contain as few as possible adjustable parameters in order to avoid overfitting. It was also demonstrated that a clear advantage of using multitask ANNs with several output units over the use of single-task ANNs with a single output unit for each property is most pronounced for relatively small data sets, whereas with large data sets, the effect can be even opposite. Surprisingly, pretraining deep ANNs using stacks of RBM models was shown to deteriorate predictive performance of QSAR models in this study.

Two massive, multitask ANNs for drug discovery have recently been reported [101,105]. One of them was trained on a data set of nearly 40 million protein-ligand measurements across 259 biological targets [101]. Another was trained on 2 million data points for 1280 biological targets [105]. In both cases, it has been shown that massively multitask ANNs trained with deep learning significantly outperform single-task methods, and their predictive performance improves as additional tasks (targets) and data points are added. This improvement is significantly influenced by both the amount of data and the number of tasks (targets). It has also been demonstrated for toxicity prediction that, by combining reactive centers, such networks can learn complex internal representation that resemble well-established toxicophores [106].

8. Conclusions

In this review, we analyzed recent developments in the application of neural networks to drug discovery: building QSAR models to predict activity profiles and drug–target interactions, binding constants with respect to various targets, drug selectivity, inhibition constants for different enzymes, toxicity profiles, ADMET and physicochemical properties, characteristics of drug-delivery systems, etc., as well as performing virtual screening. The more traditional approaches, such as 'shallow' neural networks, Bayesian, and ensemble/consensus learning, were shown to be very important tools in drug discovery. We have shown that these methods are widely used in the contemporary research and very often generate the most valuable models. Moreover, these methods allow interpretation of QSAR models and identification of the most important molecular features. Ensemble and consensus modeling may provide additional advantages by decreasing the variance of individual models as well as improving the estimation of the applicability domain of models. Neural networks are becoming

even more prominent due to recent progress in deep-learning technology. Training of millions of neurons with millions of data points that was not feasible a few years ago can now be accomplished. Deep-learning technology provides interesting and powerful complementary capability to drug discovery using neural network models, and we expect to see a rapid growth in applications in the nearest future.

9. Expert opinion

The neural networks are very important tools in drug discovery. While they initially suffered from overfitting and overtraining and incorrect model validation, these problems have now been essentially overcome. Methods such as early stopping [22], bias correction as used in Associative Neural Neworks [74,75], Bayesian regularization [24–26], and training with dropout techniques [27] allow development of highly predictive robust models. Hence, application of the traditional neural networks to drug design and increasingly other fields such as materials has matured. Neural networks are sometimes criticized as a black-box approach. However, this is as much due to use of poorly interpretable descriptors as a problem with the neural network method. There are increasingly sophisticated methods for analyzing the significance of neural network weights [31], or general purpose methods such as predicted Matched Molecular Pairs [107] allows more facile interpretation of models. Additionally, neural network models can be interpreted by analysis of the distribution of partial derivatives of ANN outputs with respect to inputs or calculation their sensitivities, as discussed above [31,32].

Neural networks, in particular those using deep-learning technology, will continue to be used actively in drug discovery in the future. They will be particularly useful for analysis of large data sets that are increasingly generated by automated high-throughput technologies so are well suited to the challenges of Big Data [108]. We expect that neural networks will be increasingly used for other challenging tasks such as force field parameterization, optimization of drug delivery systems, ADMET prediction and drug classification, prediction of synthesis difficulty, and especially for multitask learning and simultaneously prediction of multiple biological activities or properties. We also expect that learning by combining supervised and unsupervised data, learning of highly imbalanced data sets, learning of data weighted by measurement accuracies, etc., will become more commonplace. In particular, we expect that the advantages of deep-learning networks in analysis of large and complex data for which traditional statistical machine-learning methods sometimes fail will be fully exploited.

ANNs are also finding applications in augmenting expensive quantum chemistry calculations, accurate prediction of protein structures, simulation of small molecule–protein as well as protein–protein interactions, simulations of PK/PD parameters, and prediction of *in vivo* toxicity. It is feasible that future algorithm-based system biology and machine-learning approaches will be merged in a single application. For example, system biology approaches where differential equations that simulate the cells require a lot of adjustable parameters, some of which are very difficult to measure, may benefit from this fusion. Such parameters can be estimated using neural networks and be coupled with simulation outputs to identify the most likely biological system states.

However, one should not overstate the potential of deep-learning technology over traditional QSAR/QSPR for analysis of small data sets with a limited number of descriptors. The gain in the performance can come from using a big amount of previously unused related data. The gain may also arise from the ability of deep learning to create new, complex molecular descriptors through representation learning [92]. We also expect that the ability of deep learning to create multiple levels of data representations with different complexity could provide fundamentally new ways of analyzing structure–activity relationships and solving the problems of great importance for drug discovery, such as the problem of activity cliffs [109]. Indeed, very rugged and bumpy activity landscapes with numerous activity cliffs with respect to input descriptors or low-level representations might appear to be very smooth and simple with respect to high-level representations being formed in deep-learning systems, which is the essence of representation learning. The ability of metric learning, a kind of linear representation learning, to eliminate activity cliffs in activity landscapes has recently been demonstrated [110]. One can expect that nonlinear representation learning provided by neural networks should give an even greater effect.

Until recently, multilayer backpropagation neural networks (shallow or deep) and self-organizing maps formed two separate branches of development, perhaps, with an exception of VLA [78], which clustered input descriptors using SOM for neural network learning. At the present time, however, there is a clear trend towards their convergence [80]. Incorporation of SOM-like layers into deep-learning systems might endow the latter with the means of data mapping, and visualization proved to be useful for drug discovery.

Declaration of interest

IV Tetko is the CEO of BIGCHEM GmbH, which licenses OCHEM and ASNN software. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

ORCID

Igor V. Tetko http://orcid.org/0000-0002-6855-0012

References

Papers of special note have been highlighted as:
• of interest
•• of considerable interest

1. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5:115–133. doi:10.1007/BF02478259.
2. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958;65(6):386–408.

3. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by back-propagating errors. Nature. 1986;33:533–536. doi:10.1038/323533a0.
•• **The main reference for backpropagation neural networks.**
4. Fukushima K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybernetics. 1980;36:193–202. doi:10.1007/BF00344251.
5. Kohonen T. Self-organizing maps. Springer; 2001.
•• **The main reference on self-organizing maps.**
6. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci U S A. 1982;79 (8):2554–2558.
7. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504–507. doi:10.1126/science.1127647.
•• **First publication on deep learning.**
8. Bengio Y. Learning deep architectures for AI. Foundations Trends Machine Learning. 2009;2(1):1–127. doi:10.1561/2200000006.
9. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521 (7553):436–444. doi:10.1038/nature14539.
•• **Important recent review on deep learning.**
10. Hiller SA, Golender VE, Rosenblit AB, et al. Cybernetic methods of drug design. I. Statement of the problem – the perceptron approach. Comput Biomed Res. 1973;6(5):411–421.
• **First publication on the use of neural networks in drug discovery.**
11. Aoyama T, Suzuki Y, Ichikawa H. Neural networks applied to structure-activity relationships. J Med Chem. 1990;33(3):905–908.
12. Winkler DA, Burden FR. Application of neural networks to large dataset QSAR, virtual screening, and library design. Methods Mol Biol. 2002;201:325–367. doi:10.1385/1-59259-285-6:325.
13. Halberstam NM, Baskin II, Palyulin VA, et al. Neural networks as a method for elucidating structure-property relationships for organic compounds. Russian Chem Rev. 2003;72(7):629–649. doi:10.1070/RC2003v072n07ABEH000754.
14. Baskin II, Palyulin VA, Zefirov NS. Neural networks in building QSAR models. Methods Mol Biol (Clifton, NJ). 2008;458:137–158.
• **Review concerning the use of neural network in QSAR modeling.**
15. Maltarollo VG, Abf DS, Honório KM. Applications of artificial neural networks in chemical problems. In: Suzuki K, editor. Artificial neural networks-architectures and applications. INTECH Open Access Publisher; 2013. p. 203–223.
16. Dearden JC, Rowe PH. Use of artificial neural networks in the QSAR prediction of physicochemical properties and toxicities for REACH legislation. Methods Mol Biol (Clifton, NJ). 2015;1260:65–88.
17. Baskin II, Skvortsova MI, Stankevich IV, et al. On the basis of invariants of labeled molecular graphs. J Chem Inf Comput Sci. 1995;35(3):527–531.
18. Artemenko NV, Baskin II, Palyulin VA, et al. Artificial neural network and fragmental approach in prediction of physicochemical properties of organic compounds. Russian Chem Bull. 2003;52(1):20–29. doi:10.1023/A:1022467508832.
19. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representation by error propagation. In: Rumelhart DE, McClelland JL, editors. Parallel distributed processing: explorations in the microstructure of cognition, volume 1: foundations. Cambridge (MA): MIT Press; 1986. p. 318–362.
20. Riedmiller M, Braun H. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. Proc IEEE Int Conf Neural Networks. 1993;586–591.
21. Hagan MT, Menhaj M. Training feedforward networks with the Marquardt algorithm. IEEE Trans Neural Networks. 1994;5(6):989–993. doi:10.1109/72.329697.
22. Tetko IV, Livingstone DJ, Luik AI. Neural network studies. 1. Comparison of overfitting and overtraining. J Chem Inf Comput Sci. 1995;35(5):826–833.
•• **Method to avoid overfitting of neural networks, notwithstanding of their complexity, by early stopping.**
23. Srebro N, Shraibman A. Rank, trace-norm and max-norm. In: Learning theory. Springer; 2005. p. 545–560.
24. Burden F, Winkler D. Bayesian regularization of neural networks. Methods Mol Biol. 2008;458:25–44.
25. Burden FR, Winkler DA. Robust QSAR models using Bayesian regularized neural networks. J Med Chem. 1999;42(16):3183–3187. doi:10.1021/jm980697n.
•• **First application of Bayesian regularized neural networks in QSAR modeling.**
26. Burden FR, Winkler DA. An optimal self-pruning neural network and nonlinear descriptor selection in QSAR. QSAR Comb Sci. 2009;28(10):1092–1097. doi:10.1002/qsar.v28:10.
27. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. J Machine Learn Res. 2014;15(1):1929–1958.
• **Important method to avoid overfitting used in deep-learning neural network.**
28. Varnek A, Gaudin C, Marcou G, et al. Inductive transfer of knowledge: application of multi-task learning and feature net approaches to model tissue-air partition coefficients. J Chem Inf Model. 2009;49 (1):133–144. doi:10.1021/ci8002914.
29. Caruana R. Multitask learning. Mach Learn. 1997;28(1):41–75. doi:10.1023/A:1007379606734.
30. Baskin II, Halberstam NM, Mukhina TV, et al. The learned symmetry concept in revealing quantitative structure-activity relationships with artificial neural networks. SAR QSAR Environ Res. 2001;12 (4):401–416. doi:10.1080/10629360108033247.
31. Tetko IV, Villa AE, Livingstone DJ. Neural network studies. 2. Variable selection. J Chem Inf Comput Sci. 1996;36(4):794–803.
32. Baskin II, Ait AO, Halberstam NM, et al. An approach to the interpretation of backpropagation neural network models in QSAR studies. SAR QSAR Environ Res. 2002;13(1):35–41. doi:10.1080/10629360290002073.
33. Markou M, Singh S. Novelty detection: a review – part 2: neural network based approaches. Signal Process. 2003;83(12):2499–2521. doi:10.1016/j.sigpro.2003.07.019.
34. Karpov PV, Osolodkin DI, Baskin II, et al. One-class classification as a novel method of ligand-based virtual screening: the case of glycogen synthase kinase 3OI inhibitors. Bioorg Med Chem Lett. 2011;21 (22):6728–6731. doi:10.1016/j.bmcl.2011.09.051.
35. Tikhonov AN, Arsenin VY. Solutions of ill-posed problems. New York: Winston; 1977.
36. Burden FR, Ford MG, Whitley DC, et al. Use of automatic relevance determination in QSAR studies using Bayesian neural networks. J Chem Inf Comput Sci. 2000;40(6):1423–1430.
37. Burden FR, Winkler DA. Optimal sparse descriptor selection for QSAR using Bayesian methods. Qsar Comb Sci. 2009;28(6–7):645–653. doi:10.1002/qsar.v28:6/7.
• **First description of the use of feature selection using sparse Bayesian priors.**
38. Winkler DA, Burden FR. Bayesian neural nets for modeling in drug discovery. Drug Discovery Today BIOSILICO. 2004;2(3):104–111. doi:10.1016/S1741-8364(04)02393-5.
39. Mackay DJC. A practical Bayesian framework for backpropagation networks. Neural Comput. 1992;4(3):448–472. doi:10.1162/neco.1992.4.3.448.
40. Bishop CM. Neural networks for pattern recognition. Oxford (UK): Oxford University Press; 1995.
41. Figueiredo MAT. Adaptive sparseness for supervised learning. Ieee T Pattern Anal. 2003;25(9):1150–1159. doi:10.1109/TPAMI.2003.1227989.
• **Describes the mathematics involved in Bayesian regularization for regression.**
42. Bruneau P. Search for predictive generic model of aqueous solubility using Bayesian neural nets. J Chem Inf Comput Sci. 2001;41 (6):1605–1616.
43. Bruneau P, McElroy NR. logD(7.4) modeling using Bayesian regularized neural networks. Assessment and correction of the errors of prediction. J Chem Inf Model. 2006;46(3):1379–1387. doi:10.1021/ci0504014.

44. Burden FR, Winkler DA. A quantitative structure-activity relationships model for the acute toxicity of substituted benzenes to *Tetrahymena pyriformis* using Bayesian-regularized neural networks. Chem Res Toxicol. 2000;13(6):436–440.
45. Burden FR, Winkler DA. Predictive Bayesian neural network models of MHC class II peptide binding. J Mol Graph Model. 2005;23(6):481–489. doi:10.1016/j.jmgm.2005.03.001.
46. Manallack DT, Burden FR, Winkler DA. Modelling inhalational anaesthetics using Bayesian feature selection and QSAR modelling methods. Chemmedchem. 2010;5(8):1318–1323. doi:10.1002/cmdc.201000056.
47. Ma J, Sheridan RP, Liaw A, et al. Deep neural nets as a method for quantitative structure-activity relationships. J Chem Inf Model. 2015;55(2):263–274. doi:10.1021/ci500747n.
48. Salahinejad M, Le TC, Winkler DA. Aqueous solubility prediction: do crystal lattice interactions help? Mol Pharmaceut. 2013;10(7):2757–2766. doi:10.1021/mp4001958.
49. Orre R, Bate A, Lindquist M. Bayesian neural networks used to find adverse drug combinations and drug related syndromes. Persp Neural Comp. 2000;215–220.
50. Winkler DA, Burden FR. Modelling blood-brain barrier partitioning using Bayesian neural nets. J Mol Graph Model. 2004;22(6):499–505. doi:10.1016/j.jmgm.2004.03.010.
51. Polley MJ, Burden FR, Winkler DA. Predictive human intestinal absorption QSAR models using Bayesian regularized neural networks. Aust J Chem. 2005;58(12):859–863. doi:10.1071/CH05202.
52. Polley MJ, Winkler DA, Burden FR. Broad-based quantitative structure-activity relationship modeling of potency and selectivity of farnesyltransferase inhibitors using a Bayesian regularized neural network. J Med Chem. 2004;47(25):6230–6238. doi:10.1021/jm049621j.
53. Caballero J, Garriga M, Fernandez M. Genetic neural network modeling of the selective inhibition of the intermediate-conductance Ca^{2+}-activated K+ channel by some triarylmethanes using topological charge indexes descriptors. J Comput Aid Mol Des. 2005;19(11):771–789. doi:10.1007/s10822-005-9025-z.
54. Fernandez M, Caballero J, Fernandez L, et al. Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM). Mol Divers. 2011;15(1):269–289. doi:10.1007/s11030-010-9234-9.
55. Epa VC, Burden FR, Tassa C, et al. Modeling biological activities of nanoparticles. Nano Lett. 2012;12(11):5808–5812. doi:10.1021/nl303144k.
56. Winkler DA, Mombelli E, Pietroiusti A, et al. Applying quantitative structure-activity relationship approaches to nanotoxicology: current status and future potential. Toxicology. 2013;313(1):15–23. doi:10.1016/j.tox.2012.11.005.
57. Le TC, Yan B, Winkler DA. Robust prediction of personalized cell recognition from a cancer population by a dual targeting nanoparticle library. Adv Funct Mater. 2015;25(44):6927–6935. doi:10.1002/adfm.201502811.
58. Tetko IV, Luik AI, Poda GI. Applications of neural networks in structure-activity relationships of a small number of molecules. J Med Chem. 1993;36(7):811–814.
    - **First use of neural network ensembles in drug discovery.**
59. Breiman L. Random forests. Machine Learn. 2001;45(1):5–32. doi:10.1023/A:1010933404324.
60. Votano JR, Parham M, Hall LH, et al. Three new consensus QSAR models for the prediction of Ames genotoxicity. Mutagenesis. 2004;19(5):365–377. doi:10.1093/mutage/geh043.
61. Tomal JH, Welch WJ, Zamar RH. Exploiting multiple descriptor sets in QSAR studies. J Chem Inf Model. 2016;56(3):501–509. doi:10.1021/acs.jcim.5b00663.
62. Novotarskyi S, Sushko I, Korner R, et al. A comparison of different QSAR approaches to modeling CYP450 1A2 inhibition. J Chem Inf Model. 2011;51(6):1271–1280. doi:10.1021/ci200091h.
63. Nizami B, Tetko IV, Koorbanally NA, et al. QSAR models and scaffold-based analysis of non-nucleoside HIV RT inhibitors. Chemometrics Intell Lab Syst. 2015;148:134–144. doi:10.1016/j.chemolab.2015.09.011.
64. Mansouri K, Abdelaziz A, Rybacka A, et al. CERAPP: collaborative estrogen receptor activity prediction project. Environ Health Perspect. 2016.
65. Tetko IV, Sushko I, Pandey AK, et al. Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection. J Chem Inf Model. 2008;48(9):1733–1746. doi:10.1021/ci800151m.
66. Sushko I, Novotarskyi S, Körner R, et al. Applicability domain for *in silico* models to achieve accuracy of experimental measurements. J Chemom. 2010;24(3–4):202–208. doi:10.1002/cem.1296.
67. Villa AE, Tetko IV, Dutoit P, et al. Corticofugal modulation of functional connectivity within the auditory thalamus of rat, guinea pig and cat revealed by cooling deactivation. J Neurosci Methods. 1999;86(2):161–178.
68. Tetko IV, Tanchuk VY. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. J Chem Inf Comput Sci. 2002;42(5):1136–1145.
69. Tetko IV, Jaroszewicz I, Platts JA, et al. Calculation of lipophilicity for Pt(II) complexes: experimental comparison of several methods. J Inorg Biochem. 2008;102(7):1424–1437. doi:10.1016/j.jinorgbio.2007.12.029.
70. Tetko IV, Poda GI. Application of ALOGPS 2.1 to predict log D distribution coefficient for Pfizer proprietary compounds. J Med Chem. 2004;47(23):5601–5604. doi:10.1021/jm049509l.
71. Tetko IV, Poda GI, Ostermann C, et al. Large-scale evaluation of log P predictors: local corrections may compensate insufficient accuracy and need of experimentally testing every other compound. Chem Biodivers. 2009;6(11):1837–1844. doi:10.1002/cbdv.200900075.
72. Tetko IV, Sushko Y, Novotarskyi S, et al. How accurately can we predict the melting points of drug-like compounds? J Chem Inf Model. 2014;54(12):3320–3329. doi:10.1021/ci5005288.
73. Tetko IV, Varbanov HP, Galanski M, et al. Prediction of logP for Pt(II) and Pt(IV) complexes: Comparison of statistical and quantum-chemistry based approaches. J Inorg Biochem. 2016;156:1–13. doi:10.1016/j.jinorgbio.2015.12.006.
74. Novotarskyi S, Abdelaziz A, Sushko Y, et al. ToxCast EPA in vitro to in vivo challenge: insight into the Rank-i model. Chem Res Toxicol. 2016;29(5):768–775. doi:10.1021/acs.chemrestox.5b00481.
75. Abdelaziz A, Spahn-Langguth H, Werner-Schramm K, et al. Consensus modeling for HTS assays using in silico descriptors calculates the best balanced accuracy in Tox21 challenge. Frontiers Environ Sci. 2016;4(2). doi:10.3389/fenvs.2016.00002.
76. Anzali S, Gasteiger J, Holzgrabe U, et al. The use of self-organizing neural networks in drug design. Perspect Drug Discov Des. 1998;9:273–299. doi:10.1023/A:1027276425268.
77. Polanski J, Gieleciak R, Bak A. The comparative molecular surface analysis (COMSA) – a nongrid 3D QSAR method by a coupled neural network and PLS system: predicting pKa values of benzoic and alkanoic acids. J Chem Inf Comput Sci. 2002;42(2):184–191.
78. Tetko IV, Kovalishyn VV, Livingstone DJ. Volume learning algorithm artificial neural networks for 3D QSAR studies. J Med Chem. 2001;44(15):2411–2420.
79. Wang M, Li L, Yu C, et al. Classification of mixtures of Chinese herbal medicines based on a Self-Organizing Map (SOM). Mol Inform. 2016. doi:10.1002/minf.201500115.
80. Schneider P, Mueller AT, Gabernet G, et al. Hybrid network model for "deep learning" of chemical data: application to antimicrobial peptides. Mol Inform. 2016. doi:10.1002/minf.201600011.
81. Gaspar HA, Baskin II, Marcou G, et al. Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge. J Chem Inf Model. 2015;55(1):84–94. doi:10.1021/ci500575y.
82. Gaspar HA, Baskin II, Marcou G, et al. GTM-based QSAR models and their applicability domains. Mol Inform. 2015;34(6–7):348–356. doi:10.1002/minf.201400153.
83. Gaspar HA, Marcou G, Horvath D, et al. Generative topographic mapping-based classification models and their applicability domain: application to the Biopharmaceutics Drug Disposition

Classification System (BDDCS). J Chem Inf Model. 2013;53 (12):3318–3325. doi:10.1021/ci400423c.
84. Gaspar HA, Baskin II, Marcou G, et al. Stargate GTM: bridging descriptor and activity spaces. J Chem Inf Model. 2015;55 (11):2403–2410. doi:10.1021/acs.jcim.5b00398.
85. Baskin II, Palyulin VA, Zefirov NS. A neural device for searching direct correlations between structures and properties of chemical compounds. J Chem Inf Comput Sci. 1997;37(4):715–721. doi:10.1021/ci940128y.
• **First neural network with discrete convolutional architecture for searching direct correlations between structures and properties of chemical compounds.**
86. Micheli A, Sperduti A, Starita A, et al. Analysis of the internal representations developed by neural networks for structures applied to quantitative structure-activity relationship studies of benzodiazepines. J Chem Inf Comput Sci. 2001;41(1):202–218.
87. Goulon A, Picot T, Duprat A, et al. Predicting activities without computing descriptors: graph machines for QSAR. SAR QSAR Environ Res. 2007;18(1–2):141–153. doi:10.1080/10629360601054313.
88. Le TC, Conn CE, Fr B, et al. Computational modeling and prediction of the complex time-dependent phase behavior of lyotropic liquid crystals under in meso crystallization conditions. Cryst Growth Des. 2013;13(3):1267–1276. doi:10.1021/cg301730z.
89. Orre R, Bate A, Noren GN, et al. A Bayesian recurrent neural network for unsupervised pattern recognition in large incomplete data sets. Int J Neural Syst. 2005;15(3):207–222. doi:10.1142/S0129065705000219.
90. Goh WY, Lim CP, Peh KK, et al. Application of a recurrent neural network to prediction of drug dissolution profiles. Neural Comput Appl. 2002;10(4):311–317. doi:10.1007/s005210200003.
91. Bonet I, Garcia MM, Saeys Y, et al. Predicting human immunodeficiency virus (HIV) drug resistance using recurrent neural networks. In: Mira J, Alvarez JR, editors. Bio-inspired modeling of cognitive tasks, pt 1, proceedings. 2007. p. 234–243.
92. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35 (8):1798–1828. doi:10.1109/TPAMI.2013.50.
• **Important review on representation learning.**
93. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. Int Conf Artif Intelligence Stat. 2011;2011:315–323.
94. LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–2324. doi:10.1109/5.726791.
95. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–1359. doi:10.1109/TKDE.2009.191.
96. Chapelle O, Schoelkopf B, Zien A. Semi-supervised learning. Cambridge (MA): The MIT Press; 2006.
97. Varnek A, Baskin II. Chemoinformatics as a theoretical chemistry discipline. Mol Inform. 2011;30(1):20–32. doi:10.1002/minf.v30.1.
• **An important review on the use of machine learning in structure–activity modeling**
98. Varnek A, Baskin I. Machine learning methods for property prediction in chemoinformatics: quo vadis? J Chem Inf Model. 2012;52 (6):1413–1437. doi:10.1021/ci200409x.
99. Cherkasov A, Muratov EN, Fourches D, et al. QSAR modeling: where have you been? Where are you going to? J Med Chem. 2015;57 (12):4977–5010. doi:10.1021/jm4004285.
•• **Comprehensive review on QSAR modeling**
100. Ma J, Sheridan RP, Liaw A, et al. Deep neural nets as a method for quantitative structure – activity relationships. J Chem Inf Model. 2015;55(2):263–274. doi:10.1021/ci500747n.
101. Ramsundar B, Kearnes S, Riley P, et al. Massively multitask networks for drug discovery. arXiv preprint arXiv:150202072. 2015.
102. Gawehn E, Hiss JA, Schneider G. Deep learning in drug discovery. Mol Inf. 2016;35(1):3–14. doi:10.1002/minf.v35.1.
103. Markoff J. Scientists see promise in deep-learning programs. The New York Times. 2012 Nov 23.
104. Dahl GE, Jaitly N, Salakhutdinov R. Multi-task neural networks for QSAR predictions. arXiv preprint arXiv:14061231. 2014.
105. Unterthiner T, Mayr A, Ünter Klambauer G, et al. Deep learning as an opportunity in virtual screening. In: Deep Learning and Representation Learning Workshop, NIPS. 2014.
106. Unterthiner T, Mayr A, Klambauer G, et al. Toxicity prediction using deep learning. arXiv preprint arXiv:150301445. 2015.
107. Sushko Y, Novotarskyi S, Korner R, et al. Prediction-driven matched molecular pairs to interpret QSARs and aid the molecular optimization process. J Cheminform. 2014;6(1):48. doi:10.1186/1758-2946-6-6.
108. Tetko IV, Engkvist O, Koch U, et al. BIGCHEM: challenges and opportunities for Big Data analysis in chemistry. Mol Inf. 2016.
•• **An important review on Big Data in chemistry.**
109. Cruz-Monteagudo M, Medina-Franco J, Pérez-Castillo Y, et al. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? Drug Discov Today. 2014;19(8):1069–1080. doi:10.1016/j.drudis.2014.02.003.
110. Kireeva NV, Ovchinnikova SI, Kuznetsov SL, et al. Impact of distance-based metric learning on classification and visualization model performance and structure-activity landscapes. J Comput Aid Mol Des. 2014;28(2):61–73. doi:10.1007/s10822-014-9719-1.