ELSEVIER

The rise of deep learning in drug discovery

Hongming Chen[1], Ola Engkvist[1], Yinhai Wang[2], Marcus Olivecrona[1] and Thomas Blaschke[1]

[1] Hit Discovery, Discovery Sciences, Innovative Medicines and Early Development Biotech Unit, AstraZeneca R&D Gothenburg, Mölndal 43183, Sweden
[2] Quantitative Biology, Discovery Sciences, Innovative Medicines and Early Development Biotech Unit, AstraZeneca, Unit 310, Cambridge Science Park, Milton Road, Cambridge CB4 0WG, UK

Over the past decade, deep learning has achieved remarkable success in various artificial intelligence research areas. Evolved from the previous research on artificial neural networks, this technology has shown superior performance to other machine learning algorithms in areas such as image and voice recognition, natural language processing, among others. The first wave of applications of deep learning in pharmaceutical research has emerged in recent years, and its utility has gone beyond bioactivity predictions and has shown promise in addressing diverse problems in drug discovery. Examples will be discussed covering bioactivity prediction, *de novo* molecular design, synthesis prediction and biological image analysis.

Introduction

Digital data, in all shapes and sizes, is growing exponentially. According to the National Security Agency of the USA, the Internet is processing 1826 petabytes of data per day [1]. In 2011, digital information grew nine times in volume in just five years [2]; and by 2020 its amount in the world is expected to reach 35 trillion gigabytes [3]. The high demand of exploring and analyzing big data has encouraged the use of data-hungry machine learning algorithms like deep learning (DL). DL has gained huge success in a wide range of applications such as computer games, speech recognition, computer vision, natural language processing, self-driving cars, among others [4]. It is fair to say that DL is changing our everyday life. In the Gartner-selected top ten technology trends of 2018, DL-represented AI technologies were ranked at the top position [5].

Over the past decade, there has been a remarkable increase in the amount of available compound activity and biomedical data [6,7] owing to the emergence of new experimental techniques such as HTS, parallel synthesis, among others [7,8]. How to efficiently mine the large-scale chemistry data becomes a crucial problem for drug discovery. Larger data volumes in combination with increased automation technology have promoted further use of machine learning. Besides established methods like support vector machines (SVM) [9], neural networks (NN) [10] and random forest (RF) [11], which have been utilized to develop QSAR models for a long time, methods like matrix factorization [12] and DL have started to be used. DL has taken advantage of the increased amounts of data and the continuous increase of available computer power. A difference between most other machine learning methods and DL is the flexibility of the NN architecture in DL. Architectures that will be discussed in this review are convolutional neural networks (CNNs), recurrent neural networks (RNNs) and fully connected feed-forward networks. Single-layer NNs have been used in QSAR modeling for a long time [10]; and with increasing data size and computational power have made it natural to apply multilayer feed-forward networks for bioactivity predictions. A somewhat surprising development has been the use of RNNs in *de novo* design which could not be foreseen a few years ago. With the adoption of high-throughput imaging equipment, CNNs have gained remarkable success in computer vision and have become a natural choice for biological image processing. The field of applying DL in drug discovery is rapidly progressing

Corresponding author: Chen, H. (hongming.chen@astrazeneca.com)

1359-6446/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
https://doi.org/10.1016/j.drudis.2018.01.039

Please cite this article in press as: Chen, H. et al. The rise of deep learning in drug discovery, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2018.01.039

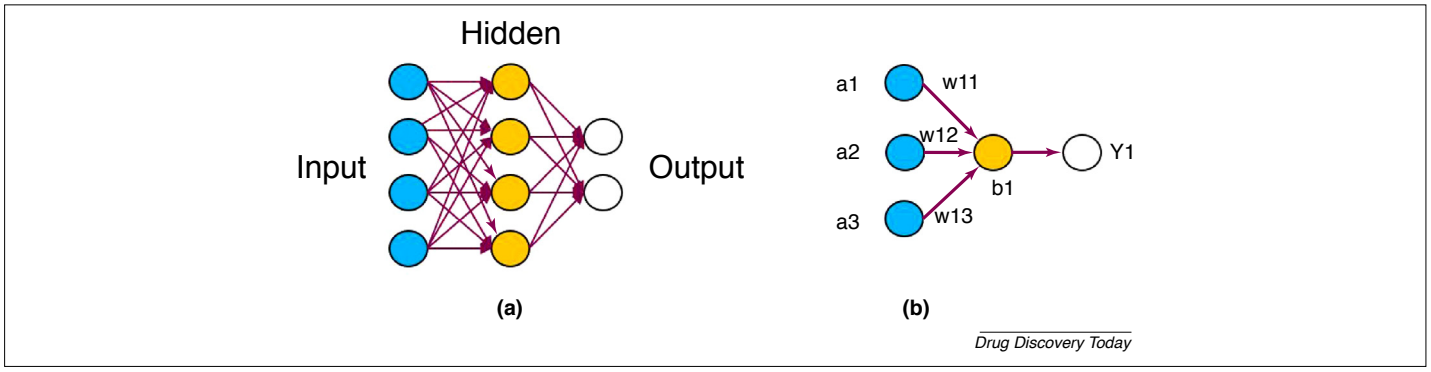

FIGURE 1

A simple illustration of neural networks (NNs). (a) A NN is composed of input, hidden and output layers. (b) The output values of a hidden unit are calculated from input values via an activation function.

with new articles published almost every week. Recently, several reviews on DL applications in computational chemistry and life sciences have been published [13–18]. Here, we focus more on DL applications in drug discovery particularly in the chemoinformatics and biological image analysis domains and highlight DL architectures used so far within drug discovery.

Principles of deep learning

DL is a class of machine learning algorithms that uses artificial neural networks (ANNs) with many layers of nonlinear processing units for learning data representations. The earliest ANN can be traced back to 1943 [19], when Warren McCulloch and Walter Pitts developed a computational model for NNs based on mathematics and algorithms called threshold logic. The basic structure of a modern ANN is represented in Fig. 1 and is inspired by the structure of the human brain. There are three basic layers in an ANN: the input layer, hidden layer and output layer. Depending on the type of ANN, the nodes, also called neurons, in neighboring layers are either fully connected or partially connected. Input variables are taken by input nodes and the variables are transformed through hidden nodes, and in the end output values are calculated at output nodes. The interrelationship between input and output values of a hidden unit can be exemplified in Fig. 1b. The output value Y_i of the node i is calculated as shown in Eq. (1).

$$Y_i = g\left(\sum_j \mathrm{W_{ij}} * a_j\right) \quad (1)$$

where a_j refers to the input variables, W_{ij} is the weight of input node j on node i and function g is the activation function, which is normally a nonlinear function (e.g., sigmoid or Gaussian function) to transform the linear combination of input signal from input nodes to an output value. The training of an ANN is done by iterative modification of the weight values in the network to optimize the errors between predicted and true value typically through the back-propagation methods [20]. The modern ANN algorithm was developed during the 1960s to the 1980s and applications have appeared since then. But the traditional ANN method suffered from problems such as overfitting, diminishing gradients, among others, and was largely replaced by other machine learning algorithms like SVM [9] and RF [11]. The recent development of DL has given ANN a renaissance. The major difference between DL and traditional ANN is the scale and complexity of the NNs. DL uses larger numbers of hidden layers whereas traditional ANNs normally can only afford one or two hidden layers owing to the limitation of computer hardware in the early days. DL can afford to use many more nodes in each layer owing to the appearance of more-powerful CPU and GPU hardware. There are also many algorithmic improvements in DL, for example using the dropout [21] and DropConnect [22] methods to address the overfitting problem, applying rectified linear unit (ReLU) [23] to avoid vanishing gradients and introducing convolutional and pooling layers as novel network architectures to enable the usage of large numbers of input variables. Most of the DL software packages are open-sourced. TensorFlow [24], Caffe [25], PyTorch [26], Keras [27] and Theano [28] are among the most popular DL packages used in the data science community. Here, we briefly introduce several popular NN architectures used in DL (Fig. 2). First is the fully connected deep neural network (DNN) which contains multiple hidden layers and each layer comprises hundreds of nonlinear process units (Fig. 2a). DNNs can take large numbers of input features and the neurons in different layers of a DNN can automatically extract features at different hierarchical levels [29].

Another very popular NN is CNN, which is widely used for image recognition (Fig. 2b). It usually contains several convolution layers and subsampling layers. The convolution layer consists of a set of filters (or kernels) that have a small receptive field and learnable parameters. During the forward pass, each filter is convoluted across the width and height of the input volume, computing the dot product between the entries of the filter and its receptive field in input volume and producing a 2D feature map of that filter. The subsampling layer is used to reduce the size of feature maps. In the end, the feature maps are concatenated into fully connected layers where neurons in neighboring layers are all connected just like in a traditional ANN to give a final output value. Owing to sharing the same parameters for each filter, a CNN largely reduces the number of free parameters learned, thus lowering the consumed memory and increasing the learning speed. It has outperformed other types of machine learning algorithms in image recognition [30].

One additional variant of an ANN (Fig. 2c) is RNN. Unlike feed-forward NNs, it allows the connection among neurons in the same hidden layer to form a directed cycle. RNNs can take sequential

Please cite this article in press as: Chen, H. et al. The rise of deep learning in drug discovery, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2018.01.039

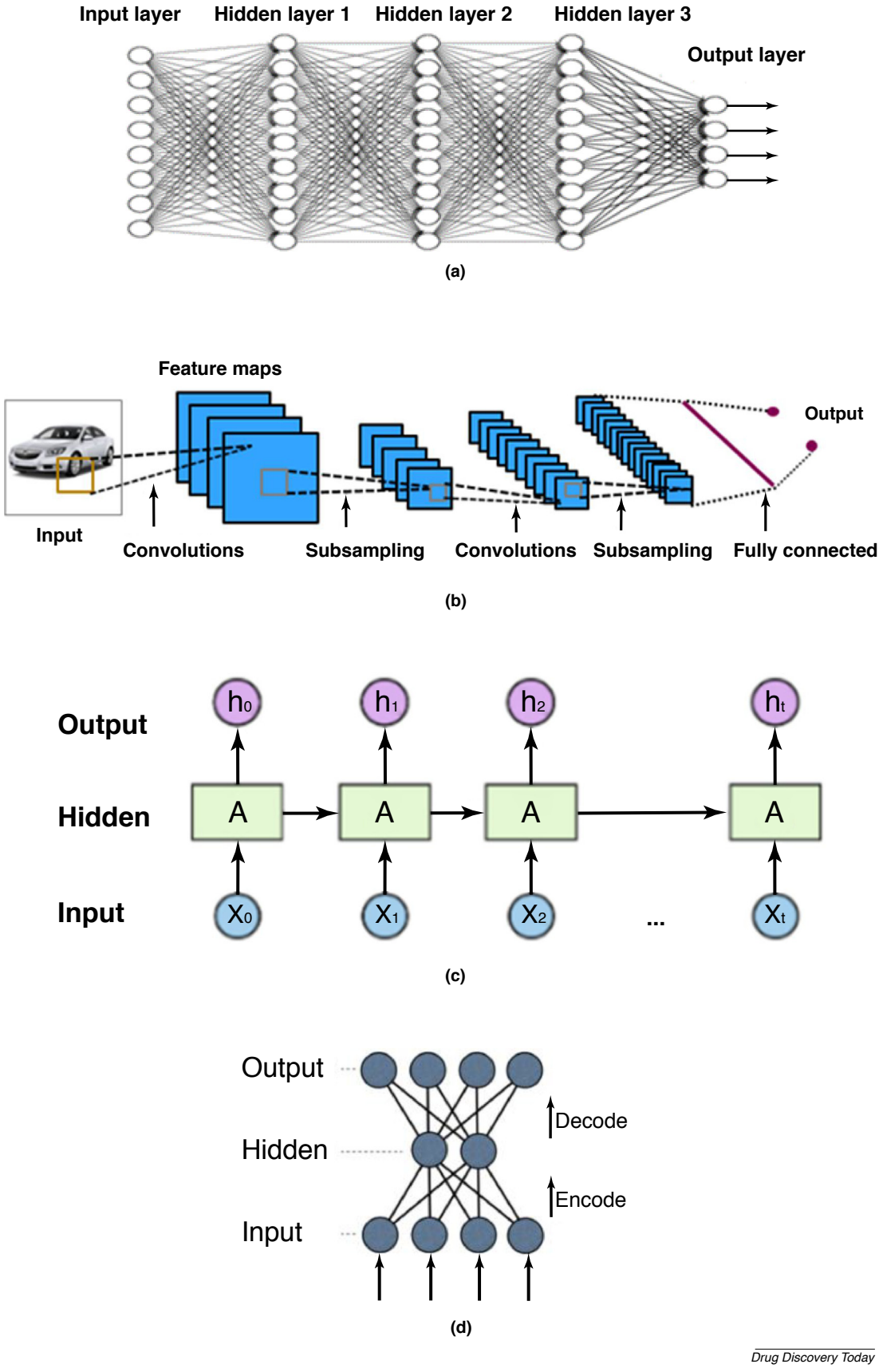

FIGURE 2

Architecture of several popular neural networks: (a) fully connected deep neural network (DNN), (b) convolutional neural network (CNN), (c) recurrent neural network (RNN) and (d) autoencoder (AE).

Please cite this article in press as: Chen, H. et al. The rise of deep learning in drug discovery, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2018.01.039

data as input features, which is very suitable for time-dependent tasks like language modeling [31]. Using a technology called long short term memory (LSTM) [32], RNNs can reduce the vanishing gradient problem.

The fourth ANN architecture shown in Fig. 2d is called auto-encoder (AE) [33]. An AE is a NN used for unsupervised learning. It contains an encoder part, which is a NN to transform the information received from the input layer into a limited number of hidden units, and then couples a decoder NN with the output layer having the same number of nodes as the input layer. Instead of predicting labels of input instances, the purpose of the decoder NN is to reconstruct its own inputs from a fewer number of hidden units. Typically, the purpose of AE is for nonlinear dimensionality reduction. Recently, the AE concept has become more widely used for learning generative models from data [34]. Below, we illustrate how these DL technologies are applied in drug discovery research.

Application of deep learning in compound property and activity prediction

Machine learning methods including ANN have been applied in compound activity prediction for a long time. Naturally, DL methods are adopted to address the activity prediction problems in the first place. When compounds are presented by the same number of molecular descriptors, the straight forward method is to use fully connected DNNs to build models. Dahl *et al*. [35] applied a DNN on the Merck Kaggle challenge dataset using a large number of 2D topological descriptors; and the DNN showed slightly better performance in 13 of the total 15 targets than the standard RF method. Some of the key learnings from the study are: (i) DNNs can handle thousands of descriptors without the need of feature selection; (ii) dropout can avoid the notorious overfitting problem faced by a traditional ANN; (iii) hyper-parameter (number of layers, number of nodes per layer, type of activation functions, etc.) optimization can maximize the DNN performance; (iv) multitask DNN models perform better than single-task models. Mayr *et al*. [36] reported their multitask DNN models that won the Tox21 challenge on a dataset comprising 12 000 compounds for 12 high-throughput toxicity assays. Similar to Dahl's architecture [35,37], dropout and ReLU activation function were used in the DNN, and model training was run in parallel on GPU machines. They used a large feature set with static descriptors (3D, 2D descriptors, pre-defined toxicophores) as well as dynamically generated extended connectivity fingerprint descriptors (ECFP) to enable DNN to make self-feature deduction during training. More interestingly, statistical association analysis was done for DNN models exclusively using ECFP, and substructures significantly associated with known toxicophores in each hidden layer can be identified. These benchmark results demonstrate the advantages of a multitask DNN compared with a single-task DNN and conventional machine learning methods.

Recently, some other benchmark studies were published to further support the conclusion. Ramsundar *et al*. carried out a systematic study [38] to build multitask DNNs and compare their performance with single-task DNN models. Their results show that multitask models constantly perform better than single-task and RF models. Koutsoukas *et al*. [39] compared a DNN model with some commonly used machine learning methods such as SVM, RF, among others, on seven datasets selected from ChEMBL [40]. DNNs were found to statistically outperform (with P value <0.01 based on the Wilcoxon's statistical test) other machine learning methods. Lenselink *et al*. [41] reported another benchmark study for comparing DNN with conventional machine learning methods RF, SVM, naive Bayesian and logic regression methods taking protein descriptors into account [i.e., the proteochemometric (PCM) study]. They investigated performance of various classification models on a dataset comprising 314 767 target–compound interactions. The DNN model turned out to be the best model in terms of BEDROC (Boltzmann-enhanced discrimination of receiver operating characteristic), and multitask and PCM implementations were shown to improve performance over single-task DNNs.

Besides the benchmark studies of DNN, Subramanian *et al*. [42] reported a study using DNN with 2D topological descriptors to build a predictive BACE activity model and achieved a classification accuracy of 0.82 and standard error of pIC_{50} ~0.53 on the validation set. Aliper *et al*. [43] built DNN models for predicting pharmacological properties of drugs and for drug repurposing leveraging transcriptomic data from the LINCS project [44], as well as the pathway information. It has been shown that, using pathway and gene-level information, DNN models achieved high accuracy in predicting drug indications, hence they could be useful for drug repurposing.

Efforts have also been made in using representation learning (i.e., enabling NNs to learn directly from the molecular structure instead of using predefined molecular descriptors). This idea was first explored by Merkwirth *et al*. in 2005 [45]. Several years later, two different methods were developed to address the problem. Lusci *et al*. [46] reported a method that employed a variant of RNN, called UGRNN, which first transforms molecular structures into vectors of the same length as the molecular representation and then passes them to a fully connected NN layer to build models. Bit values in the vectors are learned from the dataset. The UGRNN method was shown to be able to build predictive solubility models that were comparable in accuracy to models built with molecular descriptors. Xu *et al*. [47] applied the same method to model drug-induced liver injury (DILI). The DL models were built based on 475 drugs and validated on an external dataset of 198 drugs. The best model achieved an AUC of 0.955 exceeding the accuracy of previously reported DILI models.

Another type of method is called graph convolution models. The basic idea is similar to the UGRNN method, which employs NNs to automatically generate a molecular description vector and vector values are learned by training NNs. Inspired by the Morgan circular fingerprint method [48], Duvenaud *et al*. [49] proposed the neural fingerprint method as one of the first efforts in creating a graph convolution model. The workflow of this method can be seen in Fig. 3. First, the 2D molecular structure is read to form a state matrix, containing atom and bond information (based on the bonds attached to the atom) for each atom. The state matrix then goes through a convolution operation via a single-layer NN to generate a fixed length vector as the molecular representation. The convolution operation can be run at different levels by considering the contribution of neighboring atoms, which is equivalent to the circular fingerprints at different neighboring levels. The vectors generated from different convolution operations first go through a softmax transformation and then are summed up to

Please cite this article in press as: Chen, H. et al. The rise of deep learning in drug discovery, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2018.01.039

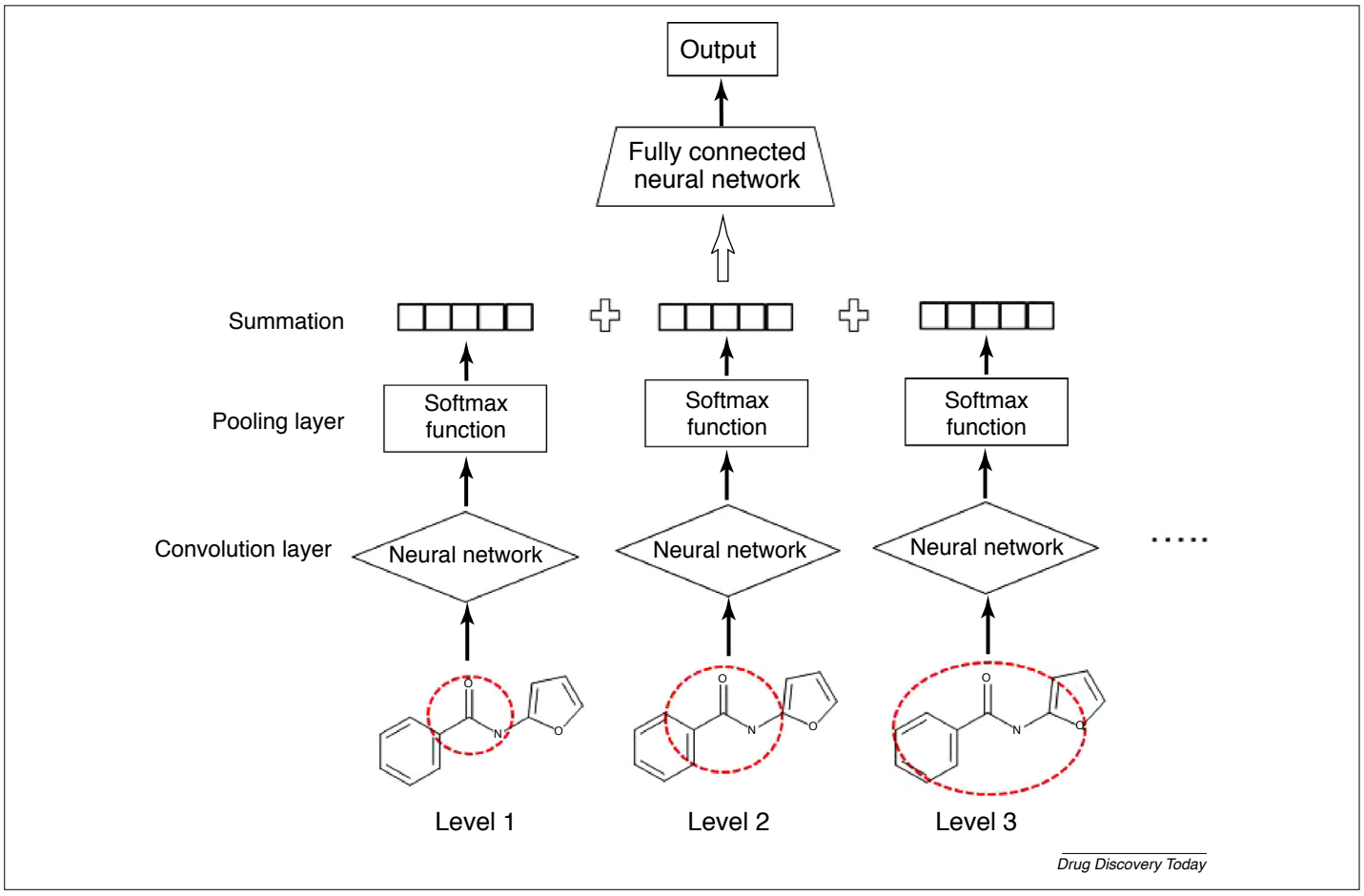

FIGURE 3

Illustration of graph convolutional neural networks (CNNs) [49]. A molecular graph first goes through a convolution operation via a single layer NN to form a vector of fixed length. The convolution operation can be run at different neighbor levels. The vectors generated from different convolution operations then go through a softmax transformation and are summed up to form the neural fingerprints of the compound. The neural fingerprint is passed through another fully connected NN layer to generate the final output. The bit values in the neural fingerprint are learned through training and are differentiable.

form the final vector for the compound, which is a neural fingerprint encoding molecular level information. The neural fingerprints are passed through another fully connected NN layer to generate the final output. The bit values in the neural fingerprint are learned through training and are differentiable. In Duvenaud's three test cases, better results were obtained using neural fingerprints than with Morgan fingerprints and, more importantly, the influential substructures in the graph convolution model can be visualized to interpret the model. The advantage of the graph convolution model is that descriptors are generated automatically during the training and do not need any predefined molecular descriptor. Such a descriptor is not a general descriptor, but is task-specific and fully differentiable and hence can potentially provide better prediction. Other molecular graph convolution methods were reported by Kearnes *et al.* [50], Xu *et al.* [51], Li *et al.* [52] and Coley *et al.* [53] to extend on Duvenaud's method. Recently, researchers from Google [54] reformulated several existing graph convolution algorithms [49,50,53,55,56] into a common framework known as a message passing neural network (MPNN) and used the MPNNs to predict quantum chemical properties.

Besides the graph-based representation learning methods, DL methods based on other types of molecular representation were also explored. Bjerrum [57] used a SMILES string as the input to LSTM RNNs to build predictive models without the need to generate molecular descriptors. More interestingly, it was observed that augmenting the dataset by using multiple SMILES strings to represent the same compound achieved better results than using canonical SMILES. Goh *et al.* [58] applied a CNN on images of 2D drawings of molecules and achieved surprisingly comparable results to DNN models trained on ECFP. Moreover [59], when the images were augmented with some basic chemical information, the model performance was further improved. The capability of learning representations from structures directly without using any predefined structure descriptor is an important feature distinguishing DL from other machine learning methods and it basically makes the traditional feature selection and reduction procedures unnecessary.

De novo design through deep learning

Another interesting application of DL in chemoinformatics is the generation of new chemical structures through NNs. Gómez-Bombarelli *et al.* proposed a novel method [60] using variational autoencoder (VAE) to generate chemical structures (Fig. 4). The first step is to use VAE to do unsupervised learning to map

Please cite this article in press as: Chen, H. et al. The rise of deep learning in drug discovery, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2018.01.039

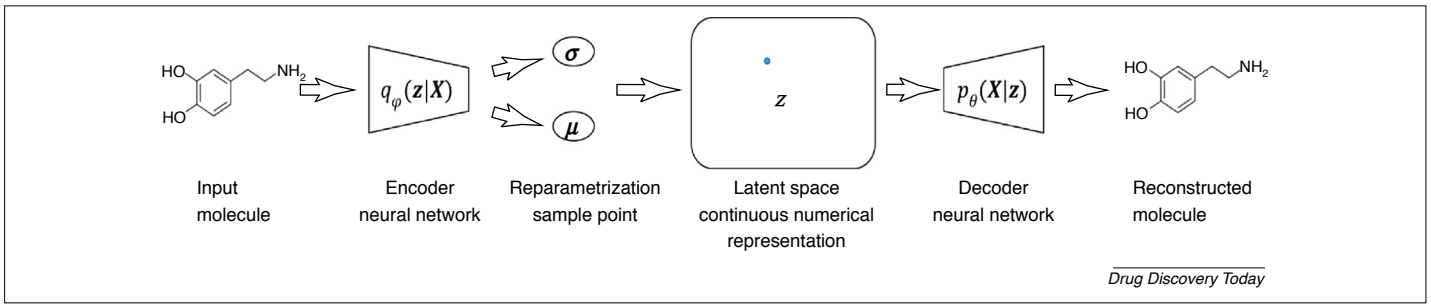

FIGURE 4

The illustration of a variational autoencoder (VAE) method. The encoder neural network (NN) converts a discrete molecule into Gaussian distribution deterministically. After the latent variables are reparameterized against the gaussian distribution with given mean and variance, a new point is sampled and fed into the decoder NN. In the generation mode, only the decoder is used to generate a new molecule from the sampled latent point.

chemical structures (SMILES strings) in the ZINC database into latent space. Once the VAE training is done, the latent vector in the latent space becomes a continuous representation of molecular structure and can be reversibly transformed to a SMILES string through the trained VAE. Generation of a new structure with desirable properties can be realized by searching optimal latent solutions in the continuous latent space via any optimization method (e.g., Bayesian optimization) and then decoding the searched latent solutions into SMILES. Following on from Gómez-Bombarelli's work, Kadurin *et al.* [61] used VAE as a molecular descriptor generator coupled with a generative adversarial network (GAN) [62], a special NN architecture, to generate new structures that were claimed to have promising specific anticancer properties. Blaschke *et al.* [63] utilized VAE to generate novel structures with predicted activity against dopamine receptor type 2.

RNNs have been very successful in the natural language processing area [31]. Segler *et al.* [64] and Yuan *et al.* [65] reported their studies using RNNs to generate novel chemical structures. After training the RNN on a large number of SMILES strings, the RNN method worked surprisingly well for generating new valid SMILES strings that were not included in the training set (Fig. 5). The RNN writes structurally valid SMILES by learning the underlying probability distribution of characters in a SMILES string and, in this case, RNN can be regarded as a generative model for molecule structures. Segler *et al.* [64] also explored the possibility of using RNNs to generate target-specific libraries by first training a general prior model and then a fine-tuned focused model through transfer learning on a small set of target-specific active compounds. In a retrospective analysis for testing on two antibioactive targets, their focused models were able to generate 18% unseen true actives for *Staphylococcus aureus* and 28% for *Plasmodium falciparum*.

Jaques *et al.* [66] applied a reinforcement learning technology, called Deep Q-learning, together with an RNN to generate SMILES with desirable molecular properties such as cLogP [67] and QED drug-likeness [68]. However, their method needed a reward function that incorporates handwritten rules to penalize undesirable types of structures, which otherwise would lead to exploitation of the reward resulting in unrealistically simple molecules. To overcome the drawback, Olivecrona *et al.* [69] proposed a policy-based reinforcement learning approach to tune the pre-trained RNNs for generating molecules with given user-defined properties. In one test example for tuning the model toward generating compounds predicted to be active against the dopamine receptor type 2, the model generated structures of which >95% were predicted to be active, including experimentally confirmed actives that have not been included in the generative model nor the activity prediction model.

The methods described above have demonstrated potentials as alternatives to the traditional rule-based approaches for *de novo* design. However, GANs and the reinforcement learning methods are known to be susceptible to mode collapse (i.e., the models only generate a single solution or a small family of similar solutions). This has been highlighted in a recent survey [70] on *de novo* structure generation using DL tools. Considerable effort [71,72] has been spent to address this issue.

Application of deep learning in predicting reactions and retrosynthetic analysis

Synthesis predictions have a long history dating back to rule-based methods in the 1960s [73]. Very recently some promising results were reported in reaction prediction using DL methods. Although there has been no explicit comparison with other machine learning methods, the results indicated that DL can achieve performance on-par with, or superior to, the rule-based methods. Schematically, two types of problems can be addressed with machine learning including DL in reaction informatics. One type is forward reaction prediction, where the products are predicted given a set of reactants, and the other type is retrosynthetic prediction, where given a final product the reaction steps that produce the product are predicted. Coley *et al.* [74] utilized NN to rank the candidate products for a set of reactions based on a training set of 15 000 reactions from US patents. The reactions were classified into templates and the trained model correctly assigned the major product rank 1 in 71.8%, rank ≤3 in 86.7% and rank ≤5 in 90.8% of cases. To overcome the coverage and efficiency issues faced with the template-based reaction prediction methods, a template-free approach was proposed [75] in a follow-up study by the same research group. They employed the Weisfeiler–Lehman difference network to score the generated candidate reactions and superior performance was achieved compared with reaction template-based methods. Segler *et al.* [76] used 3.5 million reactions as the training set for DNN. A top-ten accuracy of 97% for reaction prediction and 95% in retrosynthetic analysis were achieved. In another study [77], they combined policy networks and Monte-Carlo tree search for retrosynthetic

Please cite this article in press as: Chen, H. et al. The rise of deep learning in drug discovery, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2018.01.039

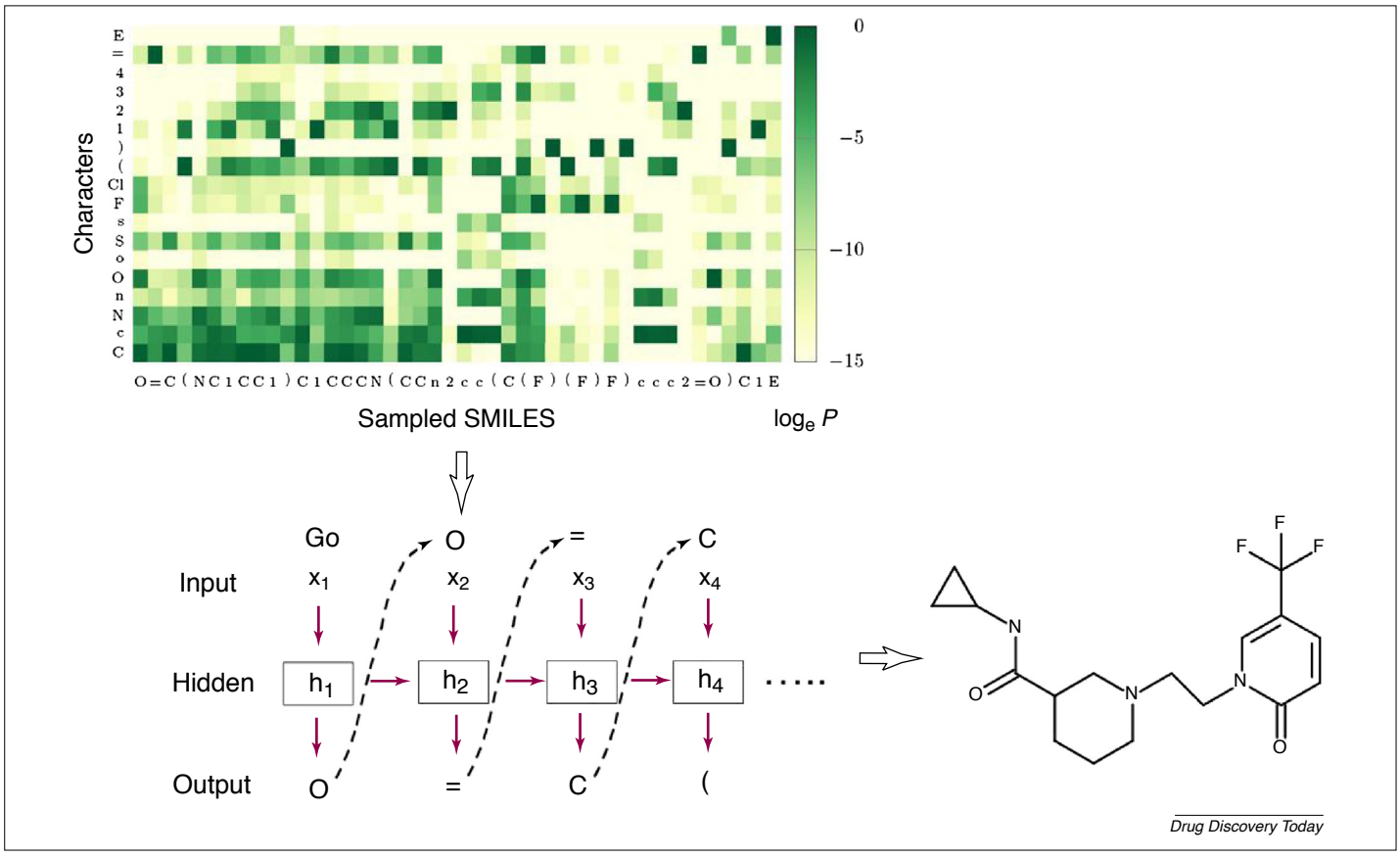

FIGURE 5

Structure generation from recurrent neural networks (RNNs). The upper plot shows how the RNN model thinks when generating the structure on the bottom right. The y axis lists all possible tokens that can be chosen at each step, the color represents the conditional probability for the character to be chosen at the current step given the previously chosen characters, and the x axis shows the character that, in this instance, was sampled. The bottom left figure demonstrates how the RNN actually works in the structure-generation mode. At each step a character was sampled based on the conditional probability distribution calculated from the RNN model and the generated character will then be used as the input character for generation of the next character.

prediction utilizing a training set consisting of 12 million reactions from scientific literature. Their system can solve twice as many molecules' retrosynthesis plans as the rule-based method. Liu *et al.* [78] used neural sequence-to-sequence models for retrosynthetic prediction. They used 50 000 reactions obtained from US patents to train the network and obtained similar accuracy to rule-based methods.

Application of convolutional neural networks to predict ligand–protein interactions

Assessing the interaction between a protein and a ligand is the crucial part of the molecular docking program and a lot of scoring functions were developed either based on forcefields or knowledge from existing protein–ligand complex structures [79]. Inspired by the success of CNNs in image analysis, several studies have been recently published in applying a CNN to score protein–ligand interaction. A typical example is the investigation done by Ragoza *et al.* [80]. The protein–ligand structures were discretized into a grid with a resolution of 0.5 Å. The grid was 24 Å on each side and centered on the binding site. Each atom was described with a function, and atom densities over the grid were generated to form the input matrix. Multilayer CNN models were defined and trained using the Caffe DL framework. The CNN scoring outperformed AutoDock Vina [81] on the CSAR inter-target pose-prediction dataset [82], but performed worse for intra-target ranking of poses. Other studies utilizing CNNs or DNNs have also been published [83–85]. Although some encouraging results have been obtained with convolutional networks, it is not clear whether they will consistently improve results compared to currently used scoring functions.

Benchmark datasets within chemoinformatics

The rapid advances made in the field of image recognition can be attributed to not only the emergence of novel algorithms but also to the existence of canonical and large datasets. The standardized dataset would allow the community to conveniently benchmark or evaluate developed machine learning methods. The yearly ImageNet Large Scale Visual Recognition Competition (ILSVRC) [86] has seen the birth of many influential CNN architectures.

Although several open-source chemoinformatics datasets [87,88] are available, their impact on machine learning method development is still limited owing to the limited size of those datasets, lack of diverse ways of splitting training and test-sets and, more importantly, lack of a standard evaluation platform for proposed new algorithms. Inspired by WordNet [89] and ImageNet [90], Wu *et al.* [91] introduced the MoleculeNet dataset by curating

Please cite this article in press as: Chen, H. et al. The rise of deep learning in drug discovery, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2018.01.039

a number of diverse collections including quantum mechanics, physicochemical, biophysics and physiological datasets, and developing a suite of software implementing many known molecule representations and machine learning algorithms. MoleculeNet is built on the open source package DeepChem [92] and provides easy access to some popular DL algorithms existing in DeepChem. This will largely facilitate comparison and development of novel machine learning algorithms in the future.

Application of deep learning in biological imaging analysis

In the drug discovery process, biological imaging and image analysis are widely used at various stages from preclinical R&D to clinical trials. Imaging enables scientists to see the phenotypes and behaviors of hosts (human or animals), organs, tissues, cells and subcellular components. Through digital image analysis, the hidden biology and pathology, as well as the drug mechanism of action, are revealed. Examples of imaging modalities are fluorescently labeled or unlabeled microscopic images, computed tomography (CT), MRI, positron emission tomography (PET), tissue pathology imaging and mass-spectrometry imaging (MSI). DL has also made its way to successes in biological image analysis and many studies reported a superior performance compared with classical classifiers.

For microscopic images, CNNs have been used [93,94] for segmenting and subtyping individual fluorescently labelled cells, as well as unlabeled imageries from phase contract microscopy [95,96]. Other traditionally laborious tasks from preclinical settings, such as cell tracking [96] and colony counting [97], could also be automated using DL. Images from tissue pathology are typically complex in nature compared with the fluorescently labeled images owing to rich tissue morphology. Nevertheless, at the cellular level, the segmentation and classification of individual cells were achieved in breast and colon tissues stained with hematoxylin and eosin (H&E) staining [98,99]. At the tissue region level, the tumor regions from H&E-stained breast tissue were identified through DL [100], whereas the extra categories of leukocytes and fat tissue can also be recognized [101]. Beyond basic image segmentation, DL has already been used for the histopathological diagnosis with H&E and the immunohistochemistry stained tissue [102,103].

The application of DL was also applied for the analysis of CT [104–106], MRI [107,108] and PET [108] imaging. Besides the popular application of the image segmentation [106,107] and classifications [104,105], its utilities have also been shown in content-based image retrieval [109] and it was reported that DL methods outperformed the popular ISOMAP and Elastic Net methods.

For the emerging MSI, similar to the application of DL in tissue pathology, tumor subtyping can be performed by high-resolution matrix-assisted laser desorption/ionization (MALDI) MSI [110]. Given that MSI can visualize the metabolic information of a tissue, sub-regions of a tumor with metabolic heterogeneity from desorption electrospray ionization (DESI) MSI can already be detected through DL [111]. Finally, in an unusual imaging area: flow cytometry, DL enabled the cell classification in real-time for high-throughput applications [112]. The training of DNNs for imaging is time-consuming and requires dedicated GPU processing. Furthermore, in the context of high-throughput imaging screening, good-quality training sets are rare. Therefore, image features trained from natural scenes and other datasets were 'borrowed' to perform biological image segmentations and classifications, and robust performances were reported [101,113].

Future development of deep learning in drug discovery

Machine learning methods and DL in particular generally need large datasets for training; however, the human brain has the capability of learning through only a few examples. How to learn with only a small amount of available data is therefore one of the hottest topics in machine learning. A DL example of exploiting auxiliary data to improve a model with only a few data points is matching networks [114], which was proposed as a variant of one-shot learning. Improved results were obtained when the auxiliary data were included. Methods like one-shot learning are relevant to drug discovery, where medicinal chemists often work on novel targets with limited data available. Altae-Tran *et al.* [115] utilized the LSTM method on chemoinformatics datasets to build models with a very small training set and promising results were reported. Very recently, a new type of architecture has been used in DL: memory augmented neural networks. The first version was the neural Turing machine. This architecture was significantly improved with a differentiable neural computer (DNC) [116]. DNCs have been applied to several problems like question-answering systems and finding the shortest path in graphs. However, these more-advanced architectures have not been applied so far in drug discovery.

Concluding remarks

Machine learning has been used since the late 1990s in drug discovery and has established itself as a useful tool in drug discovery. A recent extension of the machine learning toolbox is DL. In comparison with other methods, DL has a much more flexible architecture so it is possible to create a NN architecture tailor-made for a specific problem. A disadvantage is that DL in general needs very large training sets. A relevant question is: is DL is superior to other machine learning methods? We believe it is still too early to draw any firm conclusion, the results so far indicate that DL is superior for certain tasks like image analysis and very useful for *de novo* molecular design and reaction predictions. For tasks with structured input descriptors, DL seems to perform at least on-par with other methods. The most relevant example is bioactivity prediction where DL seems to achieve better performance overall through multitask learning. However, other machine learning methods are also improving. One example is the XGBoost [117] method, which has dominated Kaggle competitions for structured input data [118] after its introduction. Thus, in practice the choice of method used in bioactivity prediction might depend on which method the modeler is most familiar with. If different machine learning methods achieve roughly the same accuracy, the limit of what can be achieved with a machine learning model could depend on experimental uncertainty for the data and dataset size rather than the specific algorithm used.

Conflicts of interest

The authors have no conflicts of interest to declare.

Please cite this article in press as: Chen, H. et al. The rise of deep learning in drug discovery, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2018.01.039

Acknowledgments

The authors thank Christian Tyrchan, Lars Carlsson, Thierry Kogej and Clive Green for valuable discussion on deep learning and machine learning in general. This research has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement no. 676434, 'Big Data in Chemistry' (BIGCHEM). The article reflects only the authors' views and neither the European Commission nor the Research Executive Agency (RREA) are responsible for any use that could be made of the information it contains.

References

1 National Security Agency statement. Available at: https://www.nsa.gov/news-features/press-room/statements/2013-08-09-the-nsa-story.shtml
2 Gantz, J. and Reinsel, D. (2011) Extracting value from chaos. Available at: https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf
3 Gantz, J. and Reinsel, D. (2010) The digital universe decade – are you ready? Available at: https://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf
4 Howard, J. (2013) The business impact of deep learning. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 1135
5 Top Strategic Technology Trends for 2018. Available at: http://www.gartner.com/technology/research/top-10-technology-trends/
6 Papadatos, G. *et al.* (2015) Activity, assay and target data curation and quality in the ChEMBL database. *J. Comput. Aided Mol. Des.* 29, 885–896
7 Kim, S. *et al.* (2016) PubChem substance and compound databases. *Nucleic Acids Res.* 44, D1202–D1213
8 Gilson, M.K. *et al.* (2016) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* 44, D1045–D1053
9 Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Mach. Learn.* 20, 273–297
10 Salt, D.W. *et al.* (1992) The use of artificial neural networks in QSAR. *Pestic. Sci.* 36, 161–170
11 Ho, T.K. (1998) The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 832–844
12 Ammad-Ud-Din, M. *et al.* (2016) Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics* 32, i455–463
13 Ching, T. *et al.* (2017) Opportunities and obstacles for deep learning in biology and medicine. *bioRxiv* http://dx.doi.org/10.1101/142760
14 Goh, G.B. *et al.* (2017) Deep learning for computational chemistry. *J. Comput. Chem.* 38, 1291–1307
15 Gawehn, E. *et al.* (2016) Deep learning in drug discovery. *Mol. Inf.* 35, 3–14
16 Mamoshina, P. *et al.* (2016) Applications of deep learning in biomedicine. *Mol. Pharm.* 13, 1445–1454
17 Ekins, S. (2016) The next era: deep learning in pharmaceutical research. *Pharm. Res.* 33, 2594–2603
18 Baskin, I.I. *et al.* (2016) A renaissance of neural networks in drug discovery. *Expert Opin. Drug Discov.* 11, 785–795
19 McCulloch, W.S. and Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biol.* 52, 115–133
20 Dreyfus, S.E. (1973) The computational solution of optimal control problems with time lag. *IEEE Trans. Autom. Control* 18, 383–385
21 Srivastava, N. *et al.* (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958
22 Wan, L. *et al.* (2013) Regularization of neural networks using DropConnect. In *Proceedings of the 30th International Conference on Machine Learning*, (Vol. 28) (Sanjoy, D. and David, M., eds) In pp. 1058–1066, PMLR
23 Nair, V. and Hinton, G.E. (2010) Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, Omnipress. pp. 807–814
24 TensorFlow™. Available at: https://www.tensorflow.org/
25 Caffe. Available at: http://caffe.berkeleyvision.org/
26 PYTORCH. Available at: http://pytorch.org/
27 Keras. Available at: https://keras.io/
28 Theano. Available at: http://deeplearning.net/software/theano/
29 Lee, H. *et al.* (2011) Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Commun. ACM* 54, 95–103
30 Szegedy, C. *et al.* (2015) Going deeper with convolutions. *CVPR*, IEEE Computer Society pp. 1–9
31 Fernández, S. *et al.* (2007) An application of recurrent neural networks to discriminative keyword spotting. In *Proceedings of the 17th International Conference on Artificial Neural Networks*, Springer-Verlag. pp. 220–229
32 Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory. *Neural Comput.* 9, 1735–1780
33 Bengio, Y. (2009) Learning deep architectures for AI. *Found. Trends Mach. Learn.* 2, 1–127
34 Kingma, D.P. and Welling, M. (2013) Auto-encoding variational bayes. *ArXiv* 1312.6114
35 Ma, J. *et al.* (2015) Deep neural nets as a method for quantitative structure–activity relationships. *J. Chem. Inf. Model.* 55, 263–274
36 Mayr, A. *et al.* (2016) DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* http://dx.doi.org/10.3389/fenvs.2015.00080
37 Dahl, G.E. *et al.* (2014) Multi-task neural networks for QSAR predictions. *ArXiv* arXiv:1406.1231
38 Ramsundar, B. *et al.* (2017) Is multitask deep learning practical for pharma? *J. Chem. Inf. Model.* 57, 2068–2076
39 Koutsoukas, A. *et al.* (2017) Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *J. Cheminformatics* 9, 42
40 Gaulton, A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–1107
41 Lenselink, E.B. *et al.* (2017) Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminformatics* 9, 45
42 Subramanian, G. *et al.* (2016) Computational modeling of beta-secretase 1 (BACE-1) inhibitors using ligand based approaches. *J. Chem. Inf. Model.* 56, 1936–1949
43 Aliper, A. *et al.* (2016) Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharm.* 13, 2524–2530
44 NIH LINCS program. Available at: http://www.lincsproject.org/LINCS/
45 Merkwirth, C. and Lengauer, T. (2005) Automatic generation of complementary descriptors with molecular graph networks. *J. Chem. Inf. Model.* 45, 1159–1168
46 Lusci, A. *et al.* (2013) Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J. Chem. Inf. Model.* 53, 1563–1575
47 Xu, Y. *et al.* (2015) Deep learning for drug-induced liver injury. *J. Chem. Inf. Model.* 55, 2085–2093
48 Morgan, H.L. (1965) The generation of a unique machine description for chemical structures—a technique developed at Chemical Abstracts Service. *J. Chem. Doc.* 5, 107–113
49 Duvenaud, D. *et al.* (2015) Convolutional networks on graphs for learning molecular fingerprints. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, MIT Press. pp. 2224–2232
50 Kearnes, S. *et al.* (2016) Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* 30, 595–608
51 Xu, Y. *et al.* (2017) Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. *J. Chem. Inf. Model.* 57, 2672–2685
52 Li, J. *et al.* (2017) Learning graph-level representation for drug discovery. *ArXiv* arXiv:1709.03741
53 Coley, C.W. *et al.* (2017) Convolutional embedding of attributed molecular graphs for physical property prediction. *J. Chem. Inf. Model.* 57, 1757–1772
54 Gilmer, J. *et al.* (2017) Neural message passing for quantum chemistry. *ArXiv* arXiv:1704.01212
55 Li, Y. *et al.* (2015) Gated graph sequence neural networks. *ArXiv* arXiv:1511.05493
56 Kipf, T.N. and Welling, M. (2016) Semi-supervised classification with graph convolutional networks. *ArXiv* arXiv:1609.02907
57 Bjerrum, E.J. (2017) SMILES enumeration as data augmentation for neural network modeling of molecules. *ArXiv* arXiv:1703.07076
58 Goh, G.B. *et al.* (2017) Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. *ArXiv* arXiv:1706.06689
59 Goh, G.B. *et al.* (2017) How much chemistry does a deep neural network need to know to make accurate predictions? *ArXiv* arXiv:1710.02238

Please cite this article in press as: Chen, H. et al. The rise of deep learning in drug discovery, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2018.01.039

60 Gómez-Bombarelli, R. *et al.* (2016) Automatic chemical design using a data-driven continuous representation of molecules. *ArXiv* arXiv:1610.02415
61 Kadurin, A. *et al.* (2017) druGAN: an advanced generative adversarial autoencoder model for *de novo* generation of new molecules with desired molecular properties *in silico*. *Mol. Pharm.* 14, 3098–3104
62 Goodfellow, I.J. *et al.* (2014) Generative adversarial networks. *ArXiv* arXiv: 1406.2661
63 Blaschke, T. *et al.* (2017) Application of generative autoencoder in de novo molecular design. *Mol. Inf.* http://dx.doi.org/10.1002/minf.201700123
64 Segler, M.H.S. *et al.* (2018) Generating focussed molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4, 120–131
65 Yuan, W. *et al.* (2017) Chemical space mimicry for drug discovery. *J. Chem. Inf. Model.* 57, 875–882
66 Jaques, N. *et al.* (2016) Sequence Tutor: conservative fine-tuning of sequence generation models with KL-control. *ArXiv* arXiv:1611.02796
67 Leo, A. *et al.* (1971) Partition coefficients and their uses. *Chem. Rev.* 71, 525–616
68 Bickerton, G.R. *et al.* (2012) Quantifying the chemical beauty of drugs. *Nat. Chem.* 4, 90–98
69 Olivecrona, M. *et al.* (2017) Molecular *de-novo* design through deep reinforcement learning. *J. Cheminformatics* 9, 48
70 Benhenda, M. (2017) ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity? *ArXiv* arXiv:1708.08227
71 Metz, L. *et al.* (2016) Unrolled generative adversarial networks. *ArXiv* arXiv:1611.02163
72 Unterthiner, T. *et al.* (2017) Coulomb GANs: provably optimal Nash equilibria via potential fields. *ArXiv* arXiv:1708.08819
73 Corey, E.J. and Wipke, W.T. (1969) Computer-assisted design of complex organic syntheses. *Science* 166, 178–192
74 Coley, C.W. *et al.* (2017) Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* 3, 434–443
75 Jin, W. *et al.* (2017) Predicting organic reaction outcomes with Weisfeiler–Lehman network. *ArXiv* arXiv:1709.04555
76 Segler, M.H.S. and Waller, M.P. (2017) Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry* 23, 5966–5971
77 Segler, M.H.S. *et al.* (2017) Learning to plan chemical syntheses. *ArXiv* arXiv:1708.04202
78 Liu, B. *et al.* (2017) Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Central Science* 3, 1103–1113
79 Pagadala, N.S. *et al.* (2017) Software for molecular docking: a review. *Biophys. Rev.* 9, 91–102
80 Ragoza, M. *et al.* (2017) Protein–ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* 57, 942–957
81 Trott, O. and Olson, A.J. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455–461
82 Dunbar, J.B., Jr *et al.* (2011) CSAR benchmark exercise of 2010: selection of the protein–ligand complexes. *J. Chem. Inf. Model.* 51, 2036–2046
83 Gomes, J. *et al.* (2017) Atomic convolutional networks for predicting protein–ligand binding affinity. *ArXiv* arXiv:1703.10603
84 Wallach, I. *et al.* (2015) AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *ArXiv* arXiv:1510.02855
85 Pereira, J.C. *et al.* (2016) Boosting docking-based virtual screening with deep learning. *J. Chem. Inf. Model.* 56, 2495–2506
86 Russakovsky, O. *et al.* (2015) ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision* 115, 211–252
87 Mysinger, M.M. *et al.* (2012) Directory of Useful Decoys, Enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55, 6582–6594
88 Sun, J. *et al.* (2017) ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. *J. Cheminformatics* 9, 17
89 Miller, G.A. (1995) WordNet: a lexical database for English. *Commun. ACM* 38, 39–41
90 Li, F.-F. *et al.* (2009) ImageNet: constructing a large-scale image database. *J. Vision* 9, 1037
91 Wu, Z. *et al.* (2017) MoleculeNet: a benchmark for molecular machine learning. *ArXiv* arXiv:1703.00564
92 DeepChem package. Available at: https://deepchem.io/
93 Angermueller, C. *et al.* (2016) Deep learning for computational biology. *Mol. Syst. Biol.* 12, 878
94 Kraus, O.Z. *et al.* (2015) Classifying and segmenting microscopy images using convolutional multiple instance learning. *ArXiv* arXiv:1511.05286
95 Ronneberger, O. *et al.* (2015) U-net: convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 234–241
96 Ning, F. *et al.* (2005) Toward automatic phenotyping of developing embryos from videos. *IEEE Trans. Image Process.* 14, 1360–1371
97 Ferrari, A. *et al.* (2015) Bacterial colony counting by convolutional neural networks. In Engineering in Medicine and Biology Society (EMBC). *2015 37th Annual International Conference of the IEEE*, IEEE. pp. 7458–7461
98 Cireşan, D.C. *et al.* (2013) Mitosis detection in breast cancer histology images with deep neural networks. *International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer. pp. 411–418
99 Sirinukunwattana, K. *et al.* (2016) Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans. Med. Imaging* 35, 1196–1206
100 Xu, Y. *et al.* (2014) Deep learning of feature representation with multiple instance learning for medical image analysis. *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference*, IEEE. pp. 1626–1630
101 Turkki, R. *et al.* (2016) Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. *J. Pathol. Inf.* 7, 38
102 Vandenberghe, M.E. *et al.* (2017) Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Sci. Rep.* http://dx.doi.org/10.1038/srep45938
103 Litjens, G. *et al.* (2016) Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* 6, 26286
104 Bar, Y. *et al.* (2015) Deep learning with non-medical training used for chest pathology identification. *Proc. SPIE* Vol. 9414 pp. 94140V
105 Cheng, J.-Z. *et al.* (2016) Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Sci. Rep.* 6, 24454
106 Cha, K.H. *et al.* (2016) Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets. *Med. Phys.* 43, 1882–1896
107 Avendi, M. *et al.* (2016) A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Med. Image Anal.* 30, 108–119
108 Li, R. *et al.* (2014) Deep learning based imaging data completion for improved brain disease diagnosis. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 305–312
109 Liu, S. *et al.* (2014) High-level feature based PET image retrieval with deep learning architecture. *J. Nucl. Med.* 55 (Suppl. 1), 2028
110 Behrmann, J. *et al.* (2017) Deep learning for tumor classification in imaging mass spectrometry. *ArXiv* arXiv:1705.01015
111 Inglese, P. *et al.* (2017) Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer. *Chem. Sci.* 8, 3500–3511
112 Chen, C.L. *et al.* (2016) Deep learning in label-free cell classification. *Sci. Rep.* 6, 21471
113 Zhang, W. *et al.* (2016) Deep model based transfer and multi-task learning for biological image analysis. *IEEE Trans. Big Data* http://dx.doi.org/10.1109/TBDATA.2016.2573280
114 Vinyals, O. *et al.* (2016) Matching networks for one shot learning. *ArXiv* arXiv:1606.04080
115 Altae-Tran, H. *et al.* (2017) Low data drug discovery with one-shot learning. *ACS Cent. Sci.* 3, 283–293
116 Graves, A. *et al.* (2016) Hybrid computing using a neural network with dynamic external memory. *Nature* 538, 471–476
117 Chen, T. and Guestrin, C. (2016) XGBoost: a scalable tree boosting system. *ArXiv* 1603 arXiv:1603.02754
118 A Kaggle Master Explains Gradient Boosting. Available at: http://blog.kaggle.com/2017/01/23/a-kaggle-master-explains-gradient-boosting/

Please cite this article in press as: Chen, H. et al. The rise of deep learning in drug discovery, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2018.01.039