CrossMark

PERSPECTIVE

The Next Era: Deep Learning in Pharmaceutical Research

Sean Ekins [1,2]

ABSTRACT Over the past decade we have witnessed the increasing sophistication of machine learning algorithms applied in daily use from internet searches, voice recognition, social network software to machine vision software in cameras, phones, robots and self-driving cars. Pharmaceutical research has also seen its fair share of machine learning developments. For example, applying such methods to mine the growing datasets that are created in drug discovery not only enables us to learn from the past but to predict a molecule's properties and behavior in future. The latest machine learning algorithm garnering significant attention is deep learning, which is an artificial neural network with multiple hidden layers. Publications over the last 3 years suggest that this algorithm may have advantages over previous machine learning methods and offer a slight but discernable edge in predictive performance. The time has come for a balanced review of this technique but also to apply machine learning methods such as deep learning across a wider array of endpoints relevant to pharmaceutical research for which the datasets are growing such as physicochemical property prediction, formulation prediction, absorption, distribution, metabolism, excretion and toxicity (ADME/Tox), target prediction and skin permeation, etc. We also show that there are many potential applications of deep learning beyond cheminformatics. It will be important to perform prospective testing (which has been carried out rarely to date) in order to convince skeptics that there will be benefits from investing in this technique.

ABBREVIATIONS

ADME/Tox	Absorption, distribution, metabolism, excretion/toxicology
AUC	Area under the curve
DILI	Drug induced liver injury
hERG	Human ether a-go-go related gene
PLGA	Poly-lactide-co-glycolide
QSAR	Quantitative structure activity relationships
SVM	Support vector machines

INTRODUCTION

We have previously suggested that cheminformatics should look to other industries that use high performance computing approaches for inspiration (1). Six years on what is surprising is that we may not have to look to industries but instead we already have used the algorithms which we should be considering for cheminformatics (and other areas) in our everyday transactions, work or social-life online. Every time we use our smartphones with voice recognition software like Siri or read the news on them, make a purchase on the internet via Amazon or use our social network software, we take for granted that we are confronted with suggestions of other products we might like to buy or friends to connect too. Large companies such as Baidu, Google, Facebook etc. all use deep learning in facial recognition algorithms alone. We also live in an age in which self-driving cars and robot assistants are emerging rapidly after decades of research. Therefore we are literally surrounded by artificial intelligence software that use machine learning (2) that can in many ways now predict our needs before we know what they are. Perhaps for the first

✉ Sean Ekins
ekinssean@yahoo.com

1 Collaborations Pharmaceuticals, Inc, 5616 Hilltop Needmore Road, Fuquay-Varina, North Carolina 27526, USA

2 Collaborative Drug Discovery, 1633 Bayshore Highway, Suite 342, Burlingame, California 94010, USA

Springer

time we may also be directly seeing a preview of how such software tools could assist us in healthcare related research and development. We just did not know it until now.

In the area of pharmaceutical research and development and specifically that of cheminformatics there are many machine learning methods such as support vector machines (SVM), k-Nearest Neighbors, Naïve Bayesian, Decision Trees etc. (3) which have seen increasing use as our datasets have grown to become 'big data' (4–7). These methods are equally applicable in other areas and come with their own pros and cons (3) that can be used for binary classification, multiple classes or continuous data. The application of different computational approaches and machine learning algorithms to problems tends to follow the growth of datasets (8). As pharmaceutical datasets started out quite small, the methods initially used would focus on local models like pharmacophores and quantitative structure activity relationships (QSAR). In more recent years the biological data amassed from high throughput screening and high content screens has called for different tools to be used that can account for some of the issues with this bigger data (6). The power of computer processing has also increased so that more complex non-linear problems can be solved in real time with relatively inexpensive compute resources. Many of these resulting machine learning models can also be implemented on a mobile phone (9,10). In recent years, there has been increasing use of one approach called deep learning (DL), (which builds on many years of artificial neural network research) (11), that has shown powerful advantages in learning from images and languages (12). This may represent the next era of cheminformatics and pharmaceutical research in general that is focused on mining the heterogeneous big data which is accumulating using more sophisticated algorithms such as DL.

DELVING INTO DEEP LEARNING

Standard artificial neural networks (ANN) approaches use an input layer, hidden layer and output layer where each connection has a weight and these vary during training in order to connect input to output data. This method has been used widely but suffers from overfitting of data, and a poor ability to generalize with an external dataset (3), although more recent versions such as Bayesian regularized artificial neural networks are more difficult to overtrain (13). DL or deep neural networks (3) in many ways is similar to ANN in that it mimics how the brain works and takes information in an input layer but unlike ANN has many hidden layers (14) to combine signals with different weights, passes the results successively deeper in the network until an output layer (Fig. 1). The DL model is trained with a dataset by adjusting the weights to give the response expected for a certain input (e.g.

if a compound is active or inactive or the level of activity/inactivity). The ability to have multiple learnable stages makes this approach more useful for tackling more complex problems. Deep learning can be used for unsupervised learning and appears to work well with noisy data. However it still suffers from the potential to over fit data, the black box problem, as well as higher computational cost than ANN or other methods (15). There has been relatively limited application of DL to pharmaceutical problems to date and very few in the area of cheminformatics compared with other machine learning methods (11). DL tools are available in popular open source statistical software such as R (16). In addition there is TensorFlow (17), Deeplearning4j (18) while Facebook made their deep learning software (Torch) open source (19,20) followed a year later by Microsoft (CNTK) (21). Some of these methods have been summarized in a recent review (22). It should be noted that these deep learning toolkits are likely far from 'plug and play' type software tools for the average scientist which they can input their molecules and data to train a model (or for that matter any training or test datasets) and then generate predictions. It is likely that expertise in using these software toolkits is needed as well as integrating with molecular descriptor software. It is more likely that a specialized programmer / statistician / cheminformatician will be needed with knowledge of the software tools in order to generate the models which can then be made available for others to use. Existing cheminformatics and other software companies could facilitate making deep learning more accessible to non-expert users by developing accessible fully integrated tools which they can use for any dataset.

APPLICATIONS OF DEEP LEARNING IN BIOINFORMATICS

DL has seen a rapid increase in the number of publications associated with bioinformatics (23) and computational biology (22) and has been used in diverse applications (15) such as protein disorder prediction (24). This resulted in a fast approach which was comparable to other machine learning approaches based on area under the curve (AUC) and recall statistics for test datasets. While it did not out-perform the other disorder methods it had the advantage of being fast (24) (Table I). DL has also been used to refine docked protein complexes (42) based on 35,000 unbound docking complexes generated by RosettaDock and tested on 25 docking complexes not in the training set. Although this model was not compared to other methods it resulted in RMSD of 1.40 Å indicating accurate predictions (42). DL has also been used to model structural features of RNA-binding protein targets using CLIP-seq datasets and testing on 24 datasets using the area under the receiver operator characteristic to evaluate the performance which was found to be close to the state of the art

Fig. 1 A schematic of a deep learning neural network applied to cheminformatics and a single property with output as a quantitative or qualitative prediction.
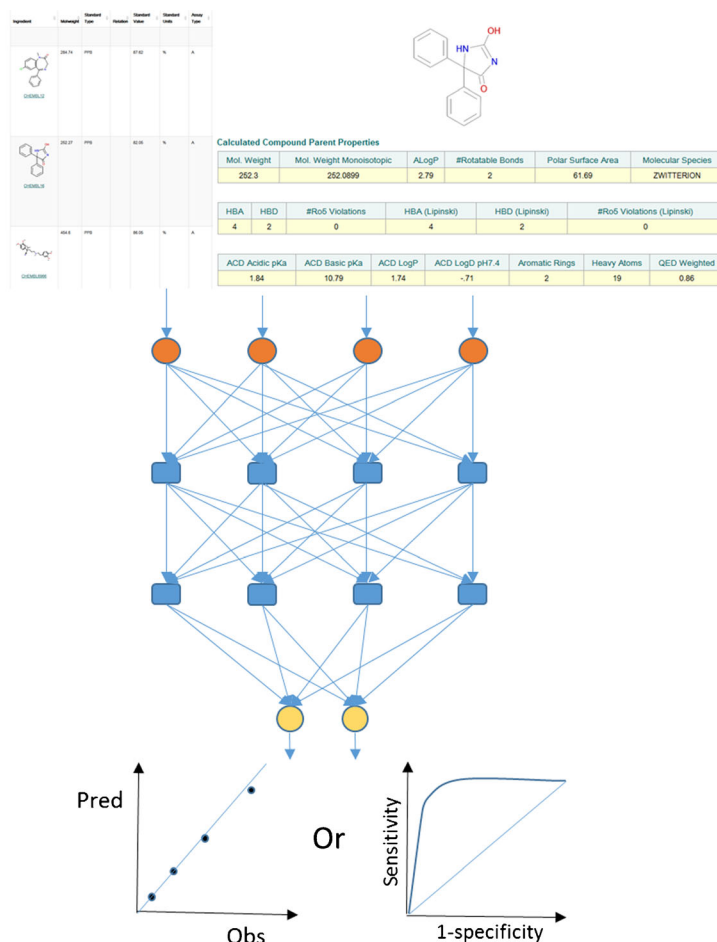
(Table I) (32). DL has been applied and compared with other methods for mechanism of action prediction from high content image analysis data (33) and it was found to be superior to SVM with (87.62 vs 20.95% accuracy). Although it should be noted that the processing time was ten times longer for DL which is likely a major limitation for learning *versus* other methods (33). The preprocessing time can likely be offset by parallelization of calculation processes.

PHARMACEUTICAL APPLICATIONS OF DEEP LEARNING

One of the earliest applications of DL to a pharmaceutically relevant problem was to predict aqueous solubility using four published datasets and was shown to compare favorably to other machine learning methods using 10 fold cross validation (25) (Table I). DL has been used to predict the site of epoxidation in molecules with average AUC > 94.0% for cross validation, although this method was not compared to additional machine learning methods (34) or used for prospective prediction. DL has been put to use with gene expression data to learn from drugs and therapeutic categories using

pathway level or landmark gene level as data reduction methods. In both cases after 10 fold cross validation deep neural networks surpassed SVM used internal testing, suggesting this as a drug repurposing approach (Table I) (37). In the area of drug formulation, predicting drug release from poly-lactide-co-glycolide (PLGA) microspheres showed deep learning to be comparable to random forest, single tree and genetic algorithms after 10 fold cross validation (43). At Merck, deep neural networks have been compared to random forests for use with large QSAR datasets and out-performed random forests for 11 out of 15 datasets (used in a Kaggle crowdsourcing competition) and 13 of 15 datasets in a second evaluation using time-split test sets (27). This utilization by a major pharmaceutical company suggests there is serious interest in the approach and it is likely other companies may have already performed similar evaluations. More frequently DL is applied to a single dataset such as drug induced liver injury. In this case multiple training and testing sets were used and comparison with normal neural networks was performed showing slight improvement with DL (26). In addition, one model was tested with 6 datasets as a form of external validation. The accuracy of DL with DILI data was around 60% for these test sets which is comparable to what has been seen elsewhere

Table I Examples of Deep Learning Studies

Example end point modelled	Dataset size	Model statistics	Summary	References
Solubility	1144, 1026, 74, 125	0.92 RMSE 0.58, 0.91 RMSE 0.60, 0.81 RMSE 0.72, 0.67 RMSE 0.90 All 10 fold cross validation	Did not compare to other machine learning methods themselves. Addition of log P in some cases improved models.	(25)
Drug Induced Liver Injury	190, 475, 1065	80.5, 88.4, 70.1, Accuracies from internal cross validation	DL Models also assessed with external test sets. DL outperformed ANN with different PaDEL and Mold2 descriptors used.	(26)
ADME and target activity	Multiple models from 2092 (microsomes) to 318,795 (hERG)	Compared DL to random forest (RF) models on external datasets (11 of 15 DL models outperform RF models from Kaggle test and 13 out of 15 additional models)	15 Kaggle datasets available consisting of activities and descriptors.	(27)
Biomarkers	62,419 records and 46 blood markers	$R^2 = 0.80$ and 82% prediction accuracy in predicting chronological age within a 10 year window.	Ensemble model performs better $R^2 = 0.82$ accuracy 83.5%	(28)
Protein contact maps	ASTRAL database release 1.75	CASP 8 and 9 data is used for test sets. CMAPpro more accurate on the CASP 8 dataset than other methods. Improvement 10% or higher.	Predicts contact maps with accuracy of 30%	(29)
Cancer diagnosis	13 published cancer gene expression datasets	10 fold cross validation average classification accuracy 46.33–100%	Principal Component Analysis was used to preprocess data	(30)
Gene expression patterns	2000 genes from Allen Developing Mouse Brain Atlas	Average AUC 0.894 versus 0.82 for bag of words for annotating gene expression pattern. Gene ontology functional annotation average AUC = 0.59 vs 0.57 for bag of words	Used deep convolutional neural network. Relative performance of different methods differs across different developing stages.	(31)
Protein disorder prediction	1111 proteins	AUC 86.8 on training set AUC 80.9 on CASP10 dataset	More sophisticated methods such as DISOPRED3 and DNdisorder have test AUC 87.2 and 82.3 respectively	(24)
RNA binding protein features	24 datasets	10 fold cross validation used AUROC values 0.71–0.99	Outperformed the GraphProt method when using base sequence, secondary and tertiary structural profiles.	(32)
High content analysis of breast cancer	148,649 rows	Leave one out cross validation Accuracy 87%	Linear SVM Accuracy 20.95, SVM using radial basis function Accuracy 21.04 Uses a public dataset for training.	(33)
Epoxidation	702	Site of epoxidation AUC 94.9% and separation of epoxidation vs non epoxidation molecules with AUC 79.3% after leave one out cross validation	The DL was compared with a logistic regression model which gave Site of epoxidation AUC 93.7% and epoxidation vs non epoxidation molecules AUC 78.9%	(34)
Tox21	11,764 training set, 296 Leaderboard set, 647 test set	DL average AUC = 0.837, SVM average AUC = 0.832, RF average AUC = 0.803	DL with multitask learning out performed single task learning on 10 of 12 assays and DL won 9 of 15 challenges with nuclear receptor and stress response panels	(35)
Tox21	11,764 training set, 296 Leaderboard set, 647 test set	Consensus model Balanced accuracies from 0.599 to 0.903	Analyzed 12 targets using associative neural networks. Built consensus models. Training and leaderboard sets were combined. Reported to have the best balanced accuracy	(36)
Refinement of docked proteins	35,000 samples of 35 unbound dimer proteins	Tested on 25 test cases across 5 proteins - RMSD 1.40 Å	No information on dataset availability	(42)
Drug repurposing	977 Landmark genes		Deep learning models trained just on gene data did not perform well at	(37)

Table I (continued)

Example end point modelled	Dataset size	Model statistics	Summary	References
	271 signaling pathways	10 fold cross validation deep learning outperforms SVM for both datasets. DL F1 score of 0.70 and SVM 0.53 for pathway 3 class problem.	classifying 12 groups of drugs. Pathways performed better.	
Label free cell classification	Not defined	5 fold Cross validation shows 16 multivariate features can provide a balanced accuracy of 96.4%	Classification of blood cells (OT-II) and cancer cells (SW-480). Deep learning outperforms logistic regression, SVM and Bayesian but all accuracies are greater than 85%	(38)
Classifying microscopy images	103 treatments – 25 images per class for training	Test set accuracy 0.96	Used an available MFC-7 breast cancer imageset BBBC021v1	(39)
Lung tumors	130 patients	Improved mean square error by 29.98% and prediction overshoot by 70.93%	Computation time of 1.54 ms might achieve real-time estimation of intra-fractional variation and better tracking for radiotherapy.	(40)
Target-ligand interaction prediction	sc-PDB, 836 targets and 2710 ligands	5 fold cross validation AUC 0.959. Out performs other published methods such as BLM-NII (AUC 0.858) and CS-PD (AUC 0.799)	Used a pairwise input neural network. % fold accuracy = 0.887	(41)

Deep learning (DL), Support Vector Machine (SVM) and artificial neural network (ANN)

using other algorithms (44). Based on a non-exhaustive assessment of several different end points relevant to pharmaceutical research, while it appears that most have seen utilization of Bayesian or SVM approaches to develop predictive models, few have so far utilized DL (Table II). Recent examples of computational models appearing in this journal alone over the past 18 months include: modeling thermodynamic proxies (76), predicting mouse liver microsomal stability (52), predicting autooxidation (77), drug solubility in human intestinal fluid (78), site of metabolism prediction in CYP2C9 (79), human skin permeability prediction (57,58), blood brain barrier penetration modeling (80), predicting clearance mechanism (58) and skin concentration due to dermal exposure (81).

Table II Representative Examples of Machine Learning Models Applied to Pharmaceutically Relevant End Points to Indicate Areas Used for Machine Learning that Could Be Useful Datasets for Potential Future Use of Deep Learning

Example end point modelled	Naïve Bayesian	Support Vector Machine	Deep Learning
Solubility	(45)	(46)	(25)
Drug Induced Liver Injury	(44)	(47)	(26)
hERG	(48,49)	(50,51)	(27)
ADME	(45,52)	(53)	(27,34)
Blood Brain Barrier penetration	(48)	(50)	
Biological Targets	(5,54)	(55)	(27,56)
Skin Permeability		(57,58)	
Transporters	(10,59–61)	(62–64)	(27)
Mutagenicity	(48,65)	(65,66)	
Formulation		(67)	
Adverse event prediction	(68)	(69,70)	
Counterfeit drug detection		(71)	
Docking		(72)	
Small molecule pKa		(3)	

Other publically accessible datasets are available in OCHEM (73), Chembench (74), CDD (75) etc

Many of these datasets could likely utilize and benefit from DL and it would be of interest to see for how many an improvement in predictions could be obtained.

THE FUTURE OF DEEP LEARNING

While there has been recent exhaustive analysis of artificial intelligence and its impact on jobs, ethical considerations and geopolitical impact (82) there have been very few discussions of the potential for using DL in pharmaceutical research (14,15). Based on the results obtained to date which admittedly have focused on internal validation with little prospective testing as seen with other machine learning methods (Table I) (83,84), DL appears promising and will likely see greater application in the years ahead. Perhaps the largest example of validation of DL models alongside other machine learning approaches is in the case of the Tox21 Challenge. DL with multitask learning (35) slightly outperformed the closest consensus artificial neural network method (36) across nuclear receptor and stress response datasets (Table I).

It is yet to be seen if DL could facilitate the ultimate robot scientist (85), as we see application to different datasets, it or other machine learning methods may become an invisible research assistant, a tool that we take for granted to perform the predictions we need before performing the experiments. This may come with challenges such as how much power do we provide to the software to make the decisions for us. Which raises the question of whether the experiments need human involvement at all? The rapid development of DL outside science suggests it is far from a new fad, and the impact is already being felt in numerous areas from fraud detection to internet search engines. So how long before DL is widespread in pharmaceutical research (14) and what can we expect? It is possible that DL could be the source of more predictive models but hurdles remain on the implementation and accessibility of models. What is clearly needed is software that is tightly integrated with the data to be modeled. This data would most frequently reside in private or public databases and could represent many different endpoints both quantitative and qualitative (Fig. 2). Therefore any efforts to bring the molecules, sources of data and DL algorithms together would greatly streamline model generation and make it more accessible to other scientists. But as with other computational modeling approaches we may also want to consider the applicability domain (86) and various factors such as the quality of the underlying data (87,88) which may determine the utility and relevance of a DL model for making a prediction (89).

A major concern would be how could DL shape research and the future of science and biomedical research in particular? It is possible that DL or any machine learning method might be able to assist by increasing the efficiency of research

Fig. 2 Using public and private data for generating deep learning models for application across vast numbers of endpoints whether quantitative or qualitative (classification models).
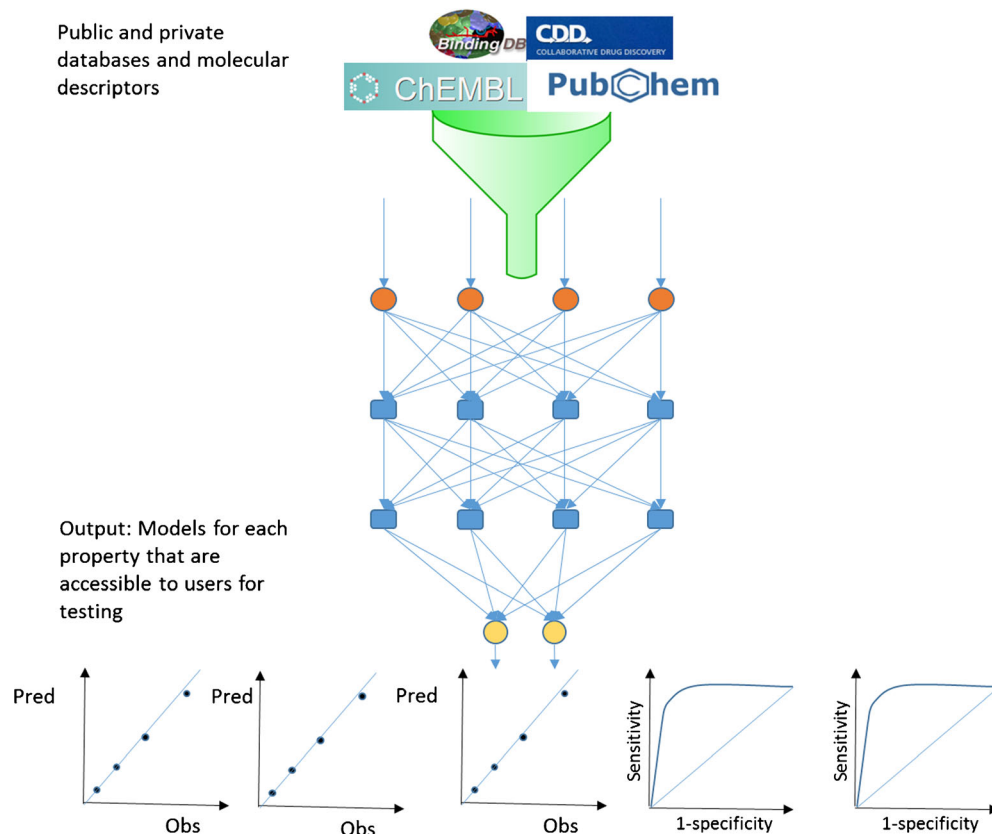
and perhaps rule out likely less successful avenues of research. This is especially important in areas where research funding is tight like rare (90) and neglected diseases (91). Any advantage that DL could provide would be welcome in these and other resource constrained areas as a side effect. Clearly we should be educating the next generation of pharmaceutical researchers to use a wide array of machine learning approaches as well as assessing the likely impact and application of DL. Developing scientists that can generate predictions *in silico* and test them *in vitro* or *in vivo* would also be welcomed. It may be only a short time before we have vice presidents of machine learning or DL in pharmaceutical and biotechnology companies. There are of course still many criticisms of these black box machine learning approaches but it is probably now accepted that with greater accuracy in prediction will come limitations in transparency.

While we have only just seen the beginning of the era of DL, we should be prepared for how it will be used and its potential impact on a wide array of potential pharmaceutically relevant endpoints (Table II). Already comparisons of DL with additional machine learning algorithms have shown that it frequently improves upon the state of the art using predominantly cross validation as the form of evaluation. We will likely see improvements in DL and maybe even alternative approaches that are superior by combining with other machine learning or other methods or data as consensus approaches. This in particular is an area we have yet to see developed for pharmaceutical applications. If we are to imagine machine learning models being ubiquitous in the pharmaceutical industry, DL may facilitate that. We predict the near future will likely see an increase in studies published in this and other journals applying DL and comparing it prospectively and retrospectively to other machine learning methods for predicting various molecule properties important for pharmaceutical research.

While at the time of writing there are over 100 DL startup companies globally, few are focused on pharmaceutical applications (92,93). We anticipate that this will be an active area as the DL connection will be one way to attract technology investors who would normally steer clear of drug discovery and pharmaceuticals as an investment. It is likely that researchers in academia and industry could immediately apply DL if it was more accessible in software and they could plug their data into it. In the interim it may require some collaboration with those more experienced data scientists using R and the various available DL toolkits. There are several areas that could perhaps see an immediate benefit, for example in the areas of predicting metabolism, interactions with P450s and transporters as well as other ADME/Tox properties there has been an increasing number of large datasets and computational machine learning models published (Table II) some of which are non-proprietary. Big datasets like melting point with over 300,000 compounds (94), DMSO solubility with over 163,000 compounds (95), all of the ChEMBL (5) datasets and of course data in PubChem (96), would represent "Big Data" (14) that could be used with DL. The data from PubChem and to a lesser extent ChEMBL, may need curation and organization prior to use in DL models. While the focus of this perspective has been predominantly molecule centric, clearly the application of DL outside the pharmaceutical industry attests to the broad array of applications and potential for impact, there is likely overlap in these domains. For example could we learn from social media data using DL methods what might be potential side effects of drugs or even new uses identified by the population. DL could be applied not only to adverse event prediction but also formulation properties, pharmacokinetic simulation, cost effectiveness and even clinical trial simulation and design. How far DL will take us and how quickly it will have an impact on research and the industry remains to be seen, but there is considerable opportunity to develop user accessible tools and apply them now to the accumulating public datasets. DL is already off to a good start in many areas and pharmaceutical researchers would do well to take a closer look and embrace it faster than they have other computational technologies in the past.

ACKNOWLEDGMENTS AND DISCLOSURES

COMPLIANCE WITH ETHICAL STANDARDS

REFERENCES

1. Ekins S, Gupta RR, Gifford E, Bunin BA, Waller CL. Chemical space: missing pieces in cheminformatics. Pharm Res. 2010;27(10): 2035–9.
2. Rost B, Radivojac P, Bromberg Y. Protein function in precision medicine: deep understanding with machine learning. FEBS Lett. 2016;590(15):2327–41.
3. Mitchell JB. Machine learning methods in chemoinformatics. Wiley Interdiscip Rev Comput Mol Sci. 2014;4(5):468–81.
4. Zhu H, Zhang J, Kim MT, Boison A, Sedykh A, Moran K. Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants. Chem Res Toxicol. 2014;27(10):1643–51.

5. Clark AM, Ekins S. Open source bayesian models: 2. Mining a "big dataset" to create and validate models with ChEMBL. J Chem Inf Model. 2015;55:1246–60.

6. Ekins S, Clark AM, Swamidass SJ, Litterman N, Williams AJ. Bigger data, collaborative tools and the future of predictive drug discovery. J Comput Aided Mol Des. 2014;28(10):997–1008.

7. Ekins S, Freundlich JS, Reynolds RC. Are bigger data sets better for machine learning? Fusing single-point and dual-event dose response data for mycobacterium tuberculosis. J Chem Inf Model. 2014;54:2157–65.

8. Ekins S, Ecker GF, Chiba P, Swaan PW. Future directions for drug transporter modelling. Xenobiotica. 2007;37(10):1152–70.

9. Clark AM, Sarker M, Ekins S. New target predictions and visualization tools incorporating open source molecular fingerprints for TB Mobile 2.0. J Cheminform. 2014;6:38.

10. Ekins S, Clark AM, Wright SH. Making transporter models for drug-drug interaction prediction mobile. Drug Metab Dispos. 2015;43:1642–5.

11. Baskin II, Winkler D, Tetko IV. A renaissance of neural networks in drug discovery. Expert Opin Drug Discovery. 2016;11:785–95.

12. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.

13. Burden F, Winkler D. Bayesian regularization of neural networks. Methods Mol Biol. 2008;458:25–44.

14. Gawehn E, Hiss JA, Schneider G. Deep learning in drug discovery. Mol Inf. 2016;35(1):3–14.

15. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of deep learning in biomedicine. Mol Pharm. 2016;13(5):1445–54.

16. Chow J-F. Things to try after useR! – Part 1: Deep Learning with H2O. 2016 Aug 8th Available from: http://www.r-bloggers.com/things-to-try-after-user-part-1-deep-learning-with-h2o/.

17. Anon. TensorFlow. 2016 Aug 8th. Available from: https://www.tensorflow.org/.

18. Anon. Deeplearning4j 2016 Aug 8th. Available from: http://deeplearning4j.org/.

19. Novet J. Facebook open-sources its cutting-edge deep learning tools. 2016 Aug 8th. Available from: http://venturebeat.com/2015/01/16/facebook-opens-up-about-more-of-its-cutting-edge-deep-learning-tools/.

20. Chintala S. FAIR open sources deep-learning modules for Torch. 2016 Aug 8th. Available from: https://research.facebook.com/blog/fair-open-sources-deep-learning-modules-for-torch/.

21. Linn A. Microsoft releases CNTK, its open source deep learning toolkit, on GitHub. 2016 Aug 8th. Available from: http://blogs.microsoft.com/next/2016/01/25/microsoft-releases-cntk-its-open-source-deep-learning-toolkit-on-github/#sm.00013j280xp1sdctrgg21w81es5ov.

22. Angermueller C, Parnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol. 2016;12(7):878.

23. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2016. doi:10.1093/bib/bbw068.

24. Deng X, Gumm J, Karki S, Eickholt J, Cheng J. An overview of practical applications of protein disorder prediction and drive for faster, more accurate predictions. Int J Mol Sci. 2015;16(7):15384–404.

25. Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. J Chem Inf Model. 2013;53(7):1563–75.

26. Xu Y, Dai Z, Chen F, Gao S, Pei J, Lai L. Deep learning for drug-induced liver injury. J Chem Inf Model. 2015;55(10):2085–93.

27. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure-activity relationships. J Chem Inf Model. 2015;55(2):263–74.

28. Putin E, Mamoshina P, Aliper A, Korzinkin M, Moskalev A, Kolosov A, et al. Deep biomarkers of human aging: application of deep neural networks to biomarker development. Aging (Albany NY). 2016;8(5):1021–33.

29. Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. Bioinformatics. 2012;28(19):2449–57.

30. Fakoor R, Ladhak F, Nazi A, Huber M. Using deep learning to enhance cancer diagnosis and classification. In: Proceeding of the 30th International conference on machine learning. Atlanta, GA: JMLR: W&CP; 2013.

31. Zeng T, Li R, Mukkamala R, Ye J, Ji S. Deep convolutional neural networks for annotating gene expression patterns in the mouse brain. BMC Bioinf. 2015;16:147.

32. Zhang S, Zhou J, Hu H, Gong H, Chen L, Cheng C, et al. A deep learning framework for modeling structural features of RNA-binding protein targets. Nucleic Acids Res. 2016;44(4):e32.

33. Kandaswamy C, Silva LM, Alexandre LA, Santos JM. High-content analysis of breast cancer using single-cell deep transfer learning. J Biomol Screen. 2016;21(3):252–9.

34. Hughes TB, Miller GP, Swamidass SJ. Modeling epoxidation of drug-like molecules with a deep machine learning network. ACS Cent Sci. 2015;1(4):168–80.

35. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. Front Environ Sci. 2016;3:80.

36. Abdelaziz A, Spahn-Langguth H, Schramm K-W, Tetko IV. Consensus modeling for HTS assays using in silico descriptors calculates the best balanced accuracy in Tox21 challenge. Front Environ Sci. 2016;4:2.

37. Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. Mol Pharm. 2016;13(7):2524–30.

38. Chen CL, Mahjoubfar A, Tai LC, Blaby IK, Huang A, Niazi KR, et al. Deep learning in label-free cell classification. Sci Rep. 2016;6:21471.

39. Kraus OZ, Ba JL, Frey BJ. Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics. 2016;32(12):i52–9.

40. Park S, Lee SJ, Weiss E, Motai Y. Intra- and inter-fractional variation prediction of lung tumors using fuzzy deep learning. IEEE J Transl Eng Health Med. 2016;4:4300112.

41. Wang C, Liu J, Luo F, Tan Y. Pairwise input neural network for target-ligand interaction prediction. IEEE Int Conf Bioinf and Biomed. 2014:67-70. doi:10.1109/BIBM.2014.6999129.

42. Akbal-Delibas B, Farhoodi R, Pomplun M, Haspel N. Accurate refinement of docked protein complexes using evolutionary information and deep learning. J Bioinform Comput Biol. 2016;14(3):1642002.

43. Zawbaa HM, Szlek J, Grosan C, Jachowicz R, Mendyk A. Computational intelligence modeling of the macromolecules release from PLGA microspheres-focus on feature selection. PLoS One. 2016;11(6):e0157610.

44. Ekins S, Williams AJ, Xu JJ. A predictive ligand-based Bayesian model for human drug induced liver injury. Drug Metab Dispos. 2010;38:2302–8.

45. Clark AM, Dole K, Coulon-Spector A, McNutt A, Grass G, Freundlich JS, et al. Open source bayesian models: 1. Application to ADME/Tox and drug discovery datasets. J Chem Inf Model. 2015;55:1231–45.

46. Cheng T, Li Q, Wang Y, Bryant SH. Binary classification of aqueous solubility using support vector machines with reduction and recombination feature selection. J Chem Inf Model. 2011;51(2):229–36.

47. Low Y, Uehara T, Minowa Y, Yamada H, Ohno Y, Urushidani T, et al. Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. Chem Res Toxicol. 2011;24(8):1251–62.

48. Ekins S. Progress in computational toxicology. J Pharmacol Toxicol Methods. 2014;69(2):115–40.

49. Wang S, Li Y, Wang J, Chen L, Zhang L, Yu H, *et al*. ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. Mol Pharm. 2012;9(4):996–1010.

50. Kortagere S, Chekmarev DS, Welsh WJ, Ekins S. New predictive models for blood brain barrier permeability of drug-like molecules. Pharm Res. 2008;25:1836–45.

51. Leong MK. A novel approach using pharmacophore ensemble/support vector machine (PhE/SVM) for prediction of hERG liability. Chem Res Toxicol. 2007;20(2):217–26.

52. Perryman AL, Stratton TP, Ekins S, Freundlich JS. Predicting mouse liver microsomal stability with "pruned" machine learning models and public data. Pharm Res. 2015;33:433–49.

53. Hou T, Wang J, Li Y. ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. J Chem Inf Model. 2007;47(6):2408–15.

54. Clark AM, Dole K, Ekins S. Open source Bayesian models: 3. Composite models for prediction of binned responses. J Chem Inf Model. 2016;56:275–85.

55. Kim S, Jin D, Lee H. Predicting drug-target interactions using drug-drug interactions. PLoS One. 2013;8(11):e80129.

56. Unterthiner T, Mayr A, Klambauer G, Hochreiter S. Toxicity prediction using deep learning. Available from: https://arxiv.org/pdf/1503.01445.pdf.

57. Baba H, Takahara J, Mamitsuka H. In silico predictions of human skin permeability using nonlinear quantitative structure-property relationship models. Pharm Res. 2015;32(7):2360–71.

58. Baba H, Takahara J, Yamashita F, Hashida M. Modeling and prediction of solvent effect on human skin permeability using support vector regression and random forest. Pharm Res. 2015;32(11):3604–17.

59. Zheng X, Ekins S, Raufman JP, Polli JE. Computational models for drug inhibition of the human apical sodium-dependent bile acid transporter. Mol Pharm. 2009;6(5):1591–603.

60. Diao L, Ekins S, Polli JE. Quantitative structure activity relationship for inhibition of human organic cation/carnitine transporter. Mol Pharm. 2010;7:2120–30.

61. Dong Z, Ekins S, Polli JE. Structure-activity relationship for FDA approved drugs as inhibitors of the human sodium taurocholate cotransporting polypeptide (NTCP). Mol Pharm. 2013;10(3):1008–19.

62. You H, Lee K, Lee S, Hwang SB, Kim KY, Cho KH, *et al*. Computational classification models for predicting the interaction of compounds with hepatic organic ion importers. Drug Metab Pharmacokinet. 2015;30(5):347–51.

63. de Cerqueira Lima P, Golbraikh A, Oloff S, Xiao Y, Tropsha A. Combinatorial QSAR modeling of P-glycoprotein substrates. J Chem Inf Model. 2006;46(3):1245–54.

64. Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF, Chen YZ. Prediction of P-glycoprotein substrates by a support vector machine approach. J Chem Inf Comput Sci. 2004;44(4):1497–505.

65. Xu C, Cheng F, Chen L, Du Z, Li W, Liu G, *et al*. In silico prediction of chemical Ames mutagenicity. J Chem Inf Model. 2012;52(11):2840–7.

66. Hansen K, Mika S, Schroeter T, Sutter A, ter Laak A, Steger-Hartmann T, *et al*. Benchmark data set for in silico prediction of Ames mutagenicity. J Chem Inf Model. 2009;49(9):2077–81.

67. Moss GP, Shah AJ, Adams RG, Davey N, Wilkinson SC, Pugh WJ, *et al*. The application of discriminant analysis and Machine Learning methods as tools to identify and classify compounds with potential as transdermal enhancers. Eur J Pharm Sci. 2012;45(1-2):116–27.

68. Vock DM, Wolfson J, Bandyopadhyay S, Adomavicius G, Johnson PE, Vazquez-Benitez G, *et al*. Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. J Biomed Inform. 2016;61:119–31.

69. Chia CC, Rubinfeld I, Scirica BM, McMillan S, Gurm HS, Syed Z. Looking beyond historical patient outcomes to improve clinical models. Sci Transl Med. 2012;4(131):131ra149.

70. Rochefort CM, Verma AD, Eguale T, Lee TC, Buckeridge DL. A novel method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data. J Am Med Inform Assoc. 2015;22(1):155–65.

71. Degardin K, Guillemain A, Guerreiro NV, Roggo Y. Near infrared spectroscopy for counterfeit detection using a large database of pharmaceutical tablets. J Pharm Biomed Anal. 2016;128:89–97.

72. Khamis MA, Gomaa W, Ahmed WF. Machine learning in computational docking. Artif Intell Med. 2015;63(3):135–52.

73. Sushko I, Novotarskyi S, Korner R, Pandey AK, Rupp M, Teetz W, *et al*. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J Comput Aided Mol Des. 2011;25(6):533–54.

74. Walker T, Grulke CM, Pozefsky D, Tropsha A. Chembench: a cheminformatics workbench. Bioinformatics. 2010;26(23):3000–1.

75. Ekins S, Hohman M, Bunin BA. Pioneering use of the cloud for development of the collaborative drug discovery (cdd) database. In: Ekins S, Hupcey MAZ, Williams AJ, editors. Collaborative computational technologies for biomedical research. Hoboken: Wiley; 2011.

76. Ekins S, Litterman NK, Lipinski CA, Bunin BA. Thermodynamic proxies to compensate for biases in drug discovery methods. Pharm Res. 2016;33(1):194–205.

77. Lienard P, Gavartin J, Boccardi G, Meunier M. Predicting drug substances autoxidation. Pharm Res. 2015;32(1):300–10.

78. Fagerberg JH, Karlsson E, Ulander J, Hanisch G, Bergstrom CA. Computational prediction of drug solubility in fasted simulated and aspirated human intestinal fluid. Pharm Res. 2015;32(2):578–89.

79. Kingsley LJ, Wilson GL, Essex ME, Lill MA. Combining structure- and ligand-based approaches to improve site of metabolism prediction in CYP2C9 substrates. Pharm Res. 2015;32(3):986–1001.

80. Wang W, Kim MT, Sedykh A, Zhu H. Developing enhanced blood-brain barrier permeability models: integrating external bio-assay data in QSAR modeling. Pharm Res. 2015;32(9):3055–65.

81. Hatanaka T, Yoshida S, Kadhum WR, Todo H, Sugibayashi K. In silico estimation of skin concentration following the dermal exposure to chemicals. Pharm Res. 2015;32(12):3965–74.

82. Anon. Special report: the return of the machinery question. In: The Economist; 2016 June 25th. Available from: http://www.economist.com/news/special-report/21700761-after-many-false-starts-artificialintelligence-has-taken-will-it-cause-mass.

83. Ekins S, Reynolds R, Kim H, Koo M-S, Ekonomidis M, Talaue M, *et al*. Bayesian models leveraging bioactivity and cytotoxicity information for drug discovery. Chem Biol. 2013;20:370–8.

84. Zhang L, Fourches D, Sedykh A, Zhu H, Golbraikh A, Ekins S, *et al*. Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. J Chem Inf Model. 2013;53(2):475–92.

85. Whelan KE, King RD. Intelligent software for laboratory automation. Trends Biotechnol. 2004;22(9):440–5.

86. Tetko IV, Bruneau P, Mewes HW, Rohrer DC, Poda GI. Can we estimate the accuracy of ADME-Tox predictions? Drug Discov Today. 2006;11(15-16):700–7.

87. Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model. 2010;50(7):1189–204.

88. Williams AJ, Ekins S, Tkachenko V. Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. Drug Discov Today. 2012;17:685–701.

89. Vracko M, Bandelj V, Barbieri P, Benfenati E, Chaudhry Q, Cronin M, *et al*. Validation of counter propagation neural network models for predictive toxicology according to the OECD principles: a case study. SAR QSAR Environ Res. 2006;17(3):265–84.

90. Ekins S, Wood J. Incentives for Starting Small Companies Focused on Rare and Neglected Diseases. Pharm Res. 2016;33:809–15.

91. Ponder EL, Freundlich JS, Sarker M, Ekins S. Computational models for neglected diseases: gaps and opportunities. Pharm Res. 2014;31(2):271–7.

92. Murnane K. What is deep learning and how is it useful? Forbes. Available from: http://www.forbes.com/sites/kevinmurnane/2016/04/01/what-is-deep-learning-and-how-is-it-useful/#715d1eaf10f0.

93. Murnane K. Thirteen companies that use deep learning to produce actionable results. Forbes. Available from: http://www.forbes.com/sites/kevinmurnane/2016/04/01/thirteen-companies-that-use-deep-learning-to-produce-actionable-results/#4e710eb07967.

94. Tetko IV, M Lowe D, Williams AJ. The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS. J Cheminform. 2016;8:2.

95. Tetko IV, Novotarskyi S, Sushko I, Ivanov V, Petrenko AE, Dieden R, *et al*. Development of dimethyl sulfoxide solubility models using 163,000 molecules: using a domain applicability metric to select more reliable predictions. J Chem Inf Model. 2013;53(8):1990–2000.

96. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, *et al*. PubChem substance and compound databases. Nucleic Acids Res. 2016;44(D1):D1202–13.