nature materials

PERSPECTIVE

https://doi.org/10.1038/s41563-019-0338-z

Exploiting machine learning for end-to-end drug discovery and development

Sean Ekins[1*], Ana C. Puhl[1], Kimberley M. Zorn[1], Thomas R. Lane[1], Daniel P. Russo[1,2], Jennifer J. Klein[1], Anthony J. Hickey[3,4] and Alex M. Clark[5]

A variety of machine learning methods such as naive Bayesian, support vector machines and more recently deep neural networks are demonstrating their utility for drug discovery and development. These leverage the generally bigger datasets created from high-throughput screening data and allow prediction of bioactivities for targets and molecular properties with increased levels of accuracy. We have only just begun to exploit the potential of these techniques but they may already be fundamentally changing the research process for identifying new molecules and/or repurposing old drugs. The integrated application of such machine learning models for end-to-end (E2E) application is broadly relevant and has considerable implications for developing future therapies and their targeting.

Santayana said, "Those who do not remember the past are condemned to repeat it". This observation applies as much to drug discovery as it does to other aspects of human endeavour[1]. The history of drug discovery is a prelude to the emerging potential of computer-assisted data exploration. One constant in drug discovery is that every few years the estimated cost to develop drugs rises further. Less than 20 years ago, developing a drug took ~12 years, cost under a billion dollars, and the biggest challenges were failures due to efficacy or toxicity-induced attrition[2]. In vitro pharmacological profiling implemented earlier in the drug discovery process helped to identify some predictable undesirable off-target activity profiles, which would hinder drug candidate development or even lead to market withdrawal if discovered after drug approval[3]. These technologies did not shorten the time or decrease the cost required to get a candidate drug to market, as now the costs of development are upwards of US$2.8 billion[4]. These methods have also not been able to predict clinical failures due to idiosyncratic toxicity, and this may be due to the lack of in vitro–in vivo correlation[5], while efficacy is as complex to predict[6].

Today's pressures are likely to be different in each company, depending on the target market and available resources, although bottlenecks are common[7]. The biggest cost and most time-consuming component associated with drug development is conducting clinical trials. This is reflected in the high prices of these chronic treatments, which puts pressure on the US healthcare and insurance systems as well as on the patients. Resolutions to these problems include efforts to speed up regulatory review and simplify clinical trials. Less than a decade ago one solution proposed was to increase the number and quality of innovative, cost-effective new medicines without incurring unsustainable research and development (R&D) costs[8]. The R&D process itself is recognized as far from the linear pathway commonly described (Fig. 1). This is clarified in a Drug Discovery, Development, and Deployment Map (4DM)[7]. Another way to possibly improve productivity in this complex environment is to implement machine learning across all areas of drug discovery and development[9] for which there are sufficient data to train models. Machine learning is a growing field of artificial intelligence that uses different statistical techniques to enable computers to learn from various data types without being explicitly programmed. This would be analogous to converting the 4DM from a 2D map into a functioning computational model that can be used to make predictions, and requires radically rethinking the whole R&D process, learning from it and optimizing it (Fig. 1). If models were available for all aspects of drug discovery and development, they could be used seamlessly to predict whether a compound was likely to be ultimately clinically viable (Fig. 1). This process could be described as end-to-end (E2E). Potential limitations of a linear combination of models might appear as errors could accumulate depending on the accuracy of each model, which may then influence the overall utility and prediction. Other optimal combinations of models could also be developed as customized pipelines are developed depending on the disease, target and therapeutic type. These efforts build on the recent proposal that machine learning will impact the future of design, synthesis, characterization and application of molecules and materials[10].

Many classification and clustering solutions in biology, medicine, precision phenotyping and clinical diagnostic support systems have leveraged machine learning methods. A subset of these methods are 'unsupervised learning' techniques that can be used to model and learn from multiomics-type data. Generically, this type of machine learning approach attempts to identify meaningful inferences from datasets that lack classification and categorical labels. For example, an approach called computational phenotyping has emerged to embrace the complexity inherent in disease mechanisms with machine learning to define accurate phenotypes and has been used to predict antibiotic resistance phenotypes in a variety of bacterial species[11]. Rather than thinking of these different efforts in isolation we will need to integrate them in the complete discovery and development pathway.

Instead of proposing a single algorithm or approach as the optimal one, we subscribe to the concept that the limits of machine learning are likely to be exposed by experimental data inconsistency and dataset size, rather than the flaws of any individual modelling framework[12]. The impact of this viewpoint is that machine learning

[1]Collaborations Pharmaceuticals, Inc., Raleigh, NC, USA. [2]The Rutgers Center for Computational and Integrative Biology, Camden, NJ, USA. [3]RTI International, Research Triangle Park, NC, USA. [4]UNC Catalyst for Rare Diseases, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [5]Molecular Materials Informatics, Inc., Montreal, Quebec, Canada. *e-mail: sean@collaborationspharma.com

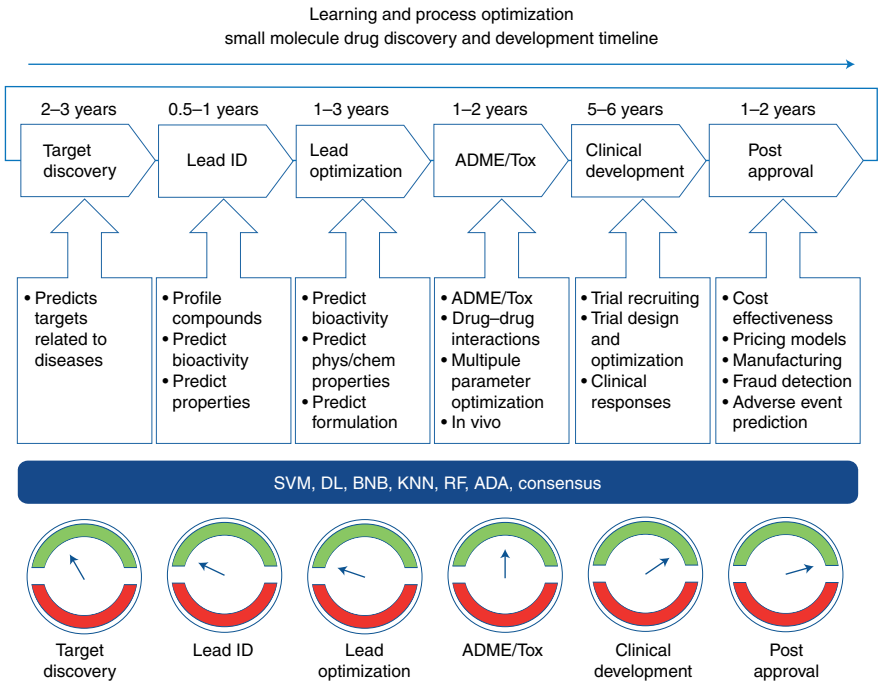

Fig. 1 | Implementing end-to-end (E2E) machine learning models at all stages of drug discovery and development illustrating some of the key areas that could be modelled. A drug discovery and development dashboard for E2E machine learning provides the go–no–go decisions based on inputs of machine learning algorithms (SVM, support vector machine; DL, deep learning; NB, naive Bayesian; KNN, *K* nearest neighbours; RF, random forest; ADA, AdaBoost) or a consensus.

can be applied for the same cost to identify treatments for a variety of diseases and may level the drug development field for smaller companies and researchers. Access to a collection of predictive models for many of the known diseases or related targets could help find additional compounds for testing that may have previously been overlooked[13]. Models for multiple targets already assist in the prediction of off-target effects[14] as well as predicting the most potent compound-target interactions[15].

The tipping point for machine learning

In order to build machine learning models, high-quality data are needed. The past ten years have seen a dramatic increase in the amount of public chemical and biological data in PubChem[16], ChEMBL[17] and other databases that include screening data. We now have millions of molecules and bioactivities for different disease targets as well as for absorption, distribution, metabolism, excretion and toxicology (ADME/Tox) properties. These data are an extremely valuable resource for drug-discovery machine learning applications. We would suggest for many of the targets and diseases we now have plentiful data as indicated by the wide use of ChEMBL and the accuracy of the models generated[18]. A considerable limitation of many databases is the data are not 'model-ready' or machine-readable[19], which is needed to successfully use any machine learning methodology. When you can download data in electronic formats, expert curation is always required to ensure compatibility of data, and domain expertise is important to build and use machine learning models. ChEMBL does a much better job of curation of data but it still takes some effort to prepare for building a machine learning model.

Machine learning models, such as support vector machines[20], *K* nearest neighbours[21], naive Bayes[22], random forest[23] and many other methods[24], have long been utilized for drug discovery. However, recent interest in deep learning or deep neural networks (DNNs) for drug discovery has catalysed interest in machine learning in this field more broadly. DNNs have been used in pattern recognition and machine learning[25], sparking their use in pharmacology and drug discovery[26] and becoming a source for numerous recent reviews[12]. DNNs have been used in various pharmaceutical applications from docking to virtual screening and beyond (Table 1), but the rise in prominence is linked to increased computational power and the availability of larger datasets. While DNNs are inspired by biological neural networks and consist of layers of interconnected neurons, much of the interest in them is centred around the flexibility of their architecture[27], which allows the generation of models for single task or multitask machine learning[28] as well as predicting drug–target interactions[29]. The use of DNNs is still in its relative infancy and has limited applications for cheminformatics as compared with other methods[30]. While DNN algorithms are increasingly available[9], they are not 'plug and play' and their use takes significant time to optimize. Also, the selection of which machine learning algorithm to use with each dataset is not readily predictable and there is really no agreement as to which algorithm is the best for cheminformatics versus other uses. One group has suggested using several 'benchmark' datasets for comparing the predictive ability of different molecular machine learning algorithms[31], while others have performed comparisons using target-related datasets from ChEMBL[18]. The assessments for DNNs could also be applied to essentially any public drug-discovery-relevant small molecule bioactivity dataset[32], but to date this algorithm has rarely been used for prospective prediction and in some respects this is a limitation of many of the machine learning drug-discovery studies published. DNNs have also been used to create novel features/descriptors from their molecular structure as an alternative to traditional molecular descriptors[33]. Two-dimensional structures of molecules have seen use as an input to predict toxicity using the Tox21 benchmark set[34], which is also part of a platform called MoleculeNet[35]. Generative DNNs have also been described for the generation virtual libraries of molecules and these enable de novo drug design with optimized

Table 1 | Areas of relevance to drug discovery and development with substantial data available where machine learning models have been applied illustrating the potential of E2E machine learning

End points modelled
Target discovery[14,15]
Molecule synthesis[36,37]
Small molecule physicochemical properties[82]
Solubility[47]
Drug induced liver injury[83]
hERG[84]
ADME properties[83]
Blood–brain barrier penetration[85]
Skin permeability[86]
Transporters[45]
Mutagenicity[87]
Drug induced liver injury[83]
In vivo pharmacokinetics[88]
Reproductive toxicology[89]
Formulation[90]
Environmental impact[91]
Pharmacoeconomics/cost effectiveness analysis/policy decisions[92]
Clinical trial: recruiting, design, optimization, success and failure[6]
Manufacturing[93]
Counterfeit drug detection[94]
Post marketing surveillance adverse event prediction[95]
Electronic health records[40]

properties[36]. However, it should be pointed out that such proof of concept studies have not synthesized molecules to validate the predictions and this needs to happen to provide evidence of their value. The closest example to this ideal scenario has purchased close analogues to the molecules generated with generative DNNs and tested them against different kinases at 10 μM, identifying several hits[37].

While there are several machine learning frameworks and tools available today aimed at using small molecules and related data, they are not at the point where they are universally accessible to all scientists. The requirement for expert users in many ways has been the Achilles' heel of cheminformatics, whereas computational tools for bioinformatics have found broader use due to their accessibility. Therefore, we need to rethink how to generally make the machine learning models for the pharmaceutical industry more usable and user-friendly to increase the number of potential users and applications. We are at a clear tipping point for machine learning and deep learning in particular, but it has taken decades to reach this point and yet full integration of these models is still likely to be a work in progress.

Machine learning models in action

Machine learning methods in the pharmaceutical industry are most commonly used for virtual screening of compounds, reducing the need to generate more high-throughput screening data by cherry-picking compounds and performing low to medium-throughput screening[38]. The same machine learning algorithms have been used widely in both pharmaceutical and toxicological research[30] (Table 1). Statistical machine learning methods have also been used to interrogate, model and learn from complex multiomics data to help to address uncertainties about the connections between different types of data[39]. For example, machine learning methods have been applied to electronic health records to accurately predict multiple medical events from different centres without site-specific data harmonization, with recent data suggesting that deep learning was comparable to regularized logistic regression in this case[40].

Several of our own recent cheminformatics prospective testing efforts have identified compounds active in vitro and in vivo against Chagas disease[13] and the Ebola virus[41] using Bayesian algorithms. This Bayesian approach has also been widely applied to ADME properties by predicting aqueous solubility, mouse liver microsomal stability[42], Caco-2 cell permeability[43], cytotoxicity[44] and interactions with transporters[45]. We have also used many different machine learning algorithms and descriptors in parallel to identify the optimum combination[46] and address complex problems facing the pharmaceutical industry related to the challenges of improving solubility[47] or metabolic stability[48] while retaining bioactivity. These challenges still persist partly because the datasets may not cover sufficient chemical space, and the test molecules could be outside the applicability of the training set of the models. Optimization and understanding the application of machine learning models is generally not trivial. Instead the field has tended to emphasize discovering the 'perfect individual model' and using various forms of cross validation to evaluate them.

We and others have recently performed several analyses using diverse drug discovery datasets and metrics to compare different machine learning methods using one type of frequently used molecular descriptor, namely FCFP6 fingerprints[49]. After fourfold cross validation and ranked normalized scores of metrics, DNNs ranked higher than all the other machine learning methods across all datasets[49]. Other researchers have also compared several machine learning approaches with different datasets from ChEMBL using random split and temporal cross validation to show the superiority of DNNs[50], or fivefold cross validation and leave out 40% as a validation set[51]. A nested cluster cross validation strategy has also been used to show that DNNs outperform these other machine learning methods[18]. We followed these studies by assessing different machine learning methods and molecular descriptors with 18,886 compounds screened against *Mycobacterium tuberculosis*[52]. This comparison demonstrated that DNNs and support vector machines appear to be superior methods regardless of the descriptor type for training and fivefold cross validation. Conversely, external testing of DNN models with a large test set did not perform as well as other machine learning methods. More recently we have evaluated these same machine learning algorithms and descriptors for multiple oestrogen receptor datasets[46]. For predicting compounds within the training set, DNNs had higher accuracy than other methods in fivefold cross validation. For external test set predictions DNN and most classic machine learning models perform similarly regardless of dataset or molecular descriptors[46]. The fact that DNN does not always outperform other methods for external testing and in several cases is not the best, is important to consider due to the computational cost of DNN. This therefore deserves more exhaustive assessment to determine which algorithm to use with each dataset. Our own efforts continue to reflect this pattern, namely that while DNN excels at cross validation assessments, it is generally no better than other machine learning methods for external testing.

Models for all diseases

The majority of global pharmaceutical companies are focused on the major diseases (for example, cancer, cardiovascular, pain, diabetes, arthritis) that conform to a robust business model, and most research scientists are similarly engaged in these endeavours. However, other diseases that in aggregate involve large patient populations and represent major unmet medical needs are gaining attention. There are neglected and tropical diseases (for example, malaria, tuberculosis and others) and rare diseases (defined as affecting less than 200,000 people in the United States) in which

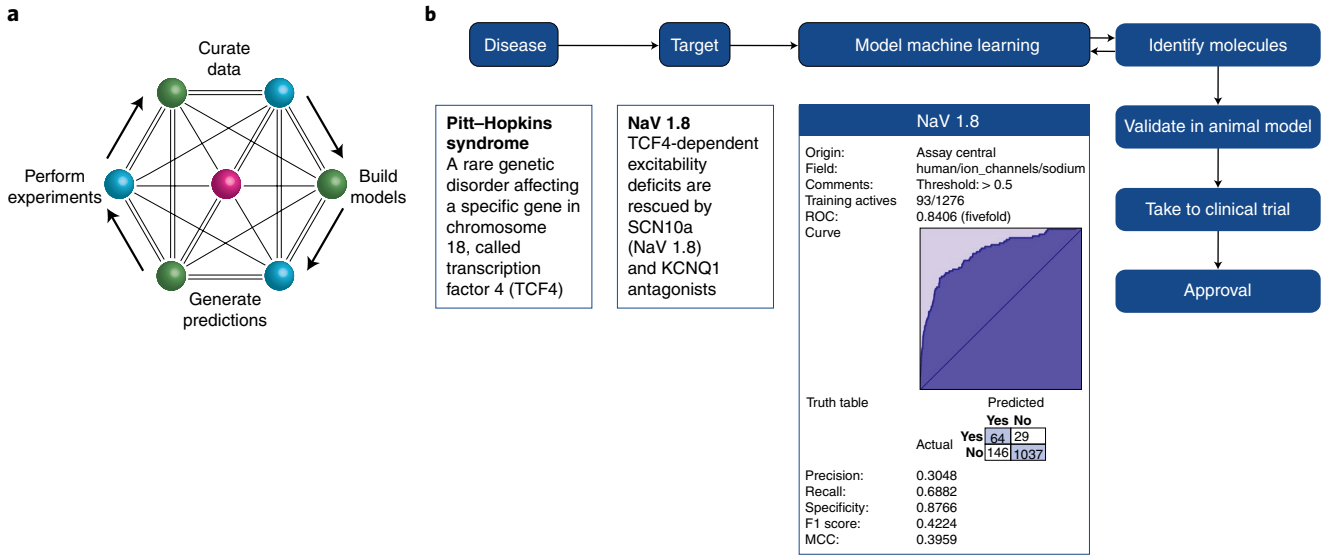

Fig. 2 | Demonstrating iterative drug discovery using machine learning. a, The prospective machine learning approach. **b**, Demonstration of linkage between disease, target and machine learning model using Pitt–Hopkins syndrome as an example[81].

interest has increased as the FDA priority voucher[53] has provided an incentive for companies to develop new treatments. While we have focused our efforts on neglected and tropical disease machine learning models for tuberculosis and malaria, these diseases are in the enviable position of having very large datasets (>300,000 compounds) from high-throughput screening that can be utilized for machine learning[54]. While NIH funding has gone into the high-throughput screening[55], until recently there have not been comparable efforts on data mining or machine learning with these data[16]. Recently, we have stressed the need to scale and 'industrialize' rare disease drug discovery[56] and move towards higher-throughput and collaborative approaches. We have also proposed that machine learning could be used to find treatments for rare diseases using an iterative approach[57]. This methodology would involve first linking the targets for rare diseases, building models for targets related to these diseases, and then use machine learning to identify additional molecules for future testing and validation (Fig. 2). Currently available chemical and biological data relevant for rare disease drug discovery are available but diffuse, existing in an array of public or private databases. Several recent efforts have focused on developing pipelines using natural language processing and human curation to mine promising targets for drug development for rare diseases[58]. These examined diseases with late onset, but clearly there is also an urgent need to address rare diseases with an early onset. Others have initiated different approaches to combat rare diseases through the development of a comprehensive global genotype-phenotype database[59], sharing genomics data[60] or other aspects of rare diseases[61], as well as assist patients and caregivers[62]. To date there are no efforts specifically using machine learning to identify drugs for rare diseases that leverage the relevant datasets for targets that are in the public domain. These approaches could learn from the work performed on neglected and tropical diseases that has used public datasets to identify new compounds[55].

Making models more accessible and interpretable

If we can combine high-quality curated screening data with cutting edge machine learning algorithms and molecular descriptors, there is the opportunity to build models that can be used to reliably predict new molecules for most areas of drug discovery. An iterative loop can be created where a group leverages its expertise with data and models to propose molecules, the experimentalists procure them and measure bioactivities, and the results are returned in a form that can be inserted directly into the model building process (Fig. 2). This type of simplistic approach, while limited to drug discovery, is amenable to both large and small company efforts and can be used across many projects simultaneously to create a pipeline of internal and external projects. We have taken this strategy with our own machine learning software called Assay Central[46] (Fig. 2), representing an accessible approach to scale drug discovery[8]. There is now increasing focus on machine learning in drug discovery, which suggests the utility of such approaches is increasing. This is exemplified by the number of deals between start-up machine learning-based drug discovery companies and big pharma, biotechs or venture capitalist investors[63]. While making machine learning models and predictions more accessible is important to demonstrate impact, efforts to increase the interpretability of these models beyond the 'black box' are critical[64]. We and others have taken different routes to improve this aspect including tools to highlight contributions of models to test molecules[65], identifying training compounds in the same neighbourhood as test molecules and scores of model applicability or overlap[46,64].

Nanoparticles and nanomedicines

Machine learning may be applicable to developing nanomedicines by exploiting large datasets, in an analogous manner to other areas of drug discovery[34]. These efforts can enable the quantitative prediction of desirable molecules before synthesis and focus research on experiments with the most promising candidates. The field of nanomedicine has led to the development of nanoinformatics and the use of data mining and machine learning to develop nano-QSARs to predict functional and structural properties of nanoparticles. A relatively wide array of machine learning approaches[34] have been applied to prediction of different biomedical properties of nanoparticles such as predicting cellular uptake, cytotoxicity, molecular loading, molecular release, nanoparticle adherence, nanoparticle size and polydispersity[66]. Computational methods can be used to predict the particle self-assembly process for targeted drug carrier nanoparticles. Quantitative structure nanoparticle assembly prediction models have been used to generate predictions of nanoassembly that were found to encapsulate drugs with high loadings and have then also been validated in cancer models[67]. Interestingly, a machine learning method has also been described to identify

clinical trials involving nanodrugs and nanodevices from ClinicalTrials.gov[68]. While drug discovery has seen decades of applications of machine learning, for nanoparticle research there are far fewer examples and data available for model building are limited to select nanomaterial databases[69]. As most of the published examples use small datasets (tens to hundreds of molecules), deep learning has limited value and has rarely been applied. Commercially available tools that could enable scientists to develop nanomedicines using these models have yet to be developed to date. The relevance of nanomaterial-related approaches is in drug formulation or delivery and should be considered an integral part of E2E.

Machine learning repurposing

Our thinking need not be limited to discovering just new molecular entities, as machine learning can help explore the patterns of known drugs and their interactions with drug targets and potentially repurpose already regulatory agency approved molecules. Much of this information on potential repurposing opportunities can already be gleaned from public sources[70]. This suggests that we may not even need to screen large numbers of compounds in the future. We now have an abundance of data, powerful and plentiful computing, public and private efforts to develop databases and models as well as accessible ways to test compounds on a fee-for-service basis. Multiscale models defining networks for a given disease can also be used to construct gene expression assays for high-throughput screening. While these are relatively nascent, as we learn more about biology their impact will also expand. For example, in a classic paradigm, inflammatory bowel disease (IBD) signatures were derived from surgical specimens and intersected with Connectivity Map[71] data representing transcriptional readouts across a number of cell lines in response to treatment with many hundreds of drugs, using a novel pattern-matching algorithm[72]. From this research the anticonvulsant drug topiramate was identified and experimentally validated as a novel treatment for IBD. The same approach has been applied to transcriptional profiles of non-small cell lung cancer (NSCLC) that identified imipramine, bepridil, and promethazine and cimetidine as NSCLC inhibitors[73]. The increasing number of drug-repositioning investigations suggests that the reuse of medications for common, rare or orphan diseases is a viable approach[26]. Machine learning can help to assess whether a drug can be repositioned for a novel indication[74]. There are likely many examples of approved drugs finding additional uses for various diseases[55], and obviously the key is to ensure that these reach the patient in a timely manner.

Repurposing efforts could be greatly assisted by obtaining data from inside companies and academic research institutes. Often a major concern about making predictions with computational models is the vastness of chemical space and the potential for selecting compounds with unfavourable properties. These limitations would be mitigated to some extent if previously approved drugs that have extensive ADME/Tox data (or better still have reached phase-I trials) could be repurposed.

The complete E2E model

In summary, while much of the focus of this Perspective has been on cheminformatics, we propose that many areas across the pharmaceutical R&D spectrum and outside of it are ripe for machine learning (Table 1). Machine learning can learn from almost any data type, such as that from research papers, patient records, images, genes, symptoms, diseases, proteins, tissues, species and drug candidates or compounds that have been shown to affect any of the preceding[75]. We could also imagine a complex interaction network between proteins upstream and downstream in the pathway that might dictate if the drug(s) will work. Proteins have isoforms and redundancy, so inhibition of one might not be enough to illicit the desired response. In the same way, inhibition of one pathway might not be enough to achieve the response, since the cell has other pathway mechanisms that would be activated to circumvent the one that has been inhibited. In this context, we can apply machine learning to the whole pathway to evaluate how a network of protein interactions will react to a perturbation in the system such as the drug that is acting on a particular target and this in turn could lead to more personalized medicine[76]. Rather than repeating the mistakes of the past, it is necessary to understand the biological context that gives rise to the disease and which gene network and proteins are operating before beginning drug discovery screens.

Using machine learning to integrate diverse, large-scale data can provide a path to predict which drug effects might best counteract the molecular networks underlying disease or result in less toxicity. This leads us to selecting the best targets and may ultimately help us to predict efficacy. Some approaches using machine learning methods have been developed to detect drug–target interactions[77], which is fundamental to both new drug discovery and drug repositioning. CRISPR-Cas9 has the potential to edit and renovate the harmful genes for personalized therapy and machine learning methods have been applied to predict the off-targets of CRISPR-Cas9 gene editing[78]. Recently, a cancer drug response profile scan (CDRscan) was developed that predicts somatic mutation profile-based drug responsiveness by linking the tumour genomic fingerprint and its sensitivity to drugs and identified 14 oncology and 23 non-oncology drugs as having new potential cancer indications that may result in treatments tailored for each individual patient[79]. In the area of drug safety, a random forest classifier was used to predict the effect of drugs on the fetus. The models successfully identified category C drugs that are likely to be harmful and those likely to be safe for fetal loss or congenital anomalies[80].

The future drug discovery and development process will use machine learning E2E that will impact training of the workforce. There will be a much heavier computational emphasis as they manage a dashboard of projects, molecules and targets across all the aspects of the process and outcomes are predicted in parallel (Fig. 1). Small pharmaceutical companies may then be able to address tens to hundreds of diseases computationally before narrowing to the most promising projects based on a wide variety of these computational models. Using machine learning more broadly across the industry could allow us to move beyond the limitations defined by researcher specialty and data silos, but it will be important to perform prospective validation of the models to demonstrate progress. These efforts will increasingly demonstrate that machine learning algorithms can help us to discover the next generation of drugs.

Received: 19 October 2018; Accepted: 7 March 2019;
Published online: 18 April 2019

References

1. Butler, L. D. et al. Current nonclinical testing paradigms in support of safe clinical trials: an IQ Consortium DruSafe perspective. *Regul. Toxicol. Pharmacol.* **87**, S1–S15 (2017).
2. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates. *Nat. Rev. Drug. Discov.* **3**, 711–715 (2004).
3. Bowes, J. et al. Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat. Rev. Drug. Discov.* **11**, 909–922 (2012).
4. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **47**, 20–33 (2016).
5. Kenna, J. G. Human biology-based drug safety evaluation: scientific rationale, current status and future challenges. *Expert Opin. Drug Metab. Toxicol.* **13**, 567–574 (2017).
6. Gayvert, K. M., Madhukar, N. S. & Elemento, O. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chem. Biol.* **23**, 1294–1301 (2016).
7. Wagner, J. A. et al. Application of a dynamic map for learning, communicating, navigating, and improving therapeutic development. *Clin. Transl. Sci.* **11**, 166–174 (2018).
8. Paul, S. M. et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **9**, 203–214 (2010).

9. Zhavoronkov, A. Artificial intelligence for drug discovery, biomarker development, and generation of novel chemistry. *Mol. Pharm.* **15**, 4311–4313 (2018).
10. Davies, D. W., Butler, K. T., Isayev, O. & Walsh, A. Materials discovery by chemical analogy: role of oxidation states in structure prediction. *Faraday Discuss.* **211**, 553–568 (2018).
11. Drouin, A. et al. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genom.* **17**, 754 (2016).
12. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today.* **23**, 1241–1250 (2018).
13. Ekins, S. et al. Machine learning models and pathway genome data base for trypanosoma cruzi drug discovery. *PLoS Negl. Trop. Dis.* **9**, e0003878 (2015).
14. Lampa, S. et al. Predicting off-target binding profiles with confidence using conformal prediction. *Front. Pharmacol.* **9**, 1256 (2018).
15. Reker, D., Rodrigues, T., Schneider, P. & Schneider, G. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc. Natl Acad. Sci. USA* **111**, 4067–4072 (2014).
16. Kim, S. et al. PubChem substance and compound databases. *Nucleic Acids Res.* **44**, D1202–1213 (2016).
17. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–1107 (2012).
18. Mayr, A. et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* **9**, 5441–5451 (2018).
19. Clark, A. M., Williams, A. J. & Ekins, S. Machines first, humans second: on the importance of algorithmic interpretation of open chemistry data. *J. Cheminform.* **7**, 9 (2015).
20. Christianini, N. & Shawe-Taylor, J. *Support Vector Machines and Other Kernel-Based Learning Methods* (Cambridge Univ. Press, 2000).
21. Shen, M., Xiao, Y., Golbraikh, A., Gombar, V. K. & Tropsha, A. Development and validation of K-nearest neighbour QSPR models of metabolic stability of drug candidates. *J. Med. Chem.* **46**, 3013–3020 (2003).
22. Bender, A. et al. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* **2**, 861–873 (2007).
23. Susnow, R. G. & Dixon, S. L. Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *J. Chem. Inf. Comput. Sci.* **43**, 1308–1315 (2003).
24. Mitchell, J. B. Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4**, 468–481 (2014).
25. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
26. Aliper, A. et al. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharm.* **13**, 2524–2530 (2016).
27. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **55**, 263–274 (2015).
28. Wu, K., Zhao, Z., Wang, R. & Wei, G.-W. Top*P–S*: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. *J. Comput. Chem.* **39**, 1444–1454 (2018).
29. Wen, M. et al. Deep-learning-based drug-target interaction prediction. *J. Proteome Res.* **16**, 1401–1409 (2017).
30. Ekins, S. The next era: Deep learning in pharmaceutical research. *Pharm. Res.* **33**, 2594–2603 (2016).
31. Wu, Z. et al. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
32. Altae-Tran, H., Ramsundar, B., Pappu, A. S. & Pande, V. Low data drug discovery with one-shot learning. *ACS Cent. Sci.* **3**, 283–293 (2017).
33. Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A. & Zhavoronkov, A. 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 (2017).
34. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
35. Rifaioglu, A. S. et al. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Brief Bioinform.* https://doi.org/10.1093/bib/bby061 (2018).
36. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**, eaap7885 (2018).
37. Putin, E. et al. Adversarial threshold neural computer for molecular de novo design. *Mol. Pharm.* **15**, 4386–4397 (2018).
38. McGaughey, G. B. et al. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **47**, 1504–1519 (2007).
39. Johnson, K. W. et al. Enabling precision cardiology through multiscale biology and systems medicine. *JACC Basic Transl. Sci.* **2**, 311–327 (2017).
40. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **1**, 18 (2018).
41. Ekins, S. et al. Machine learning models identify molecules active against Ebola virus in vitro. *F1000Research* **4**, 1091 (2015).
42. Perryman, A. L., Stratton, T. P., Ekins, S. & Freundlich, J. S. Predicting mouse liver microsomal stability with "pruned' machine learning models and public data. *Pharm. Res.* **33**, 433–449 (2015).
43. Clark, A. M. et al. Open source Bayesian models: 1. Application to ADME/Tox and drug discovery datasets. *J. Chem. Inf. Model.* **55**, 1231–1245 (2015).
44. Perryman, A. L. et al. Naive Bayesian models for vero cell cytotoxicity. *Pharm. Res.* **35**, 170 (2018).
45. Sandoval, P. J., Zorn, K. M., Clark, A. M., Ekins, S. & Wright, S. H. Assessment of substrate dependent ligand interactions at the organic cation transporter OCT2 using six model substrates. *Mol. Pharmacol.* **94**, 1057–1068 (2018).
46. Russo, D. P., Zorn, K. M., Clark, A. M., Zhu, H. & Ekins, S. Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction. *Mol. Pharm.* **15**, 4361–4370 (2018).
47. 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.* **53**, 1563–1575 (2013).
48. Stratton, T. P. et al. Addressing the metabolic stability of antituberculars through machine learning. *ACS Med. Chem. Lett.* **8**, 1099–1104 (2017).
49. Korotcov, A., Tkachenko, V., Russo, D. P. & Ekins, S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery datasets. *Mol. Pharm.* **14**, 4462–4475 (2018).
50. Lenselink, E. B. et al. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminform.* **9**, 45 (2017).
51. Koutsoukas, A., Monaghan, K. J., Li, X. & Huan, J. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *J. Cheminform.* **9**, 42 (2017).
52. Lane, T. et al. Comparing and validating machine learning models for mycobacterium tuberculosis drug discovery. *Mol. Pharm.* **15**, 4346–4360 (2018).
53. Ridley, D. B. Priorities for the priority review voucher. *Am. J. Trop. Med. Hyg.* **96**, 14–15 (2017).
54. Ekins, S. et al. Bayesian models leveraging bioactivity and cytotoxicity information for drug discovery. *Chem. Biol.* **20**, 370–378 (2013).
55. Hernandez, H. W. et al. High throughput and computational repurposing for neglected diseases. *Pharm. Res.* **36**, 27 (2018).
56. Ekins, S. Industrializing rare disease therapy discovery and development. *Nat. Biotechnol.* **35**, 117–118 (2017).
57. Ekins, S. & Perlstein, E. O. Doing it all – how families are reshaping rare disease research. *Pharm. Res.* **35**, 192 (2018).
58. Chen, B. & Altman, R. B. Opportunities for developing therapies for rare genetic diseases: focus on gain-of-function and allostery. *Orphanet. J. Rare Dis.* **12**, 61 (2017).
59. Trujillano, D. et al. A comprehensive global genotype-phenotype database for rare diseases. *Mol. Genet. Genomic Med.* **5**, 66–75 (2017).
60. Thompson, R. et al. RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. *J. Gen. Intern. Med.* **29**, 780–787 (2014).
61. Rath, A. et al. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum. Mutat.* **33**, 803–808 (2012).
62. *Rare Disease InfoHub* https://rarediseases.oscar.ncsu.edu (2018).
63. Fleming, N. How artificial intelligence is changing drug discovery. *Nature* **557**, 55–57 (2018).
64. Chuang, K. V. & Keiser, M. J. Adversarial controls for scientific machine learning. *ACS Chem. Biol.* **13**, 2819–2821 (2018).
65. Marchese Robinson, R. L., Palczewska, A., Palczewski, J. & Kidley, N. Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. *J. Chem. Inf. Model.* **57**, 1773–1792 (2017).
66. Jones, D. E., Ghandehari, H. & Facelli, J. C. A review of the applications of data mining and machine learning for the prediction of biomedical properties of nanoparticles. *Comput. Methods Programs Biomed.* **132**, 93–103 (2016).
67. Shamay, Y. et al. Quantitative self-assembly prediction yields targeted nanomedicines. *Nat. Mater.* **17**, 361–368 (2018).
68. de la Iglesia, D. et al. A machine learning approach to identify clinical trials involving nanodrugs and nanodevices from ClinicalTrials.gov. *PLOS ONE* **9**, e110331 (2014).
69. Tropsha, A., Mills, K. C. & Hickey, A. J. Reproducibility, sharing and progress in nanomaterial databases. *Nat. Nanotechnol.* **12**, 1111–1114 (2017).
70. Baker, N. C., Ekins, S., Williams, A. J. & Tropsha, A. A bibliometric review of drug repurposing. *Drug Discov. Today* **23**, 661–672 (2018).
71. Lamb, J. et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).

72. Dudley, J. T. et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med.* **3**, 96ra76 (2011).
73. Schadt, E. E., Buchanan, S., Brennand, K. J. & Merchant, K. M. Evolving toward a human-cell based and multiscale approach to drug discovery for CNS disorders. *Front. Pharmacol.* **5**, 252 (2014).
74. Napolitano, F. et al. Drug repositioning: a machine-learning approach through data integration. *J. Cheminform.* **5**, 30 (2013).
75. Cruz, S. et al. In silico HCT116 human colon cancer cell-based models en route to the discovery of lead-like anticancer drugs. *Biomolecules* **8**, 56 (2018).
76. Fröhlich, H. et al. From hype to reality: data science enabling personalized medicine. *BMC Med.* **16**, 150 (2018).
77. Chen, R., Liu, X., Jin, S., Lin, J. & Liu, J. Machine learning for drug-target interaction prediction. *Molecules* **23**, 2208 (2018).
78. Lin, J. & Wong, K. C. Off-target predictions in CRISPR-Cas9 gene editing using deep learning. *Bioinformatics* **34**, i656–i663 (2018).
79. Chang, Y. et al. Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci. Rep.* **8**, 8857 (2018).
80. Boland, M. R., Polubriaginof, F. & Tatonetti, N. P. Development of A machine learning algorithm to classify drugs of unknown fetal effect. *Sci. Rep.* **7**, 12839 (2017).
81. Rannals, M. D. et al. Psychiatric risk gene transcription factor 4 regulates intrinsic excitability of prefrontal neurons via repression of SCN10a and KCNQ1. *Neuron* **90**, 43–55 (2016).
82. Zang, Q. et al. In silico prediction of physicochemical properties of environmental chemicals using molecular fingerprints and machine learning. *J. Chem. Inf. Model.* **57**, 36–49 (2017).
83. Hong, H., Thakkar, S., Chen, M. & Tong, W. Development of decision forest models for prediction of drug-induced liver injury in humans using a large set of FDA-approved drugs. *Sci. Rep.* **7**, 17311 (2017).
84. Korotcov, A., Tkachenko, V., Russo, D. P. & Ekins, S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Mol. Pharm.* **14**, 4462–4475 (2017).
85. Wang, W., Kim, M. T., Sedykh, A. & Zhu, H. Developing enhanced blood-brain barrier permeability models: integrating external bio-assay data in QSAR modeling. *Pharm. Res.* **32**, 3055–3065 (2015).
86. 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.* **32**, 3604–3617 (2015).
87. Xu, C. et al. In silico prediction of chemical Ames mutagenicity. *J. Chem. Inf. Model.* **52**, 2840–2847 (2012).
88. Huang, W. et al. Prediction of human clearance based on animal data and molecular properties. *Chem. Biol. Drug Des.* **86**, 990–997 (2015).
89. Basant, N., Gupta, S. & Singh, K. P. QSAR modeling for predicting reproductive toxicity of chemicals in rats for regulatory purposes. *Toxicol. Res.* **5**, 1029–1038 (2016).
90. Alhalaweh, A. et al. Computational predictions of glass-forming ability and crystallization tendency of drug molecules. *Mol. Pharm.* **11**, 3123–3132 (2014).
91. Miller, T. H. et al. Prediction of bioconcentration factors in fish and invertebrates using machine learning. *Sci. Total Environ.* **648**, 80–89 (2019).
92. Rose, S., Bergquist, S. L. & Layton, T. J. Computational health economics for identification of unprofitable health care enrollees. *Biostatistics* **18**, 682–694 (2017).
93. Calderon, C. P., Daniels, A. L. & Randolph, T. W. Deep convolutional neural network analysis of flow imaging microscopy data to classify subvisible particles in protein formulations. *J. Pharm. Sci.* **107**, 999–1008 (2018).
94. Degardin, K., Guillemain, A., Guerreiro, N. V. & Roggo, Y. Near infrared spectroscopy for counterfeit detection using a large database of pharmaceutical tablets. *J. Pharm. Biomed. Anal.* **128**, 89–97 (2016).
95. Page, D. et al. Identifying adverse drug events by relational learning. *Proc. Conf. AAAI Artif. Intell.* **2012**, 790–793 (2012).

Acknowledgements

In memory of Rebecca J. Williams. J. Freundlich, R. J. G. Arnold, P. Madrid, J. Lage de Siqueira-Neto, A. Williams, A. Tropsha, A. Gerlach, J. Gerlach, D. Chipman, A. Davidow and M. Hupcey are kindly acknowledged for discussions and some of the collaborations described herein. S.E. acknowledges funding to Collaborations Pharmaceuticals, Inc., from NIGMS R44 GM122196-02A1, NINDS 1R43NS107079-01, NINDS 3R43NS107079-01S1, NCATS 1UH2TR002084-01 and FY2018 UNC Research Opportunities Initiative (ROI) award. Research reported in this publication was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under award number R43NS107079. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests

S.E. is founder and CEO, A.C.P., K.M.Z., T.L. and J.J.K. are employees, and D.P.R. and A.M.C. are consultants of Collaborations Pharmaceuticals, Inc. A.M.C. is also the founder and owner of Molecular Materials Informatics, Inc. A.J.H. has no conflicts of interest.

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to S.E.