Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Biology

Recent advances in drug repurposing using machine learning

Fabio Urbina, Ana C. Puhl and Sean Ekins

Abstract

Drug repurposing aims to find new uses for already existing and approved drugs. We now provide a brief overview of recent developments in drug repurposing using machine learning alongside other computational approaches for comparison. We also highlight several applications for cancer using kinase inhibitors, Alzheimer's disease as well as COVID-19.

Addresses

Collaborations Pharmaceuticals, Inc., 840 Main Campus Drive, Lab 3510, Raleigh, NC, 27606, USA

Corresponding author: Ekins, Sean (sean@collaborationspharma.com)

Current Opinion in Chemical Biology 2021, **65**:74–84

This review comes from a themed issue on **Machine Learning in Chemical Biology**

Edited by **Connor W. Coley** and **Xiao Wang**

For a complete overview see the Issue and the Editorial

Available online 16 July 2021

https://doi.org/10.1016/j.cbpa.2021.06.001

Keywords

Machine learning, Drug repurposing, Drug discovery.

Abbreviations

Aβ, amyloid beta; AD, Alzheimer's disease; BACE1, β-site APP-cleaving enzyme 1; CMap, connectivity map; DNN, deep neural networks; FDA, Food and Drug Administration; GSK3β, glycogen synthase kinase 3β.

Introduction

Bringing a new small molecule drug to the clinic is a lengthy and costly process that has been widely described [1] taking anywhere from over a decade and may range from over $700 million to $2.7 billion by recent estimates [2]. Naturally, any advancements in technology which could accelerate the process or decrease the cost significantly would be a true paradigm shift. Drug repurposing is a simple concept with a long history which promises a great deal in this respect [3]. Drug repurposing aims to develop a new use for a drug beyond its original purpose and has been applied more widely as Food and Drug Administration (FDA)-approved drug libraries became more commercially accessible, allowing academics and small companies to run medium to high-throughput screens on thousands of molecules [4]. In several cases, new uses for existing drugs were found and some candidates have progressed to the clinic [5–7]. Drug companies also started to pay more attention to repositioning compounds that perhaps were either on the shelf or failed for another indication. Interestingly, this view of repurposing molecules is not new, a bibliometric analysis of PubMed showed that, since at least the 1940s, many molecules have been tested against a large number of diseases. For example, 189 drugs were linked with over 300 diseases [3]. More recently, one approach that has been swiftly gaining traction over the past 5–10 years is computational drug repurposing. This method became especially relevant in 2020 as scientists attempted to find a quick solution for treating COVID-19, and drug repurposing offered a potentially faster path to treatment [8]. Several computational tools, such as similarity searching, docking, pharmacophores, and machine learning can be used for small molecule drug discovery. We will focus primarily on machine learning in this review which can encompass a wide array of data resources and computational approaches (Figure 1). In our view, this requires several steps such as identification and curation of dataset/s, building and validation (internal and preferably external) of the machine learning models, scoring and prioritization of a drug library (or clinical assets), before final validation of these predictions *in vitro* and/or *in vivo* (see graphical abstract). Previous reviews by us or others have described some of the earlier computational repurposing successes, data sources available to enable this [9,10] as well as several of the opportunities and limitations of repurposing industry-provided clinical stage molecules [11]. As will be seen, machine learning applications can still be quite broad in how they are implemented, whether with small molecule, genomics, image assay data or combinations of these or other data types (Figure 1). We will also describe several areas where there have been significant repurposing efforts, beginning with COVID-19, kinases for cancer and Alzheimer's disease (AD) as well as for other diseases.

COVID-19 repurposing

Since the outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the late 2019, there have been hundreds of papers and likely many more preprints describing computational approaches or short communications proposing compounds with likely activity

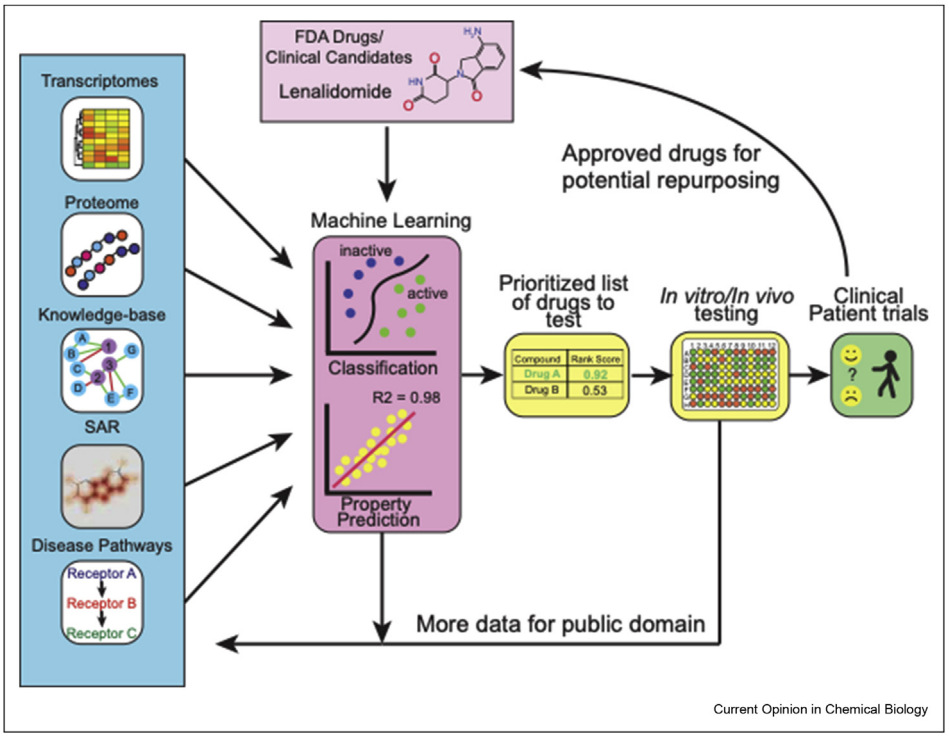
against SARS-CoV-2 [8]. The year 2020 brought drug repurposing a level of recognition that was unimaginable before the COVID-19 epidemic. As researchers were generally isolated from their laboratories, this led to many using computational tools to propose compounds for possible testing against the virus. To see thousands of researchers turning their attention to the COVID-19 problem and the flood of papers that followed, the expectation is that a drug will surely follow. In the case of treatments for COVID-19, however, while there have been a few small successes, and many clinical trials ongoing, a drug candidate has yet to emerge from all this computational repurposing. A number of causes likely contributed to the lack of success stories. Although researchers used computational tools to propose compounds for testing against the virus, the lack of BSL3 laboratory availability (during 2020) often meant little to no experimental data existed to back up their proposals other than some computational score or a docked molecule in a protein (and sometimes both). This has perhaps also sown more confusion as the general public is unable to discern the difference between a computational prediction and one that is experimentally verified. In very few of these computational-derived examples, compounds were tested *in vitro* against a target protein or in virus-infected whole cells. There were also many examples of studies in which a small number of drugs or other molecules were selected for various reasons (not computationally) and then tested [12], and these have tended to outnumber the computational repurposing papers. Although not all of the following approaches use machine learning, several other computational approaches will also be described.

One of the earliest molecules to be computationally repurposed using a knowledge graph approach was the AAK1 and JAK1/2 kinase inhibitor baricitinib [13] (Table 1). The mechanism was eventually validated *in vitro* and in human patients [14,15] as well as in combination with hydroxychloroquine [16]. As the outbreak progressed there were also examples of high-throughput screens against SARS-CoV-2 targets or in infected cells as well [17–22]. A cheminformatics and protein interaction map identified FDA and clinical stage compounds binding to sigma-1 and 2 receptors which act as host factors for COVID-19. The most potent compound identified in Vero cells was PB28 (concentration of drug required for 90% inhibition (IC_{90}) 280 nM), which is not an approved drug, while several approved drugs such as clemastine and cloperastine were also identified (Table 1), albeit with much weaker μM activity[23]. To some extent the use and efficacy of atypical antipsychotics (likely targeting sigma receptors) was also shown using real-world data analysis of hospitalized SARS-CoV-2 patients [24]. For all the clinical failures associated with hydroxychloroquine [25,26], it has still served as a starting point for computational repurposing. In one example it was used as a template for ligand-based virtual screening of 3981 molecules that had obtained regulatory approval worldwide. Amodiaquine (antimalarial) was the most active *in vitro* in Vero EG cells along with zuclopenthixol (typical antipsychotic) and nebivolol (beta blocker) [27]. Docking FDA drugs in the M^{pro} structure has identified dipyridamole (IC_{50} 0.53 μM) and this showed activity in Vero cells (EC_{50} ~0.1 μM), subsequently it was taken to a small clinical trial with eight patients and demonstrated positive responses [28]. There have been many other examples of computational docking in a target followed by assessment of binding to the target or assay of activity, but in most cases the activity of infected cells is not described [29,30]. Another approach has been to repurpose molecules that were previously computationally identified using machine learning for another virus then applying them to SARS-CoV-2. This is an approach we have taken and in due course demonstrated that the Ebola inhibitors tilorone, quinacrine and pyronaridine may also be sub-μM inhibitors of SARS-CoV-2-infected A549-ACE2 cells [12]. As yet there has not been a 'home-run' that has derived from the repurposing approach for SARS-CoV-2 although it has intensified the interest in kinase inhibitors as potential antivirals.

Kinase repurposing and cancer applications

The human genome encodes about 634 kinases (pseudokinases included), which are enzymes that transfer a polar phosphate group from adenosine triphosphate (ATP) to regulatory serine, threonine, or tyrosine residues on proteins and have a role in hundreds of diseases [31,32]. To date, the FDA has approved 62 small molecule protein kinase inhibitors which have a market in the tens of billions of dollars per year [31,32]. Most of these are delivered orally apart from temsirolimus (IV) and netarsudil (eye drops). As of 2019, 11 of them inhibited protein-serine/threonine protein kinases, 2 are dual-specificity protein kinases, 11 target non-receptor protein-tyrosine kinases, and 28 block receptor protein-tyrosine kinases. Forty-six are used in the treatment of cancers (e.g., leukaemias, breast and lung cancers), whereas eight are for non-malignancies such as myelofibrosis, polycythaemia vera, chronic immune thrombocytopenia, rheumatoid arthritis, renal graft versus host disease, idiopathic pulmonary fibrosis, glaucoma, Crohn disease and ulcerative colitis [31,32]. There are many hundreds of kinases, so this creates challenges to ensure selectivity depending on the target or the degree of polypharmacology that is permissible. The wealth of publicly available data on this target class also represents an opportunity to leverage to design and develop new classes of compounds as well as repurposing existing kinase inhibitors for other kinases not tested. The market for anticancer drugs is considerable [33]. Commercial interest in kinase inhibitors is at an

Figure 1



Overview of computational machine learning approaches for drug repurposing. SAR, structure activity relationship.

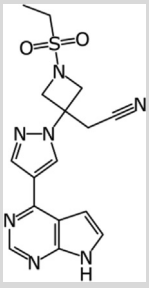
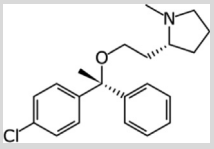
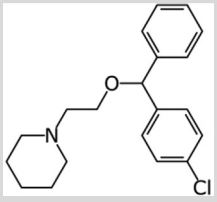
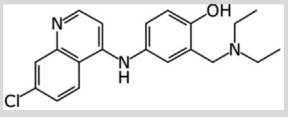
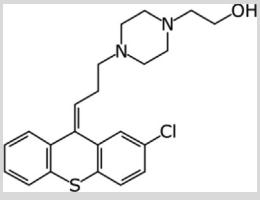
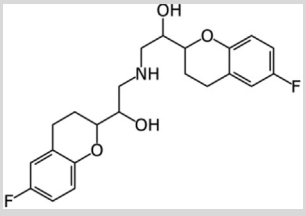
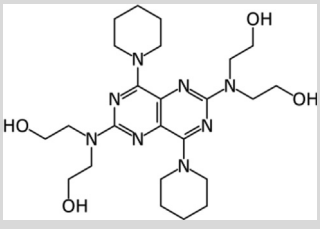
all-time high in recent years with Merck acquiring ArQule for $2.7 billion for a single clinical stage candidate (ARQ531 a BTK inhibitor) and acquiring Peloton Therapeutics for $2.2 billion. Pfizer bought Array Biopharma for $10.6 billion for their pipeline which included two marketed kinase inhibitors, and in early 2019 Eli Lilly bought Loxo Oncology for $8 billion which has several kinase inhibitors in their pipeline. There still appears a significant commercial opportunity in this target area.

Kinases are therefore a very heavily studied class of proteins in cancer and yet a large part of the kinome is 'dark' with no drugs or probes [34]. There is hence considerable interest in accelerating the identification of molecules that might be probes or potential drugs for proteins belonging to the protein kinase families that are currently unannotated. There have been notable efforts to create public structure activity datasets such as the Published Kinase Inhibitor Set as a resource for developing new probes for kinases [35]. Illuminating the Druggable Genome program began in 2014 with the goal of developing techniques and information about protein targets for this protein class as well as G-protein-coupled receptor, ion channels and nuclear receptor classes of proteins which had not been characterized or for which molecules or biologics had not been identified [36–38]. The number of proteins that are listed on the Illuminating the Druggable Genome website as of February 2021 is 328 (162 are kinases [39,40]). Currently, there are only 18 previously dark kinases for which chemical probes have been identified. In parallel with this effort, another NIH common fund project called the Library of Integrated Network-based Cellular Signatures consortium has generated public data which relate to how cells respond to various genetic or environmental stressors with results obtained in cultured and primary cells whose state has been perturbed by a 'perturbagen' [41]. These data collected by 15 institutions have been made available to the community through many different tools and workflows [41] which enables various different types of evaluations [42].

For example, one of the datasets and tools in the Library of Integrated Network-based Cellular Signatures is the connectivity map (CMap) which brings together data on genes, thousands of drugs and disease states. The original basis for the CMap was 164 drugs and tool compounds tested in 3 cell lines with microarray data that were subsequently used in several repurposing projects to identify new uses for old drugs [43–45]. The approach was

Table 1

Examples of computationally derived molecules from drug repurposing for SARS-CoV-2 [97].

Structure	Drug	Method	*In vitro*/*in vivo* activity	Reference
	Baricitinib	Knowledge graph	Significantly reduced viral load in human liver spheroids at 400 and 800 nM	[13–16]
	Clemastine	Protein interaction map	pIC50 5.67	[23]
	Cloperastine	Protein interaction map	pIC50 5.27	[23]
	Amodiaquine	Ligand-based virtual screen	EC_{50} 0.13 and 5.4 μM	[27]
	Zuclopenthixol	Ligand-based virtual screen	EC_{50}1.35 and 15 μM	[27]
	Nebivolol	Ligand-based virtual screen	EC_{50}2.72 and 2.8 μM	[27]
	Dipyridamole	Docking in M^{pro}	IC_{50} 0.53 μM, EC_{50} ~0.1 μM	[28]

pIC_{50}, negative log of the IC_{50} value when converted to molar; EC_{50}, half maximal effective concentration; IC_{50} , half maximal inhibitory concentration.

recently expanded with a reduced expression profile limited to 1000 landmark genes and tested with ~19,000 small molecules (as well as thousands of short hairpin ribonucleic acid (shRNAs), cDNAs, and biologics) in 6 cancer cell lines. A connectivity score was developed with three confidence measures [46]. Such a tool provides the capability to predict the mechanism of action of molecules based on similarity profiles of compounds of known function. In total, 1902 of these compounds were mapped to known protein targets. When using an external set of compounds, 76% could be connected in the correct perturbagen class. Some drugs were also found to associate to multiple targets for which they were similarly active. CMap was used to discover the mechanism of action for new small molecules and this was demonstrated for kinase inhibition with several drugs. Working in the opposite direction, to discover a molecule specific for another protein kinase was also demonstrated by querying the database for signatures of what knockdown of the target would look like and then find molecules sharing a similar signature [46]. Such studies have inspired others to use compound-induced transcriptomic data generated from cell lines in order to predict compound activity to molecular targets. Random forest models generated with these data for a small subset of 69 genes had similar results to or outperformed Morgan fingerprints [47]. The same group has also demonstrated how the transcriptomic data can also be used with a generative adversarial network to design molecules that address desired targets (although it should be mentioned that these two studies were theoretical and lacked experimental verification using molecule synthesis and testing) [48].

It can be summarized that de-orphaning kinases is a very slow process involving biochemical, enzymatic and cellular profiling often followed by assessment of phosphorylation, cellular target engagement and structural analyses which may occasionally result in a single molecule illuminating more than one of the dark kinases [40]. There are now many data sources relevant to kinases that could be integrated to develop computational approaches for assisting in de-orphaning dark kinases that have not been attempted or validated. In most cases the computational tools leverage molecule similarity analysis. For example, Virtual Kinome profiler was used to profile ~37 million compound−kinase pairs and predict 151,708 compounds against 248 kinases simultaneously [49]. This approach identified 19 small molecule inhibitors (including already approved drugs) of EGFR, HCK, FLT1 and MSK1 protein kinases [49]. Earlier work by the same group identified 4 new off targets (FRK, FYN A, ABL1, SLK) for tivozanib, an approved VEGF tyrosine kinase inhibitor approved to treat renal cell carcinoma [50]. Others have focused on developing the optimal kinase library for screening which made use of molecule similarity to existing small molecule kinase inhibitors [34]. Some researchers have used machine learning to map the activity of compounds across the kinome using the t-distributed stochastic neighbour embedding algorithm. Validation of this model focused on the FLT3 kinase [51]. Some of the largest kinome-wide screening efforts have combined public and proprietary data and modelled 280 kinases, with most random forest models having Area under the curve (AUCs) >0.7. The models were evaluated with leave out tests but not prospective testing [52]. Other groups have narrowed their efforts to modelling a single kinase, such as a model for JAK2 inhibitors which was used to select 13 compounds, out of which 6 had IC_{50} values less than 100 nM [53]. Different algorithms have been used including deep neural networks (DNNs) [54], Bayesian and Recursive partitioning models [55], other quantitative structure activity relationship methods [56] or deep generative models [57]. All of these computational tools could be potentially of use for drug repurposing. For example, a multitask deep learning model built with data for 391 kinases were used to find targets for clinically approved and investigational kinase drugs for which the researchers validated their predictions [58].

Most groups working on computational predictions for kinases have focused on identifying novel molecules. Some have also taken different approaches combining different data types or using whole cell or genomic data. An early study demonstrated how machine learning could be used to predict cancer cell sensitivity using genomic and chemical properties as well as proposed the potential for repurposing but without providing any examples of this [59]. A more recent version of this (although not focused on kinases) used deep learning with transcriptomic and molecular data in order to identify pimozide for repurposing for non-small cell cancer as well as tested *in vitro* (IC_{50} 19.54 μM in A549 cells) which while weak, was at a comparable activity level to gemcitabine [60]. One group developed a platform called 3D-REMAP to use ligand binding site information and protein-ligand docking to propose a cardiovascular drug levosimendan as a kinase inhibitor (RIOK1, K_d 0.82 μM as well as inhibited FLT3, MAP2K5, PIP5K1A, GAK and KIT) and potential repurposing candidate [61]. We have recently generated Bayesian machine learning models for a rare cancer called chordoma using datasets from two independent cell-based screens. These were then used to prioritize a small number of clinical stage kinase inhibitors and demonstrated that one of these AZD2014 (mTOR1 and mTOR2) possessed comparable sub-μM *in vitro* activity as the approved drug afatinib (EGFR, currently in a clinical trial for this disease) in two chordoma cell lines [62].

Machine learning approaches can also be used with multiple omics datasets to predict a therapeutic response signature which can be used to optimize patient treatment and this application was then validated with a wide

array of kinase inhibitors, both approved and clinical candidates [63]. There is therefore considerable scope for using these types of computational machine learning approaches for repurposing of pre-existing kinase inhibitors or clinical stage compounds for cancer and potentially other disease indications.

Alzheimer's disease repurposing

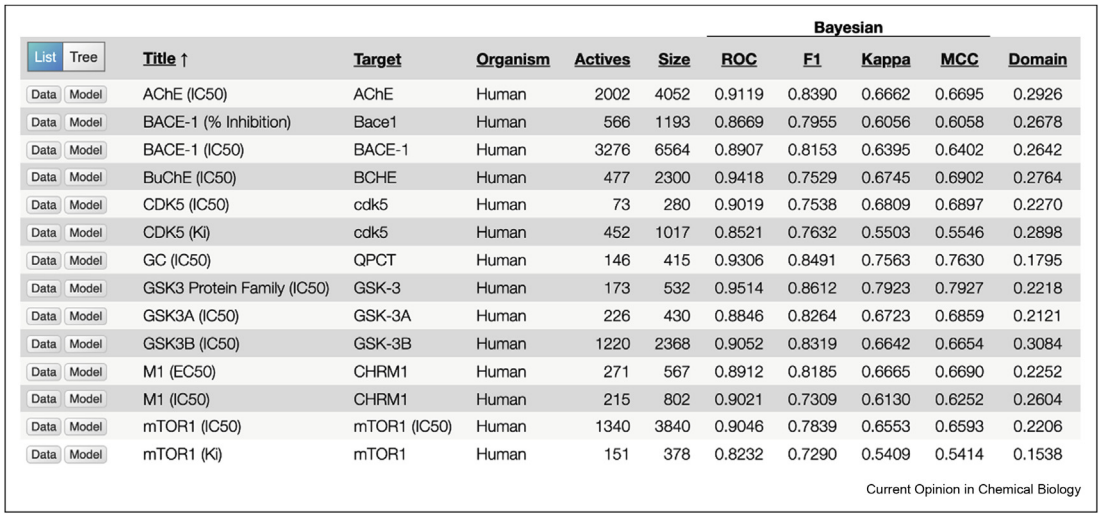
AD is widely known as the most common type of dementia, it is an irreversible neurodegenerative disease that is also the sixth leading cause of death in the United States, affects over 5.1 million adults over age 65, and costs over $150 billion per year [64]. AD is a typical age-dependent neurodegenerative disease that affects 5% of individuals >65 years, 20% of those >85 years and more than one-third of those >90 years [65]. However, approximately 200,000 individuals under age 65 have younger-onset AD, though there is greater uncertainty about the younger-onset estimate. Therefore, in the absence of proper preventative and therapeutic efforts, its prevalence will continue to increase as life expectancy increases. The mechanisms underlying the AD pathophysiology are still unclear. Aggregation of tau and amyloid beta (Aβ) proteins as well as decreased acetylcholine are hallmarks of the disease and the focus of many studies [66]. The only therapies currently on the market for treatment of AD, namely four cholinesterase inhibitors (tacrine, rivastigmine, donepezil and galantamine) and one *N*-methyl-D-aspartate receptor inhibitor (memantine), are only symptomatic and do not affect the underlying disease mechanisms or alter the disease course [64]. Therefore, drug repurposing may offer a potential avenue to identify additional clinical stage molecules.

A recent study described an approach using machine learning to quantify the pathology and AD severity as well as using neuronal cell lines treated with 80 kinase inhibitors to access the gene expression profile [67]. Some of the best scoring compounds included ruxolitinib (JAK1/2) and regorafenib (RET, VEGFRs) which were assessed for mechanism; however, experimental validation against AD was not tested *in vitro* or *in vivo* [67]. Post-mortem examination of brains with AD shows increased levels of activated Glycogen Synthase Kinase 3β (GSK3β) [68,69]. GSK3β is a proline-directed serine/threonine kinase that is involved in several intracellular signalling cascades, and is involved in a host of cellular functions, including gene transcription, glucose regulation and apoptosis [70–72]. GSK3β figures prominently in AD as the main kinase responsible for the phosphorylation of the microtubule-stabilizing protein tau [73–75]. Extracellular Aβ deposits are also connected to the activity of GSK3β because this activity is linked to the improper processing of the amyloid-precursor protein APP [76]. The β-site APP-cleaving enzyme 1 (BACE1) is responsible for the pathogenic β-cleavage of APP that results in Aβ peptides, and GSK3β inhibition downregulates expression of the BACE1 gene reducing Aβ production [77]. GSK3 inhibitors are of interest to the pharmaceutical industry for their potential use in various diseases, including diabetes, cancer, AD, Parkinson's disease, psychiatric diseases, and stem cell proliferation [78]. There are dozens of commercially available GSK3 inhibitors, of varying potency and selectivity; however, most of these inhibitors are not suitable as therapeutics due to their toxicity [79]. Only a few GSK3β inhibitors have made it to clinical trials for AD and only two of these have been tested for efficacy in treating AD, and neither one has shown great effect in treating the disease [80,81]. We have curated and validated a Bayesian machine learning model (using extended connectivity fingerprints) for inhibition of GSK3β which has over 2300 molecules (from ChEMBL) and a fivefold cross-validation receiver operating characteristic (ROC) > 0.90 [82]. This computational model was used to prioritize commercial compounds for testing. The clinical candidate ruboxistaurin was identified as an inhibitor of GSK3β and was validated using the ADP-Glo™ Kinase Assay with an IC_{50} of 39 nM. A secondary *in vitro* assay using the Z'LYTE™ Kinase Assay demonstrated an IC_{50} of 155 nM [82]. The IC_{50} for ruboxistaurin against GSK3β was higher than for the initial target PKCβ (6 nM) but lower than the closely related kinase PKCα (360 nM). Ruboxistaurin has been independently patented as an inhibitor of GSK3β for the treatment of neurological or psychiatric disorders (US 9,265,764 B2). To date, we have curated and validated Bayesian machine learning models for 10 AD targets (Figure 2). In general, the fivefold cross-validation statistics are ROC >0.8 and Matthews correlation coefficient (MCC) >0.5. There are definitely limitations as some datasets are small, but those with thousands of molecules look ideal. These computational models could then be used to find novel modulators of these proteins which could be used alone or in combination to treat AD. This type of approach could also be taken for additional AD targets to develop a suite of machine learning models that will represent structure–activity data from public sources.

Antibacterial machine learning

As has been documented many times, antibiotic drug discovery has been profoundly difficult with very low hit rates from high-throughput screening. There is also an urgent need for new drugs due to the occurrence of drug resistance. The lack of market support has also made it commercially treacherous for any companies working in this area. Computational machine learning approaches have been widely applied for antibacterial drug discovery for over a decade. We and others [83] have published extensively using the naïve Bayesian approach (with functional connectivity fingerprint descriptors) for tuberculosis drug discovery [84,85]. These models

Figure 2



List / Tree	Title ↑	Target	Organism	Actives	Size	Bayesian ROC	Bayesian F1	Bayesian Kappa	Bayesian MCC	Domain
Data Model	AChE (IC50)	AChE	Human	2002	4052	0.9119	0.8390	0.6662	0.6695	0.2926
Data Model	BACE-1 (% Inhibition)	Bace1	Human	566	1193	0.8669	0.7955	0.6056	0.6058	0.2678
Data Model	BACE-1 (IC50)	BACE-1	Human	3276	6564	0.8907	0.8153	0.6395	0.6402	0.2642
Data Model	BuChE (IC50)	BCHE	Human	477	2300	0.9418	0.7529	0.6745	0.6902	0.2764
Data Model	CDK5 (IC50)	cdk5	Human	73	280	0.9019	0.7538	0.6809	0.6897	0.2270
Data Model	CDK5 (Ki)	cdk5	Human	452	1017	0.8521	0.7632	0.5503	0.5546	0.2898
Data Model	GC (IC50)	QPCT	Human	146	415	0.9306	0.8491	0.7563	0.7630	0.1795
Data Model	GSK3 Protein Family (IC50)	GSK-3	Human	173	532	0.9514	0.8612	0.7923	0.7927	0.2218
Data Model	GSK3A (IC50)	GSK-3A	Human	226	430	0.8846	0.8264	0.6723	0.6859	0.2121
Data Model	GSK3B (IC50)	GSK-3B	Human	1220	2368	0.9052	0.8319	0.6642	0.6654	0.3084
Data Model	M1 (EC50)	CHRM1	Human	271	567	0.8912	0.8185	0.6665	0.6690	0.2252
Data Model	M1 (IC50)	CHRM1	Human	215	802	0.9021	0.7309	0.6130	0.6252	0.2604
Data Model	mTOR1 (IC50)	mTOR1 (IC50)	Human	1340	3840	0.9046	0.7839	0.6553	0.6593	0.2206
Data Model	mTOR1 (Ki)	mTOR1	Human	151	378	0.8232	0.7290	0.5409	0.5414	0.1538

Alzheimer's disease Bayesian machine learning models generated with ECFP6 descriptors and fivefold cross-validation statistics [82]. ROC, receiver operator characteristic; F1 score, is the harmonic mean of the precision and recall; kappa, cohen's kappa coefficient; MCC, matthews correlation characteristic; Domain, comparison of model training set with ChEMBL; AChE, acetylcholinesterase; BACE1, beta-secretase-1; BuChE, butyrylcholinesterase; CDK5, **c**yclin dependent kinase 5GC; QPCT, glutaminyl-peptide cyclotransferase; GSK-3, glycogen synthase kinase; M1, CHRM1, muscarinic acetylcholine receptor M_1; mTOR, mammalian target of rapamycin.

based on whole cell screening data from *Mycobacterium tuberculosis* were used to score (repurpose) molecules from the GSK antimalarial dataset [86] and from the seven compounds that were selected and tested, five had MIC $\leq$ 2 μg/mL, the most active being 0.0625 μg/mL [87]. Further examples include using 2 different whole cell models to score 3 vendor libraries and then 124 actives were found from testing 550 compounds [88], while after filtering a library of >150,000 molecules, 11 out of 48 compounds were active [89]. Overall hit rates were 15–71% for suggested compounds, much higher than the 0.6–1.5% from random library HTS [87–89]. We have also developed machine learning models with 18,886 molecules (with activity cut offs of 10 μM, 1 μM and 100 nM) [90] and used them to evaluate multiple machine learning methods (including deep learning and support vector machines). Bayesian machine learning models were on a par with DNNs with external test sets [90]. These models could be used to virtually screen FDA-approved drugs, although actual high-throughput screens were carried out over a decade ago, since then hundreds of new non-antibacterial drugs have been approved, but it is likely that many of them have never been tested against *M. tuberculosis*, and this represents a future repurposing opportunity.

A recent DNN trained on *Escherichia coli* data for 2335 molecules was used to score >107 million molecules and one of these, halicin (a repurposed *c*-Jun kinase inhibitor SU3327), had an minimum inhibitory concentration (MIC) of 2 μg/mL [91]. This molecule also showed broad-spectrum bactericidal activity as well as activity in a mouse model of *Acinetobacter baumannii* infection. After screening other large compound collections, two further molecules with broad-spectrum antibacterial activity were identified [91]. An example of a much larger library of compounds used for machine learning was 74,567 molecules tested against *E. coli* in order to generate models with gradient boosting, random forests, feed forward neural networks, support vector machines and Sammon mapping (average classification accuracy for internal testing 77.5–83.2%) [92]. These computational models were used to score a library of 5000 compounds that was also tested *in vitro*. It was not described how each method performed with this external test set, although actives that were correctly predicted were shown, of which several had MIC <1 μg/mL [92]. Although this approach was applied to identify novel antibacterials, it could also be applied for repurposing, yet this was not apparently attempted. These efforts further suggested how ligand-based machine learning approaches can search through the vast chemical space of molecules to prioritize small numbers of compounds to test and provide a higher hit rate than random screening for antibiotic drug discovery.

Applying machine learning for drug repurposing to other diseases and applications

High-throughput image-based assays used to generate 'fingerprints' have been used with DNNs on a

glucocorticoid receptor translocation dataset of over half a million compounds which was then used to predict compounds against a kinase target (50-fold enrichment) and a CNS target (289-fold enrichment) [93]. Rather than a drug repurposing application per se, this is more of a data repurposing demonstration. However, this approach could clearly be used for drug repurposing projects. The therapeutic performance mapping system was used with rat proteomic data to identify repurposed molecule combinations for nerve root avulsion [94]. This systems proteomics approach combines network maps and mathematical models and was used to select compounds for *in vitro* testing and the combination of acamprosate and ribavirin was found to be neuroprotective when tested *in vivo* as it accelerated nerve regeneration and recovery [94]. A Bayesian machine learning approach used transcriptional data and *in vitro* screening data to identify synergistic drug combinations for use in malaria by developing compound pathway signatures [95]. Thirty-five compound pairs were selected and tested (precision 83.5% and recall 65.1%) demonstrating that hydroxyzine and tacrolimus as well as raloxifene and thioridazine had high levels of synergy [95]. As a final example, a computational target prediction approach (cheminformatic similarity ensemble approach) was used to predict the activity of pharmaceutical excipients against 3117 medically relevant proteins. Those target excipient pairs with high probability were prioritized and 69 were tested *in vitro*. Nineteen were active against one of 12 targets (36% success rate) [96]. Thus, many of these molecules had relevant activities and in some cases the concentration of excipient would be at a level to have a physiological effect (e.g., cetylpyridinium chloride and the dopamine receptor 3) [96]. This could be considered an excipient repurposing application. These repurposing applications are literally scratching the surface of what may be possible with different datasets that are used to train machine learning models for various applications relevant to drug repurposing.

Discussion

Drug repurposing using machine learning has grown steadily over the past decade as we have started to see the application of DNNs and the revolution in computer power provided by graphics processing units. This has in turn led to the growth in venture capital-backed artificial intelligence drug discovery companies using purely computational approaches to discover drugs or repurpose existing drugs. As we have described herein there has been considerable use of computational approaches (including machine learning) to COVID-19, cancer, AD and antibacterial drug discovery among other areas. Although it could still be considered as nascent in terms of the development and applications, we are already starting to see some of the benefits, namely leveraging the large amounts of public data that are available in the public domain to identify new uses for these drugs that are already approved. This may offer faster identification of drugs for new disease applications in a manner that may also be more cost effective. It is likely that the use of drug repurposing using machine learning will become even more mainstream after episodes like COVID-19. What is of course key to continued viability of this approach is that the computational predictions are not used in isolation but are always backed up by convincing *in vitro* and *in vivo* verification of the predictions. In conclusion, repurposed drugs reaching patients will be the ultimate mark of success of this approach.

Funding

We kindly acknowledge National Institutes of Health funding: R44GM122196-02A1 from NIGMS, 3R43AT010585-01S1 from National Center for Complementary and Integrative Health (NCCAM), R41GM131433-01A1 from National Institute of General Medical Sciences (NIGMS) and 1R43ES031038-01 from National Institute of Environmental Health Sciences. Research reported in this publication was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Number R43ES031038. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: SE is the owner and FU and ACP are employees of Collaborations Pharmaceuticals, Inc.

Acknowledgements

The authors kindly acknowledge Dr Patricia Vignaux, Dr Thomas Lane, Ms Kimberley Zorn, Ms Eni Minerali, Mr Daniel Foil and Ms Jennifer Klein as well as other colleagues and collaborators for their assistance with some of the research on machine learning repurposing studies described by our group in recent years.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL: **How to improve R&D productivity: the pharmaceutical industry's grand challenge**. *Nat Rev Drug Discov* 2010, **9**:203–214.

2. Wouters OJ, McKee M, Luyten J: **Estimated research and development investment needed to bring a new medicine to market, 2009-2018**. *J Am Med Assoc* 2020, **323**:844–853.

3. Baker NC, Ekins S, Williams AJ, Tropsha A: **A bibliometric review of drug repurposing**. *Drug Discov Today* 2018, **23**:661–672.

4. Ekins S, Gerlach J, Zorn KM, Antonio BM, Lin Z, Gerlach A: **Repurposing approved drugs as inhibitors of Kv7.1 and**

Nav1.8 to treat pitt hopkins syndrome. *Pharm Res* 2019, **36**: 137.

5. Boucherie DM, Duarte GS, Machado T, Faustino PR, Sampaio C, Rascol O, Ferreira JJ: **Parkinson's disease drug development since 1999: a story of repurposing and relative success**. *J Parkinsons Dis* 2021, **11**:421–429.

6. Toshner M, Spiekerkoetter E, Bogaard H, Hansmann G, Nikkho S, Prins KW: **Repurposing of medications for pulmonary arterial hypertension**. *Pulm Circ* 2020, **10**. 2045894020941494.

7. Shameer K, Readhead B, Dudley JT: **Computational and experimental advances in drug repositioning for accelerated therapeutic stratification**. *Curr Top Med Chem* 2015, **15**:5–20.

8. Ekins S, Mottin M, Ramos P, Sousa BKP, Neves BJ, Foil DH, Zorn KM, Braga RC, Coffee M, Southan C, *et al.*: **Deja vu: stimulating open drug discovery for SARS-CoV-2**. *Drug Discov Today* 2020, **25**:928–941.

9. Ekins S, Williams AJ, Krasowski MD, Freundlich JS: **In silico repositioning of approved drugs for rare and neglected diseases**. *Drug Discov Today* 2011, **16**:298–310.

10. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z: **A survey of current trends in computational drug repositioning**. *Briefings Bioinf* 2016, **17**:2–12.

11. Southan C, Williams AJ, Ekins S: **Challenges and recommendations for obtaining chemical structures of industry-provided repurposing candidates**. *Drug Discov Today* 2013, **18**:58–70.

12. Puhl AC, Fritch EJ, Lane TR, Tse LV, Yount BL, Sacramento CQ, Fintelman-Rodrigues N, Tavella TA, Maranhão Costa FT, Weston S, *et al.*: **Repurposing the Ebola and marburg virus inhibitors tilorone, quinacrine, and pyronaridine: in vitro activity against SARS-CoV-2 and potential mechanisms**. *ACS Omega* 2021, **6**:7454–7468.

13. * Richardson P, Griffin I, Tucker C, Smith D, Oechsle O, Phelan A, Rawling M, Savory E, Stebbing J: **Baricitinib as potential treatment for 2019-nCoV acute respiratory disease**. *Lancet* 2020, **395**:e30–e31.
This short article highlighted the repurposing of baricitinib against SARS-CoV-2 using a knowledge graph approach.

14. * Stebbing J, Krishnan V, de Bono S, Ottaviani S, Casalini G, Richardson PJ, Monteil V, Lauschke VM, Mirazimi A, Youhanna S, *et al.*: **Mechanism of baricitinib supports artificial intelligence-predicted testing in COVID-19 patients**. *EMBO Mol Med* 2020, **12**, e12697.
This short article highlighted the *in vitro* activity of baricitinib against SARS-CoV-2.

15. Lenz HJ, Richardson P, Stebbing J: **The emergence of baricitinib: a story of tortoises versus hares**. *Clin Infect Dis* 2021, **72**: 1251–1252.

16. Titanji BK, Farley MM, Mehta A, Connor-Schuler R, Moanna A, Cribbs SK, O'Shea J, DeSilva K, Chan B, Edwards A, *et al.*: **Use of baricitinib in patients with moderate to severe coronavirus disease 2019**. *Clin Infect Dis* 2021, **72**:1247–1250.

17. Coelho C, Gallo G, Campos CB, Hardy L, Wurtele M: **Biochemical screening for SARS-CoV-2 main protease inhibitors**. *PLoS One* 2020, **15**, e0240079.

18. Smith E, Davis-Gardner ME, Garcia-Ordonez RD, Nguyen TT, Hull M, Chen E, Baillargeon P, Scampavia L, Strutzenberg T, Griffin PR, *et al.*: **High-throughput screening for drugs that inhibit papain-like protease in SARS-CoV-2**. *SLAS Discov* 2020, **25**:1152–1161.

19. Rothan HA, Teoh TC: **Cell-based high-throughput screening protocol for discovering antiviral inhibitors against SARS-COV-2 main protease (3CLpro)**. *Mol Biotechnol* 2021, **63**: 240–248.

20. Gorshkov K, Chen CZ, Xu M, Carlos de la Torre J, Martinez-Sobrido L, Moran T, Zheng W: **Development of a high-throughput homogeneous AlphaLISA drug screening assay for the detection of SARS-CoV-2 nucleocapsid**. *ACS Pharmacol Transl Sci* 2020, **3**:1233–1241.

21. Chen CZ, Shinn P, Itkin Z, Eastman RT, Bostwick R, Rasmussen L, Huang R, Shen M, Hu X, Wilson KM, *et al.*: **Drug repurposing screen for compounds inhibiting the cytopathic effect of SARS-CoV-2**. *Front Pharmacol* 2020, **11**: 592737.

22. Dittmar M, Lee JS, Whig K, Segrist E, Li M, Kamalia B, Castellana L, Ayyanathan K, Cardenas-Diaz FL, Morrisey EE, *et al.*: **Drug repurposing screens reveal cell-type-specific entry pathways and FDA-approved drugs active against SARS-Cov-2**. *Cell Rep* 2021, **35**:108959.

23. ** Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, O'Meara MJ, Rezelj VV, Guo JZ, Swaney DL, *et al.*: **A SARS-CoV-2 protein interaction map reveals targets for drug repurposing**. *Nature* 2020, **583**:459–468.
An early publication describing the repurposing of molecules for SARS-CoV-2 with experimental validation.

24. ** Gordon DE, Hiatt J, Bouhaddou M, Rezelj VV, Ulferts S, Braberg H, Jureka AS, Obernier K, Guo JZ, Batra J, *et al.*: **Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms**. *Science* 2020, **370**.
An early publication describing the repurposing of molecules for SARS-CoV-2 with clinical record validation.

25. Boulware DR, Pullen MF, Bangdiwala AS, Pastick KA, Lofgren SM, Okafor EC, Skipper CP, Nascene AA, Nicol MR, Abassi M, *et al.*: **A randomized trial of hydroxychloroquine as postexposure prophylaxis for COVID-19**. *N Engl J Med* 2020, **383**:517–525.

26. Roomi S, Ullah W, Ahmed F, Farooq S, Sadiq U, Chohan A, Jafar M, Saddique M, Khanal S, Watson R, *et al.*: **Efficacy of hydroxychloroquine and tocilizumab in patients with COVID-19: single-center retrospective chart review**. *J Med Internet Res* 2020, **22**, e21758.

27. ** Bocci G, Bradfute SB, Ye C, Garcia MJ, Parvathareddy J, Reichard W, Surendranathan S, Bansal S, Bologa CG, Perkins DJ, *et al.*: **Virtual and in vitro antiviral screening revive therapeutic drugs for COVID-19**. *ACS Pharmacol Transl Sci* 2020, **3**:1278–1292.
An example of a study that used computational predictions for drugs with SARS-CoV-2 activity which was then validated *in vitro*.

28. Liu X, Li Z, Liu S, Sun J, Chen Z, Jiang M, Zhang Q, Wei Y, Wang X, Huang YY, *et al.*: **Potential therapeutic effects of dipyridamole in the severely ill patients with COVID-19**. *Acta Pharm Sin B* 2020, **10**:1205–1215.

29. Feng S, Luan X, Wang Y, Wang H, Zhang Z, Wang Y, Tian Z, Liu M, Xiao Y, Zhao Y, *et al.*: **Eltrombopag is a potential target for drug intervention in SARS-CoV-2 spike protein**. *Infect Genet Evol* 2020, **85**:104419.

30. Virdi RS, Bavisotto RV, Hopper NC, Frick DN: **Discovery of drug-like ligands for the Mac1 domain of SARS-CoV-2 Nsp3**. *bioRxiv* 2020.

31. Roskoski R. *Protein kinase inhibitors*, **2020**; 2020.

32. Roskoski Jr R: **Properties of FDA-approved small molecule protein kinase inhibitors: a 2020 update**. *Pharmacol Res* 2020, **152**:104609.

33. Shibata S, Matsushita M, Saito Y, Suzuki T: **Anticancer drug prescription patterns in Japan: future directions in cancer therapy**. *Ther Innov Regul Sci* 2018, **52**:718–723.

34. Moret N, Clark NA, Hafner M, Wang Y, Lounkine E, Medvedovic M, Wang J, Gray N, Jenkins J, Sorger PK: **Cheminformatics tools for analyzing and designing optimized small-molecule collections and libraries**. *Cell Chem Biol* 2019, **26**:765–777. e763.

35. Elkins JM, Fedele V, Szklarz M, Abdul Azeez KR, Salah E, Mikolajczyk J, Romanov S, Sepetov N, Huang XP, Roth BL, *et al.*: **Comprehensive characterization of the published kinase inhibitor set**. *Nat Biotechnol* 2016, **34**:95–103.

36. Nguyen DT, Mathias S, Bologa C, Brunak S, Fernandez N, Gaulton A, Hersey A, Holmes J, Jensen LJ, Karlsson A, *et al.*: **Pharos: collating protein information to shed light on the druggable genome**. *Nucleic Acids Res* 2017, **45**:D995–D1002.

37. Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, Gomez SM, Guha R, Hersey A, Holmes J, *et al.*: **Unexplored therapeutic opportunities in the human genome**. *Nat Rev Drug Discov* 2018, **17**:317–332.

38. Rodgers G, Austin C, Anderson J, Pawlyk A, Colvis C, Margolis R, Baker J: **Glimmers in illuminating the druggable genome**. *Nat Rev Drug Discov* 2018, **17**:301–302.

39. Berginski ME, Moret N, Liu C, Goldfarb D, Sorger PK, Gomez SM: **The Dark Kinase Knowledgebase: an online compendium of knowledge and experimental results of understudied kinases**. *Nucleic Acids Res* 2021, **49**:D529–D535.

40. Tamir TY, Drewry DH, Wells C, Major MB, Axtman AD: **PKIS deep dive yields a chemical starting point for dark kinases and a cell active BRSK2 inhibitor**. *Sci Rep* 2020, **10**:15826.

41. Keenan AB, Jenkins SL, Jagodnik KM, Koplev S, He E, Torre D, Wang Z, Dohlman AB, Silverstein MC, Lachmann A, *et al.*: **The library of integrated network-based cellular signatures NIH program: system-level cataloging of human cells response to perturbations**. *Cell Syst* 2018, **6**:13–24.

42. Niepel M, Hafner M, Mills CE, Subramanian K, Williams EH, Chung M, Gaudio B, Barrette AM, Stern AD, Hu B, *et al.*: **A multicenter study on the reproducibility of drug-response assays in mammalian cell lines**. *Cell Syst* 2019, **9**:35–48 e35.

43. Zimmer M, Lamb J, Ebert BL, Lynch M, Neil C, Schmidt E, Golub TR, Iliopoulos O: **The connectivity map links iron regulatory protein-1-mediated inhibition of hypoxia-inducible factor-2a translation to the anti-inflammatory 15-deoxy-delta12,14-prostaglandin J2**. *Cancer Res* 2010, **70**:3071–3079.

44. Lamb J: **The Connectivity Map: a new tool for biomedical research**. *Nat Rev Cancer* 2007, **7**:54–60.

45. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, *et al.*: **The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease**. *Science* 2006, **313**:1929–1935.

46. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, *et al.*: **A next generation connectivity map: L1000 platform and the first 1,000,000 profiles**. *Cell* 2017, **171**:1437–1452. e1417.

47. Baillif B, Wichard J, Mendez-Lucio O, Rouquie D: **Exploring the use of compound-induced transcriptomic data generated from cell lines to predict compound activity toward molecular targets**. *Front Chem* 2020, **8**:296.

48. Mendez-Lucio O, Baillif B, Clevert DA, Rouquie D, Wichard J: **De novo generation of hit-like molecules from gene expression signatures using artificial intelligence**. *Nat Commun* 2020, **11**:10.

49. * Ravikumar B, Timonen S, Alam Z, Parri E, Wennerberg K, Aittokallio T: **Chemogenomic analysis of the druggable kinome and its application to repositioning and lead identification studies**. *Cell Chem Biol* 2019, **26**:1608–1622. e1606.
An example of using computational approaches to repurpose kinase inhibitors.

50. Cichonska A, Ravikumar B, Parri E, Timonen S, Pahikkala T, Airola A, Wennerberg K, Rousu J, Aittokallio T: **Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors**. *PLoS Comput Biol* 2017, **13**, e1005678.

51. Janssen APA, Grimm SH, Wijdeven RHM, Lenselink EB, Neefjes J, van Boeckel CAA, van Westen GJP, van der Stelt M: **Drug discovery maps, a machine learning model that visualizes and predicts kinome-inhibitor interaction landscapes**. *J Chem Inf Model* 2019, **59**:1221–1229.

52. Merget B, Turk S, Eid S, Rippmann F, Fulle S: **Profiling prediction of kinase inhibitors: toward the virtual assay**. *J Med Chem* 2017, **60**:474–485.

53. Yang M, Tao B, Chen C, Jia W, Sun S, Zhang T, Wang X: **Machine learning models based on molecular fingerprints and an extreme gradient boosting method lead to the discovery of JAK2 inhibitors**. *J Chem Inf Model* 2019, **59**:5002–5012.

54. Xiao T, Qi X, Chen Y, Jiang Y: **Development of ligand-based big data deep neural network models for virtual screening of large compound libraries**. *Mol Inform* 2018, **37**, e1800031.

55. Fang J, Yang R, Gao L, Yang S, Pang X, Li C, He Y, Liu AL, Du GH: **Consensus models for CDK5 inhibitors in silico and their application to inhibitor discovery**. *Mol Divers* 2015, **19**:149–162.

56. Kong Y, Bender A, Yan A: **Identification of novel aurora kinase A (AURKA) inhibitors via hierarchical ligand-based virtual screening**. *J Chem Inf Model* 2018, **58**:36–47.

57. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, Terentiev VA, Polykovskiy DA, Kuznetsov MD, Asadulaev A, *et al.*: **Deep learning enables rapid identification of potent DDR1 kinase inhibitors**. *Nat Biotechnol* 2019, **37**:1038–1040.

58. Li X, Li Z, Wu X, Xiong Z, Yang T, Fu Z, Liu X, Tan X, Zhong F, Wan X, *et al.*: **Deep learning enhancing kinome-wide polypharmacology profiling: model construction and experiment validation**. *J Med Chem* 2020, **63**:8723–8737.

59. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, Saez-Rodriguez J: **Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties**. *PLoS One* 2013, **8**, e61318.

60. Li B, Dai C, Wang L, Deng H, Li Y, Guan Z, Ni H: **A novel drug repurposing approach for non-small cell lung cancer using deep learning**. *PLoS One* 2020, **15**, e0233112.

61. Lim H, He D, Qiu Y, Krawczuk P, Sun X, Xie L: **Rational discovery of dual-indication multi-target PDE/kinase inhibitor for precision anti-cancer therapy using structural systems pharmacology**. *PLoS Comput Biol* 2019, **15**, e1006619.

62. Anderson E, Havener TM, Zorn KM, Foil DH, Lane TR, Capuzzi SJ, Morris D, Hickey AJ, Drewry DH, Ekins S: **Synergistic drug combinations and machine learning for drug repurposing in chordoma**. *Sci Rep* 2020, **10**:12982.

63. Stetson LC, Pearl T, Chen Y, Barnholtz-Sloan JS: **Computational identification of multi-omic correlates of anticancer therapeutic response**. *BMC Genom* 2014, **15**(Suppl 7):S2.

64. Mielke MM, Vemuri P, Rocca WA: **Clinical epidemiology of Alzheimer's disease: assessing sex and gender differences**. *Clin Epidemiol* 2014, **6**:37–48.

65. Lee JE, Han PL: **An update of animal models of Alzheimer disease with a reevaluation of plaque depositions**. *Exp Neurobiol* 2013, **22**:84–95.

66. West S, Bhugra P: **Emerging drug targets for Abeta and tau in Alzheimer's disease: a systematic review**. *Br J Clin Pharmacol* 2015, **80**:221–234.

67. * Rodriguez S, Hug C, Todorov P, Moret N, Boswell SA, Evans K, Zhou G, Johnson NT, Hyman BT, Sorger PK, *et al.*: **Machine learning identifies candidates for drug repurposing in Alzheimer's disease**. *Nat Commun* 2021, **12**:1033.
An early application of machine learning for Alzheimer's disease.

68. Leroy K, Yilmaz Z, Brion JP: **Increased level of active GSK-3beta in Alzheimer's disease and accumulation in argyrophilic grains and in neurones at different stages of neurofibrillary degeneration**. *Neuropathol Appl Neurobiol* 2007, **33**:43–55.

69. Pei JJ, Tanaka T, Tung YC, Braak E, Iqbal K, Grundke-Iqbal I: **Distribution, levels, and activity of glycogen synthase kinase-3 in the Alzheimer disease brain**. *J Neuropathol Exp Neurol* 1997, **56**:70–78.

70. Zhang Y, Huang NQ, Yan F, Jin H, Zhou SY, Shi JS, Jin F: **Diabetes mellitus and Alzheimer's disease: GSK-3beta as a potential link**. *Behav Brain Res* 2018, **339**:57–65.

71. Beurel E, Jope RS: **The paradoxical pro- and anti-apoptotic actions of GSK3 in the intrinsic and extrinsic apoptosis signaling pathways**. *Prog Neurobiol* 2006, **79**:173–189.

72. Turenne GA, Price BD: **Glycogen synthase kinase3 beta phosphorylates serine 33 of p53 and activates p53's transcriptional activity**. *BMC Cell Biol* 2001, **2**:12.

73. Ishiguro K, Omori A, Takamatsu M, Sato K, Arioka M, Uchida T, Imahori K: **Phosphorylation sites on tau by tau protein kinase I, a bovine derived kinase generating an epitope of paired helical filaments**. *Neurosci Lett* 1992, **148**:202–206.

74. Hanger DP, Hughes K, Woodgett JR, Brion JP, Anderton BH: **Glycogen synthase kinase-3 induces Alzheimer's disease-like phosphorylation of tau: generation of paired helical filament epitopes and neuronal localisation of the kinase**. *Neurosci Lett* 1992, **147**:58–62.

75. Lovestone S, Reynolds CH, Latimer D, Davis DR, Anderton BH, Gallo JM, Hanger D, Mulot S, Marquardt B, Stabel S, *et al.*: **Alzheimer's disease-like phosphorylation of the microtubule-associated protein tau by glycogen synthase kinase-3 in transfected mammalian cells**. *Curr Biol* 1994, **4**:1077–1086.

76. Uemura K, Kuzuya A, Shimozono Y, Aoyagi N, Ando K, Shimohama S, Kinoshita A: **GSK3beta activity modifies the localization and function of presenilin 1**. *J Biol Chem* 2007, **282**:15823–15832.

77. Ly PT, Wu Y, Zou H, Wang R, Zhou W, Kinoshita A, Zhang M, Yang Y, Cai F, Woodgett J, *et al.*: **Inhibition of GSK3beta-mediated BACE1 expression reduces Alzheimer-associated phenotypes**. *J Clin Invest* 2013, **123**:224–235.

78. Maqbool M, Hoda N: **GSK3 inhibitors in the therapeutic development of diabetes, cancer and neurodegeneration: past, present and future**. *Curr Pharm Des* 2017, **23**:4332–4350.

79. Walz A, Ugolkov A, Chandra S, Kozikowski A, Carneiro BA, O'Halloran TV, Giles FJ, Billadeau DD, Mazar AP: **Molecular pathways: revisiting glycogen synthase kinase-3beta as a target for the treatment of cancer**. *Clin Cancer Res* 2017, **23**:1891–1897.

80. Matsunaga S, Fujishiro H, Takechi H: **Efficacy and safety of glycogen synthase kinase 3 inhibitors for Alzheimer's disease: a systematic review and meta-analysis**. *J Alzheimers Dis* 2019, **69**:1031–1039.

81. Kuroki H, Anraku T, Kazama A, Bilim V, Tasaki M, Schmitt D, Mazar AP, Giles FJ, Ugolkov A, Tomita Y: **9-ING-41, a small molecule inhibitor of GSK-3beta, potentiates the effects of anticancer therapeutics in bladder cancer**. *Sci Rep* 2019, **9**:19977.

82. Vignaux P, Minerali E, Foil DH, Puhl AC, Ekins S: **Machine learning for discovery of GSK3β inhibitors**. *ACS Omega* 2020, **5**:26551–26561.

83. Prathipati P, Ma NL, Keller TH: **Global Bayesian models for the prioritization of antitubercular agents**. *J Chem Inf Model* 2008, **48**:2362–2370.

84. Ekins S, Bradford J, Dole K, Spektor A, Gregory K, Blondeau D, Hohman M, Bunin B: **A collaborative database and computational models for tuberculosis drug discovery**. *Mol Biosystems* 2010, **6**:840–851.

85. Ekins S, Kaneko T, Lipinksi CA, Bradford J, Dole K, Spektor A, Gregory K, Blondeau D, Ernst S, Yang J, *et al.*: **Analysis and hit filtering of a very large library of compounds screened against Mycobacterium tuberculosis**. *Mol Biosyst* 2010, **6**:2316–2324.

86. Gamo F-J, Sanz LM, Vidal J, de Cozar C, Alvarez E, Lavandera J-L, Vanderwall DE, Green DVS, Kumar V, Hasan S, *et al.*: **Thousands of chemical starting points for antimalarial lead identification**. *Nature* 2010, **465**:305–310.

87. Ekins S, Reynolds R, Kim H, Koo M-S, Ekonomidis M, Talaue M, Paget SD, Woolhiser LK, Lenaerts AJ, Bunin BA, *et al.*: **Bayesian models leveraging bioactivity and cytotoxicity information for drug discovery**. *Chem Biol* 2013, **20**:370–378.

88. Ekins S, Reynolds RC, Franzblau SG, Wan B, Freundlich JS, Bunin BA: **Enhancing hit identification in *Mycobacterium tuberculosis* drug discovery using validated dual-event bayesian models**. *PloS One* 2013, **8**, e63240.

89. Ekins S, Casey AC, Roberts D, Parish T, Bunin BA: **Bayesian models for screening and TB mobile for target inference with *Mycobacterium tuberculosis***. *Tuberculosis* 2014, **94**:162–169.

90. Lane T, Russo DP, Zorn KM, Clark AM, Korotcov A, Tkachenko V, Reynolds RC, Perryman AL, Freundlich JS, Ekins S: **Comparing and validating machine learning models for *Mycobacterium tuberculosis* drug discovery**. *Mol Pharm* 2018, **15**:4346–4360.

91. ** Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, MacNair CR, French S, Carfrae LA, Bloom-Ackermann Z, *et al.*: **A deep learning approach to antibiotic discovery**. *Cell* 2020, **180**:688–702 e613.

An important article describing how a deep learning ligand-based model was able to identify a new use for an existing molecule.

92. Ivanenkov YA, Zhavoronkov A, Yamidanov RS, Osterman IA, Sergiev PV, Aladinskiy VA, Aladinskaya AV, Terentiev VA, Veselov MS, Ayginin AA, *et al.*: **Identification of novel antibacterials using machine learning techniques**. *Front Pharmacol* 2019, **10**:913.

93. ** Simm J, Klambauer G, Arany A, Steijaert M, Wegner JK, Gustin E, Chupakhin V, Chong YT, Vialard J, Buijnsters P, *et al.*: **Repurposing high-throughput image assays enables biological activity prediction for drug discovery**. *Cell Chem Biol* 2018, **25**:611–618. e613.

This article describes how image analysis data can be used to predict activity of compounds

94. Romeo-Guitart D, Fores J, Herrando-Grabulosa M, Valls R, Leiva-Rodriguez T, Galea E, Gonzalez-Perez F, Navarro X, Petegnief V, Bosch A, *et al.*: **Neuroprotective drug for nerve trauma revealed using artificial intelligence**. *Sci Rep* 2018, **8**:1879.

95. * KalantarMotamedi Y, Eastman RT, Guha R, Bender A: **A systematic and prospectively validated approach for identifying synergistic drug combinations against malaria**. *Malar J* 2018, **17**:160.

This article describes how a machine learning approach can be used to predict synergistic activity.

96. * Pottel J, Armstrong D, Zou L, Fekete A, Huang XP, Torosyan H, Bednarczyk D, Whitebread S, Bhhatarai B, Liang G, *et al.*: **The activities of drug inactive ingredients on biological targets**. *Science* 2020, **369**:403–413.

This article described the use of similarity ensemble approach to predict the activities of inactive ingredients in drugs which were then validated *in vitro*.

97. Muratov EN, Amaro R, Andrade CH, Brown N, Ekins S, Fourches D, Isayev O, Kozakov D, Medina-Franco JL, Merz KM, *et al.*: **A critical overview of computational approaches employed for COVID-19 drug discovery**. *Chem Soc Rev*. 2021 Jul 2, https://doi.org/10.1039/d0cs01065k. Epub ahead of print. PMID: 34212944.