Received: 15 September 2020 | Revised: 27 November 2020 | Accepted: 10 December 2020

DOI: 10.1002/wcms.1513

ADVANCED REVIEW

WIREs COMPUTATIONAL MOLECULAR SCIENCE WILEY

Machine learning, artificial intelligence, and data science breaking into drug design and neglected diseases

José Peña-Guerrero[1] | Paul A. Nguewa[1] | Alfonso T. García-Sosa[2]

[1]Department of Microbiology and Parasitology, University of Navarra, ISTUN Institute of Tropical Health, IdiSNA (Navarra Institute for Health Research), Pamplona, Navarra, Spain

[2]Institute of Chemistry, University of Tartu, Tartu, Estonia

Correspondence

Paul A. Nguewa, Department of Microbiology and Parasitology, University of Navarra, ISTUN Institute of Tropical Health, IdiSNA (Navarra Institute for Health Research), Pamplona, E-31008 Navarra, Spain.
Email: panguewa@unav.es
Alfonso T. García-Sosa, Institute of Chemistry, University of Tartu, Ravila 14a, Tartu 54011, Estonia.
Email: alfonso.tlatoani.garcia.sosa@ut.ee

Funding information

"la Caixa" Foundation, Grant/Award Number: LCF/PR/PR13/11080005; Departamento de Educación, Gobierno de Navarra; EU COST Action, Grant/Award Numbers: CA18217, CA18218; EU Project unCoVer, Grant/Award Number: DLV-101016216; Fundación Caja Navarra; Fundación Roviralta; Gobierno de Navarra-Salud, Grant/Award Number: 12/2017; Haridus- ja Teadusministeerium, Grant/Award Number: IUT34-14; Inversiones Garcilaso de la Vega; Laser Ebro; Spanish Ministry of Education, Culture, and Sport (PhD fellowship), Grant/Award Number: FPU17/03304; Ubesol

Abstract

Machine learning (ML) is becoming capable of transforming biomolecular interaction description and calculation, promising an impact on molecular and drug design, chemical biology, toxicology, among others. The first improvements can be seen from biomolecule structure prediction to chemical synthesis, molecular generation, mechanism of action elucidation, inverse design, polypharmacology, organ or issue targeting of compounds, property and multi-objective optimization. Chemical design proposals from an algorithm may be inventive and feasible. Challenges remain, with the availability, diversity, and quality of data being critical for developing useful ML models; marginal improvement seen in some cases, as well as in the interpretability, validation, and reuse of models. The ultimate aim of ML should be to facilitate options for the scientist to propose and undertake ideas and for these to proceed faster. Applications are ripe for transformative results in understudied, neglected, and rare diseases, where new data and therapies are strongly required. Progress and outlook on these themes are provided in this study.

This article is categorized under:

Structure and Mechanism > Computational Biochemistry and Biophysics
Structure and Mechanism > Molecular Structures

KEYWORDS

Artificial intelligence, drug design, data science, drug discovery, machine learning, molecular design, neglected diseases

1 | INTRODUCTION

Machine learning (ML) and artificial intelligence (AI) are being used in a large variety of fields given improvements in data science and their strong improvement over other methods in areas such as image recognition and processing. We are slowly seeing their first steps coming to fruition in medicinal and computational chemistry[1] in their new development phase. Small and large companies as well as groups in academia are deploying ML for predicting properties, reactivities, synthesis of compounds,[2] drug targets prediction,[3] compound synergy for antimalarials, crop protection,[4] among others.

Deep learning (DL) and other learning techniques with the advent of computational power have accelerated ML development in several fields including new materials[5] and computational chemistry.[6,7] ML can also play an important role in the discovery and design of compounds for rare and neglected tropical diseases (NTDs) if there is progress over other methods and if such a progress is well-applied and understood. There are several advantages and challenges that are becoming clearer in this fast-growing field.

Arguably, ML has existed in chemistry for a long time in the initial correlations between activity of substances and their substituents,[8] and correlations of reaction rates of chemicals and the physicochemical properties of their functional groups. Indeed, quantitative structure–activity relationships (QSAR), k-nearest neighbor neural networks (kNN), support vector machines (SVM), and random forest (RF) algorithms have been used in chemistry for several decades already.[8] However, DL has seen an explosion in interest in chemistry given the development of algorithms, computing power, and ability to use the growing number of chemical and biological data.[1,6,7,8]

Even the Hammett equation, an early example of QSAR, has been refined by ML by considering a larger variety of reaction and multisubstitution patterns that were limited in the original equation to a similar core and specific group and position replacements.[9] Hammett parameters ρ and σ for rate constants for benzylbromides reacting with thiols and ammonium salt decomposition were optimized by global regression in two experimental datasets, as well as in a synthetic dataset with computational activation energies of 2,400 SN_2 reactions with various nucleophiles, leaving groups (–H, –F, –Cl, –Br) and functional groups (–H, $–NO_2$, –CN, $–NH_3$, $–CH_3$). Mean absolute error rates of around 1.5 kcal/mol are reached for the methods used, but the ML methods achieve this convergence faster, though some require more parameters than the original interpretation. Kernel ridge regression required one parameter for each training point, while the Hammett model had only as many parameters as there are reactions and set of substituents. These ML equations are an improvement because even if the original formula is simpler and easier to interpret, more reactions can now be studied, an average of rates can be used, and this allows for a more general equation that can be better to use than the first incarnation.

The "imitation game" or *Turing test* describes a remote human interrogator in a fixed time frame, having to distinguish between a computer and a human solely on their replies to various questions as asked by the interrogator.[10] Through a series of these tests, a computer's success at "thinking" and not merely parroting can be measured by its probability of being misidentified as the human subject.[10] In the concept of molecular design, one could think of a substance proposed by a "computer" or ML/AI algorithm based on a design challenge proposed by a chemist or biologist and that such proposed compound is novel, interesting, accessible, non-trivial, and "human-like" in its creativity.

2 | ML MODELS AND APPROACHES

A defining character of ML is hyperparametrization and tuning of these variables in order to minimize a loss function, that is, minimize the difference between a modeled and an observable function. This loss function minimization can take the form of gradients and several options are available to achieve early or late convergence.[1]

Central to ML in chemistry are the data representation of substances, usually in text notation such as SMILES or SELFIES,[11] as well as descriptors or features that can be atom-based, molecular based, topological, voxel, and so forth. These representations and features affect the way relationships among them can be found. Classical QSAR studies would put emphasis on reducing the number of features per data row (chemical) in order not to overfit the final correlating equation. However, in ML there is less requirement to perform these feature selections and indeed feature augmentation can occur, though this may come at the risk of reducing the interpretability of ML models.

The *Naïve Bayes (NB) algorithm* is based on the Bayesian theorem assuming that for a given target value, the description of each predictor can be performed independently to other predictors. The final prediction is achieved by considering all descriptor-based properties.[12]

The *RF classifier* is built on decision trees. Each tree is independently constructed and each node is split using the best among a subset of predictors randomly chosen at the node.[13]

J48 is a decision tree algorithm that uses a tree pruning approach, which generates fewer but more easily interpretable results. The *J48* algorithm takes one attribute of the data and splits the set of samples into subsets, one for every value of the attribute. The attribute with the maximum information gain will be selected to make the decision.[14]

The *Sequential Minimization Optimization (SMO)* algorithm is widely used for training support vector machines. SMO, an iterative classifier, breaks up the quadratic programming optimization problem into smaller issues, finally solved analytically. The SMO algorithm is simple, easy to use and fast.[15]

A convolutional neural network (CNN) model is largely used for graphical data, with pixels or vector representations, for example, inputs a 3D grid of each protein−ligand complex wherein every grid point stores atom densities in order to calculate protein–ligand binding.[16]

A deep neural network (DNN) is composed of several hidden layers of neurons that adapts (learns) new conditions, and represents the type of ML when the system uses many layers of nodes to derive high-level functions from input information.[17]

If we take SMILES representations as words (the SMILES for benzene is c1ccccc1, for example), then generating new SMILES strings can be regarded as natural language generation and a recurrent neural network (RNN, typical tool for such tasks) can be employed.[18] An RNN has a temporal memory enabling the capture of long range dependencies within a message and as reviewed recently by Elton et al., RNNs are also an important architecture for molecular generation, especially in drug discovery.[18]

Boosting is a ML technique that can help improve other algorithms by giving stronger weights to correct predictions.[19] Light-gradient and distributed optimized gradient (XGBoost[20]) boosting methods are derived from these and can often be the best performers numerically in competitions by just combining other algorithms and increasing the predicted accuracy by a few fractions of hundreds or thousands of a percentile. However, it is arguable that such a minor increase in numerical performance is indeed an increase in performance if the results are virtually the same and do not allow better control of inputs.

ML can also be used to reduce dimensionality of data sets, that is, to distil the most representative features in multidimensional data such as in t-distributed stochastic neighbor embedding (t-SNE), a useful nonlinear dimensionality reduction ML technique for visualization of multidimensional data in 2D or 3D[21]]. This has advantages over other widely-used methods such as principal component analysis (PCA) because t-SNE tries to conserve the local structure of data points and their vicinities, whereas PCA (also an unsupervised data method as t-SNE) tries to conserve the global structure of data points and is highly affected by outliers. PCA is used in noise filtering, feature extractions, stock market predictions, and gene data analysis; t-SNE is used in computer security research, music analysis, cancer research, bioinformatics, and biomedical signal processing.[22]

All data science and ML techniques require large quantities of balanced, unbiased, high-quality, diverse, curated, and representative data. It is critical for ML to have large, quality datasets, and biases are evident in several widely used databases, giving rise to probably spurious relationships.[23] Benford's Law is a relationship between the first, second, and third or more significant digits of data spanning multiple orders of magnitude, and constitutes one way of statistically comparing distributions of data to test for following of natural phenomena distribution patterns.[24] It has even been used to detect fraudulent data.[24] Typical and important properties in molecular design and development also follow these distributions, as shown for log$P$, pK_a, and solubility of drugs, of National Cancer Institute (NCI) and larger sets of compounds, both experimentally determined and calculated.[24] Using these types of tests can help determine whether there is enough (diverse) data in a training set to resemble a natural distribution, and also to help guide data generation toward a natural distribution of phenomena. One way of improving reproducibility and transparency in computational chemistry calculations has been proposed by using blockchain technology.[25] A small molecular dynamics simulation was performed on a carbon monoxide molecule allowing also the immutability of calculations.[25] This would enable to trace and repeat a calculation without any interference or modification.

Active learning is defined as ML where the predictions guide the next data points to acquire, that is, the most valuable next experiment.[26] Alternative learning approaches include meta-learning,[27] transfer-learning,[28] multitask learning,[29] few-shot learning,[30] as well as generative models.[31]

For classification tasks, metrics commonly used to measure the classification results are based on the true and false, positives and negatives given their correct or incorrect predicted classification (TP, FP, TN, and FN, respectively. Performance of ML methods is usually tracked or tuned using accuracy ($\mathrm{TP} + \mathrm{TN}/[\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}]$), precision ($\mathrm{TP}/[\mathrm{TP} + \mathrm{FP}]$), recall or sensitivity or true positive rate ($\mathrm{TP}/[\mathrm{TP} + \mathrm{FN}]$), specificity or selectivity or true negative rate ($\mathrm{TN}/[\mathrm{TN} + \mathrm{FP}]$), Matthews' correlation coefficient (MCC, $(\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN})/\surd((\mathrm{TN} + \mathrm{FN})\cdot(\mathrm{TN} + \mathrm{FP})\cdot(\mathrm{TP} + \mathrm{FN})\cdot(\mathrm{TP} + \mathrm{FP}))$), F1-scores (2 * (precision + recall)/(precision * recall)), balanced accuracy (BAC, [sensitivity + specificity]/2), and area under the receiver-operator curve (ROC-AUC), among others. Metrics are not the only factor in determining the quality of a prediction. The reusability and interpretability or control over a method are as or even more important. Of course, data representations, features, algorithms, layer constructions, ML theory and implementations, are all active areas of research and development[328].[11]

Even if there is no mathematical definition of interpretability, it can be defined as the degree to which a human can understand the cause of a decision,[33] or as the degree to which a human can consistently predict a model's result.[34] In a ML model with high interpretability, it is easier for a person to understand why certain predictions or decisions were made.[35]

3 | APPLICATIONS OF ML IN BIOMOLECULAR FIELDS

True molecular design is an expensive and slow process of repeated trial and error. Design of catalysts can help in a multitude of reactions to make possible or more feasible, important and expensive conversion of substances to products. Catalyst design can also be aided by ML using gradient based optimization and alchemical transformations.[36] For example, comparison of a genetic algorithm (GA) with Monte Carlo and random sampling for optimization of core–shell nanoparticles showed that the fraction of nanoparticles found with energies within 0.25 eV of the optimal d-band reaches a point where GA finds more than twice as many particles with the desired fitness as other sampling methods (37% of fit particles after sampling 15% of space).[37]

Gómez-Bombarelli et al.[32] showed how a generative adversarial neural network (GAN) coupled with a discriminator neural network (NN) can guide the automated design of compounds based on a continuous data-driven representation of molecules. Inverse design of compounds can then be based on optimization of required properties and then synthesis (also in automated design of materials), rather than synthesizing many compounds (expensive) to then optimize their properties.[38,39] For the first study, a deep neural network was trained on hundreds of thousands of structures to build a coupled encoder, decoder, and predictor, where continuous representations of molecules automatically generated novel chemical structures by performing simple operations in the latent space, such as decoding random vectors, perturbing known chemical structures, or interpolating between molecules. Gradient-based optimization efficiently guided the search for optimized functional compounds for drug-like molecules and molecules with fewer that nine heavy atoms. Using an objective function of $5 \times \text{QED} - \text{SAS}$ (QED is a desirability measure called “Qualitative Estimate of Drug-likeness,” and SAS is a synthetic accessibility score), a starting compound with 0.65%, 3.56%, and 18.06% QED, SAS, and percentile, respectively, is transformed to an ending compound with improved values of 0.89%, 2.09%, and 98.23%.[32]

Of special interest are AI-designed compounds being coupled with automated chemical synthesis to speed the identification of novel lead compounds,[40,41] such as self-driving laboratories that iteratively design, execute, and learn from experiments in a fully autonomous loop.[42] A “mobile robot chemist” has been presented in the Bayesian optimization-led search for improved photocatalysts for hydrogen production from water,[43] where the free-roaming robot operated autonomously over 8 days, performing 688 experiments within a 10 variable experimental space, driven by a batched Bayesian search algorithm.[43] The hope is that ML and AI driven automated chemistry reduces the speed and complexity of numerous simultaneous variables and conditions can be harnessed for other chemical discovery projects.

ML has also been used for a general-purpose neural network potential (ANI) that approaches CCSD(T)/CBS accuracy on thermodynamic benchmarks by training a network to DFT data followed by transfer learning techniques retraining on gold standard QM calculations (CCSD(T)/CBS) and thus optimally scanning chemical space. The resulting potential is 10^9 times faster than CCSD(T)/CBS calculations and has comparable accuracy.[44] The neural network potential ANI-1ccX NNP has been used to study protein-ligand binding affinities[45] by representing the intramolecular forces of protein-bound drugs within molecular dynamics simulations. These potentials are reported to be capable of predicting the protein–ligand binding pose and conformational component of the absolute Gibbs energy of binding for a set of drug molecules. Molecular mechanics overestimated by a large number (circa 4.7 kcal/mol) the conformational energy for the drug erlotinib binding to its target, while the NNP predicts a more moderate number.[45] This ANI functional allows to make use of the molecular mechanic's (MM) well-characterized parameters for protein and solution while using the NN-derived potential for the intramolecular interactions, which does not need the slow and laborious (and prone to error) development for each ligand of harmonic/torsional/improper terms used in conventional force fields. These and other techniques may help in increasing accuracy in modeling studies, helping to reduce false predictions while maintaining quick calculation times. However, since the ANI potential was not developed with charged groups, these are problematic for the NN potential though they are also a source of error in conventional forcefields.

CNNs have been shown to give less prediction errors than kernel ridge regression (KRR) and RF with normalized mean average errors of 0.0494 for the CNNs compared to 0.136 for KRR and 0.239 for RF.[46] It is thus apparent that different ML techniques will be better suited to different tasks based on the type of data, molecular representation (descriptor or atom or image-based), activity landscape, and so forth.

It is also common for ML methods to not improve, or only slightly improve the results that can be obtained through other more established modeling techniques. However, there are also large steps in improvement observed in some applications using ML, such as in competitions for structural prediction, as well as protein–ligand binding prediction. For the latter, the Drug Design Grand Challenge 4 (D3R GC) found good submissions for pose prediction from docking as well as from ML (docking: 60% of all submissions in the first stage had a median pose RMSD <2.5 Å, and pose prediction results were of high accuracy across nearly all submissions for beta secretase 1 [BACE1]).[47] For affinity prediction however, many methods underperform the null model of ligands simply ranked based on molecular weight. For the 10 ML submissions, a RF regression null model built with ChEMBL data as a baseline is outperformed by many of the ML methods for enzymes. Similar performance was observed for affinity prediction for the BACE target for methods whether utilizing ML or not, though for the CatS enzyme, methods that used ML tended to perform better than those that did not.[47] For protein structure prediction, AlphaFold by DeepMind uses large amounts of genomic and protein structure database information to predict folded protein structures that have outperformed all other techniques in the Critical Assessment of Protein Structure Prediction (CASP).[48] From that data, AlphaFold DNNs predict the distances between pairs of amino acids and the angles between chemical bonds that connect those amino acids. The NN predicts a distribution of distances between every pair of residues in a protein, which are then combined into a score that estimates how accurate a proposed protein structure is. A separate neural network uses all distances in aggregate to estimate how close the proposed structure is to the right answer.[48] These scoring functions are used to search the protein landscape to find structures that matched the predictions. They trained a generative NN to invent new fragments, which were used to continually improve the score of the proposed protein structure. The second part of the method optimized scores through a gradient descent resulting in highly accurate structures. While the next best method that used contact information and sampling reported accurately 14 out of 43 domains, AlphaFold produced high-accuracy structures with template modeling scores of 0.7 or higher for 24 out of 43 free modeling domains.[48] AlphaFold2 approaches experimental accuracy for CASP14, and revolutionized the field of protein structure prediction, leading to news headlines of "solving protein folding". Accurate protein (and nucleic acid) structure prediction can be an important step in the description, interpretation, and calculation of biomolecular interactions. Biomolecular interactions are a foundation stone of intra- and intermolecular and intra- and intercellular, tissue, organ and organism communication and bioactivity.

A dataset compilation such as the QM9 dataset based on the massive GDB database of combinatorial exploration of the chemical space gives overall higher accuracy in energy prediction than PC9, an equivalent dataset with only H, C, N, O and F and up to nine heavy atoms of the PubChemQC project. However, PC9 encompasses more chemical diversity and a stronger ability to predict energies, as determined through statistics of bonding distances and chemical functions, as well as Kernel Ridge Regression, Elastic Net, and the Neural Network model provided by SchNet.[49]

ML can also be used to rationalize the pathway and mechanism of inhibitors. Molecular dynamics (MD) simulations have been used as input to a DNN that is able to identify the relationship between the functional properties of ligand-receptor complexes and the ensemble of conformations that they sample within MD trajectories.[50] ML is thus used for the ligand-specific functional mechanisms of g-protein coupled receptors (GPCRs, an important class of proteins for drug development) by the pharmacological classification of ligands bound to the 5-HT 2A and D2 subtypes of class-A GPCRs from the serotonin and dopamine families.[50] The extra cellular loop ECL1 is known to play a critical role in class-A GPCR activation, adopting ligand-dependent conformations that are important for function.[51] The N-terminus of the intracellular loop ICL3 is also critically important for GPCR activation as this site directly interacts with the G protein involved in the opening of the intracellular side of the receptor observed in active GPCRs.[52] Besides being able to discriminate between ligand classes, the dynamics of these motifs are different between the 5-HT 2A R and D2R systems. This study also shows how important motifs can be used as collective variables and be extracted from the network to help in further analysis. ML can also help in analyzing and enhancing MD simulations by helping to efficiently sample the underlying free energy surface and kinetics.[53] Such tools can improve MD methods for explaining pathways and mechanisms of action of compounds. Challenges remain toward interpretability (what did the ML model learn?), transferability (would the ML model work on another system?), and if an ML model can be used to generate Boltzmann-weighted samples that were previously unexplored (sampling challenge).[53]

Other ML methods can also give good results. A study found that convolutional neural networks (CNN), where atom properties are used instead of pixels, are more accurate than DNN for predicting quantum chemical energies.[54] 2-D CNNs and position specific scoring matrices for amino acids were also of use in bioinformatics for finding proteins with the important molecular function in transmembrane proteins of being engaged in electron transport.[55] This study found better performance than kNN, RF, LibSVM, and QuickRBF, with good values for sensitivity (80.3%), specificity (94.4%), and accuracy (92.3%), in addition to a MCC of 0.71 for an independent dataset. De-orphaning proteins, or finding functions for them, is an important area of biomedical research in order to find targets for modulation of biochemical events.

DL can also be used to detect S-sulfenylation sites from protein sequence information alone by coding them as natural language sentences comprising biological subwords and then consequentially employ them to perform classification.[56] The performance statistics with an independent dataset gave sensitivity, specificity, accuracy, MCC, and area under the curve (AUC) rates of 85.71%, 69.47%, 77.09%, 0.5554, and 0.833, respectively. These methods are expected to increase in use given the need to predict the function of proteins based on their sequence, as there are still many more proteins with sequence information determined rather than structural information.

The response of human cancer cell to drugs and their synergy has been predicted by an interpretable DL model able to stratify ER-positive breast cancer patient clinical outcomes, as well as predicted combinations improving progression-free survival in patient-derived xenograft models (overall survival median of 48.2 months, $n = 104$ with DrugCell vs. median overall survival of 33.6 months, $n = 118$ without DrugCell).[57] ML techniques have not reached clinical practice in large numbers because they have lacked interpretability and focused on monotherapies. DrugCell was trained on the responses of 1,235 tumor cell lines to 684 drugs and their combinations, showing how to construct interpretable models for predictive medicine.

4 | APPLICATIONS OF ML IN DRUG DESIGN

ML has been used for predicting target identification and validation, identification of prognostic biomarkers, and analysis of digital pathology data in clinical trials and for drug discovery.[58,59] Issues being raised are challenges in the interpretability and repeatability of ML-generated results, as well as the need for large amounts of data. Consensus scoring across different targets using ML has also been performed, showing advantages in scoring of structure-based virtual screening over previous methods,[60] most likely due to hedging results of predictions from several methods. Different docking programs perform differently over each target in the DUD-E dataset.[61] A transfer-learning approach called boosting consensus scoring (BCS) uses individual tree ensemble classifiers (one for each target) and trained by gradient boosting using active and decoy compound labels from DUD-E and the docking scores from the eight programs as feature inputs.[60] BCS outperformed both the individual programs and the traditional consensus methods, with a ROC-AUC value of 0.85 and enrichment factor at 1% (EF1) of 29 compared to the mean consensus of ROC-AUC = 0.83 and EF1 = 26. These represent a small, though statistically significant improvement.

Drug design is essentially a multiobjective search for the best balance of binding energy, solubility, bioavailability, toxicity, and so forth. These several, sometimes conflicting objectives can be solved for using a Pareto front. Global error metrics for model quality may not be predictive of discovery performance and so Pareto shell error (bands about the Pareto front) has been proposed for assessing ML methods to help judge the suitability of a model for proposing candidates (non-global, better focused on top candidate acquisition in multiobjective candidate discovery).[62] Recurrent neural networks (RNNs) can be trained as generative models for molecular structures, similar to statistical language models in natural language processing. This has been performed for de novo compound design that recovered inhibitors in two cases: for *Staphylococcus aureus* (model reproduced 14% of 6,051 hold-out test molecules), and against *Plasmodium falciparum*.[31] An approach for automated drug design is to use RNNs as SMILES generators and train them with the learning procedure called transfer learning. First, the initial model is trained on a large generic data set of molecules to learn the general syntax of SMILES, followed by fine-tuning on a smaller set of molecules coming from, for example, a lead optimization program.[63] Amabilino et al. showed that data set sizes containing at least 190 molecules are needed for effective GRU-RNN-based molecular generation using transfer learning.[63]

Boosting as a ML technique giving stronger weights to algorithms with correct predictions among an ensemble of algorithms has been performed in AL-Boost,[64] allowing to distinguish between drugs and nondrugs as well as between organ or tissue anatomical therapeutical chemical classification system (ATC) disease categories. In addition, dimensionality reduction with t-distributed Stochastic Neighbor Embedding (t-SNE) also achieved such compartmentalization of drugs/nondrugs/organs or disease categories, such as between Nervous System and Antineoplastic drugs.[64] ML of this type can help better target compounds for their specific site of action and delivery routes, which are critical to deliver a compound to its desired point of action and avoiding side-effects or toxicity.

Molecular dynamics fingerprints (MDFPs) have been used as an orthogonal descriptor for training ML models (RF, gradient tree boosting [GTB], SVM, and meta-learner classifiers) to classify small molecules into substrates and nonsubstrates of P-gp (P-glycoprotein 1, a protein involved in the efflux of substances from the cell and important contributor to drug resistance for antibiotics and cancer).[65] MDFPs were evaluated on an in-house dataset ($n = 3{,}930$) and a ChEMBL dataset ($n = 1{,}114$ compounds) and compared to commonly used 2D molecular descriptors including structure-based and property-based descriptors. ML methods trained on these descriptors were found to produce good classification models (accuracy >0.7), although not better than the models using simpler and less computationally expensive descriptors such as PropertyFP or water-based MDFPs. The most relevant features in the membrane–solute and protein–ligand MDFPs were the solvent-accessible solvent area (SASA) of the solute and Lennard-Jones energy terms and not specific behavior of the compound in the membrane or the interactions with the protein. These approaches highlight the use by ML of information content-rich data such as MD simulations, though acquiring the right (relevant) information and the degree of complexity of data are important factors.

Deep NN for QSAR applications are gaining in appeal, and these methods can in cases show better performance than traditional approaches.[66] However, the gain in performance may not be much greater to justify the additional time and hardware investment needed for these methods.[67] The Machine Learning for Pharmaceutical Discovery and Synthesis (MLPDS) consortium comprising the Massachusetts Institute of Technology (MIT) and 13 chemical and pharmaceutical companies (Merck, Sunovion, Janssen, AstraZeneca, Eli Lilly, LEO, Amgen, Pfizer, GSK, BASF, Bayer, Novartis, and WuXi) is a data-driven synthesis planning program to discuss how predictive models can be integrated into medicinal chemistry synthesis workflows.[68] Standardized metrics, shared data sets, common benchmarks, fundamental advances to representation, robustness in low-data scenarios, and generalizability were identified in order to create more robust machine-learning-based synthetic tools. Hybrid ML and expert-encoded tools may be able to present useful aspects where many of the current Computer Aided Synthesis Planning (CASP) tools are developed for planning routes using robust, reproducible chemistry. These tools not only suggest transformations that an experienced chemist could not identify but also aim to enable chemists to lighten the cognitive burden of synthesis planning. A deep generative model, generative tensorial reinforcement learning (GENTRL) was used to discover in 21days potent inhibitors of the discoidin domain receptor 1 (DDR1), a kinase target implicated in fibrosis and other diseases.[69] The most potent compound with nanomolar activity had favorable pharmacokinetic properties even if it resembled several known inhibitors.[70] There have also been reports of the first AI-created compound to enter Phase I clinical trials in a collaboration between Exscientia and Sumitomo Dainippon Pharma.[71,72] In this case, a compound was designed that acted as a long-acting 5-HT1a agonist for obsessive–compulsive disorder treatment (OCD).

Another GAN coupled with a discriminator NN has also been used to design antimicrobial peptides with the most potent having a minimum inhibitory concentration of 3.1 μg/mL, twice as strong as ampicillin.[73] GANs have also been used for creating latent space and inverse QSAR that was searched to generate novel compounds with predicted activity against the dopamine receptor type 2.[74] Fragment-based structure generation has been proposed through knowledge and use of synthetic complexity, chemically valid structures, novelty, and diversity of compounds.[75] It has been used in multiparameter optimization in de novo design based on an actor/critic model both using bidirectional long short-term memory (LSTM) networks,[76] where the AI method learns how to generate new compounds with desired properties from an initial set of lead molecules and improving them by replacing their fragments.Thus obtained were 93% of the generated molecules as chemically valid and over 33% satisfied the targeted objectives, while there were none in the initial set.[76]

Image-based prediction of properties/activities has also been proposed as a DL QSAR using transfer learning to build models for antagonists of the progesterone receptor.[77] DeepSnap-DL uses images of three dimensional structures with multiple angles as input data into DL classification by optimization of parameters and image adjustment from 3D-structures.[77] ML has also been used for predicting drug repurposing,[59,78] aiming to exploit compounds that have already passed clinical trials for safety, though they may require additional consideration of the different administration and dosing patterns for the repurposed use from those of their original indication.

Success for active learning can also be defined by the medicinal chemistry-important improvement of retrieval of novel active chemotypes[79–81] or robustness against enriching false assay positives.[82] Some small studies imply that active learning performed optimizations in a more systematic and explorative manner than other methods,[82] though larger studies are required for firmer conclusions.

When the COVID-19 pandemic struck and led to worldwide lockdowns, several groups collaborated for finding lead compounds against the SARS-CoV-2 MPro protease using AI and ML, fragments, and X-ray crystal structures from the Diamond light source.[83] These initial micromolar hits were then further optimized to nanomolar potent compounds, demonstrating the use of distributed research efforts and ML methods used in different locations for contributing to solve a pressing and urgent health need. This also highlights the neglect about concerns made for years about the probability of dangerous pandemics.

Promiscuity cliffs have also been proposed by the ML study of fragments on multiple target ligands. It can be challenging to provide well-defined rules given the large number of pan-assay interfering compounds (PAINS) and their highly variable substructures, yet ML has been used to predict PAINS compounds that are promiscuous and distinguish them from others that are mostly inactive.[84] Activity cliffs can be useful to medicinal chemists for exploring regions of chemical space and functional groups substitutions that provide "jumps" in activity.[85] However, they also affect QSAR and ML techniques, so special consideration must be used to avoid outliers, noise, and misclassifications in the underlying data, as well as using the most appropriate modeling tool for discontinuous data landscapes.[85] Protein/ligand interaction fingerprints (IFPs) were found to capture more binding mode-relevant information than atom environment fingerprints in an active learning setting.[86]

A well-established technique in drug design and discovery is virtual screening with structure-based methods such as molecular docking,[87] especially coupled to experimental validation and prediction.[88–91] Several programs and scoring functions are available for this task (predicting ligand binding poses and ranking) and they differ in their accuracy for a given protein/ligand pair.[92,93] A study presents a bespoke NN where the choice of the most adequate docking program (search method)/scoring function for complex data in the PDBBind was calculated by joining two NN, one based on a Graph Convolutional Network on voxelized representations of protein binding sites and another based on a fully-connected NN based on ligand ECFP fingerprints and 2-D descriptors.[94] The resulting NN gave the lowest average and the lowest RMSD of the generated poses, as well as the largest number of poses with a lower RMSD than the X-ray resolution of the corresponding crystal. These are valuable predictions in order to best choose the most adequate combination of search algorithm/scoring function for a given system or class of systems; in this case, the best results being obtained by a data split consisting of sampling 20% of proteins in each non-overlapping PFAM cluster.[94]

D-COID is another attempt at building a training dataset with the aim to generate highly compelling decoy complexes that are individually matched to active complexes[95] given that challenging decoys or negatives are not commonly used. An earlier well-known dataset is the DUD-E decoy compilation[61] that may include hidden bias.[96] vScreenML was trained as a general-purpose classifier for virtual screening built on the XGBoost framework. The authors state that virtual screening with their technique can go from a typical 12% hit rate to up to 10 out of 23 compounds with an inhibition constant $IC_{50} < 50$ micromolar, including one compound with $IC_{50} = 280$ nanomolar ($K_i = 173$ nM).

A typical problem in drug discovery is that there are not enough data or the data are too sparse and do not cover well chemical space. Precisely for this scenario, an optimized chemical language model (CLM) implementing a NN with long short-term memory (LTSM) with transfer learning, augmented data, and temperature sampling enabled generating SMILES of molecules at the interface of natural products and synthetic compounds.[97] Other techniques include data augmentation, basically increasing the data available by transforming and generating features (descriptors) based on the data already available. Using protein information to augment data has been shown to be able to improve DL post-processing of virtual screening results for generalization (make better predictions on protein/ligand complexes from a different distribution to the training data) by forcing to include protein/ligand information into the model.[98] A problem for virtual screening is a large proportion of false positives. Including more stringent decoys matched by molecular properties and binding conformations in an XGBoost procedure (gradient-boosted decision trees) showed notable separation of the scores assigned to active/decoy complexes along with a slight increase in MCC of 0.57.[95] Notably, almost all the predicted compounds from this method showed experimental inhibition of the target acetylcholinesterase (AChE), with several of them at nanomolar potency.

5 | ML APPLIED TO RARE, NEGLECTED, AND UNDERSTUDIED DISEASES

Given the advance in techniques and data, ML can also find good use in the discovery and design of rare and neglected disease therapeutics. Small molecules, biologics, as well as new targets are amenable to be tackled. Well-known targets can have a new approach, such as in the (non-conserved) allosteric binding sites of *Mycobacterium tuberculosis* fumarate hydratase.[99,100] Multiple target design[21,101–103] can also be well suited for finding patterns of compounds with required interactions. Multiple organ dysfunction syndrome (MODS) is commonly observed in pediatric intensive care units and sepsis is a suspected leading cause, but the full understanding of the epidemiology and outcome of MODS in children is limited by inconsistent definitions and populations studied. Three hundred children each year are affected by Diffuse intrinsic pontine glioma (DIPG). Hepatocellular carcinoma is a rare form of liver cancer responsible for 662,000 deaths each year.[104] All three are being studied through database analysis and interactions of 12,000 compounds with targets and next-generation sequencing, matching molecular and genomic characteristics of rare disease against databases revealing which compounds may act on which characteristics.[104]

Tourette's syndrome (TS) has also been approached by compiling data using functional connectivity magnetic resonance imaging to examine whole-brain functional networks in children and adults with Tourette's.[105] Diagnostic classification via multivariate classification methods showed that brain networks in childhood TS appeared "older" and brain networks in adulthood TS appeared "younger" in comparison with typically developing individuals.

Suicide is a neglected public health problem in Northern Europe with multiple causes. Suicide risk has been studied using ML and single-payer health care registry data from tens of thousands of people in Denmark,[106] where classification trees and random forests showed sex-specific differences in risk for suicide, physical health being more important to men's suicide risk than women's suicide risk. Psychiatric disorders and possibly associated medications were important to suicide risk. Diagnoses and medications measured 2 years before suicide were more important indicators of suicide risk than when measured 6 months earlier. Individuals in the top 5% of predicted suicide risk appeared to account for 32.0% of all suicide cases in men and 53.4% of all cases in women.[106]

Parents of children with a rare disease, patient groups, and foundations can now quickly form their own teams of researchers and clinicians from scratch, raise money to fund a portfolio of complementary research strategies, and then increasingly use external companies to independently replicate their findings.[107] Datasets have been compiled that are ready for use with ML, such as with Charcot–Marie Tooth 1A (CMT1A), which is caused by mutations in the gene PMP22 (first identified in 1992). Advances in this way have also been produced for the ultra-rare disease multiple sulfatase deficiency (MSD) that has no treatment, caused by mutations in the gene that encodes the human C (alpha)-formylglycine generating enzyme, sulfatase-modifying factor 1 (SUMF1); as well as for the case of the Cystic Fibrosis Foundation that is the benchmark for how a large rare disease foundation could help bring a treatment to market and see a return on investment.[107]

Prediction and associations of preterm birth and its subtypes with eicosanoid enzymatic pathways and inflammatory markers have been studied using ML: biomarkers were organized into biological groups and by enzymatic pathways and applied adaptive elastic-net and random forest to evaluate the accuracy of each group for predicting preterm birth cases[108]; adaptive elastic-net identified 5-oxoeicosatetraenoic acid, resolvin D1, 5,6-epoxy-eicsatrienoicacid, and 15-deoxy-12,14-prostaglandin J2 as the most predictive compounds for spontaneous preterm birth.[108]

Problems with fertility are rising in developed and developing countries. Fertility associated proteins can be found in bone marrow and peripheral blood, postnatal mammalian ovary and daily sperm production.[109] Fertility-GRU incorporates gated recurrent units and position-specific scoring matrix profiles to predict the function of fertility-related proteins, reducing overfitting in the data set by adding dropout layers in the DL model, and producing predictions with a cross-validation accuracy of 85.8% and an independent accuracy of 91.1%, sensitivity of 90.5%, specificity of 91.7%, and an MCC of 0.82.

5.1 | ML and drug development for NTDs: Drug discovery, drug repurposing for NTDs

NTDs are communicable diseases highly prevalent in tropical and subtropical areas in 149 countries. The World Health Organization (WHO) has identified 20 NTDs caused by bacteria, helminths and protozoa parasites or viruses, such as Buruli ulcer, dengue, dracunculiasis (guinea-worm disease), echinococcosis, foodborne trematodiasis, human African trypanosomiasis (sleeping sickness), Chagas disease (American trypanosomiasis), leishmaniasis (Cutaneous and Visceral), leprosy, lymphatic filariasis (elephantiasis), onchocerciasis (river blindness), rabies, schistosomiasis (snail fever), soil-transmitted helminthiasis (intestinal worms), taeniasis/cysticercosis (pork tapeworm), blinding trachoma, yaws,[110] and "new" NTDs (mycetoma, chromoblastomycosis and other deep mycoses, scabies and other ectoparasitic infestations, and snakebite envenoming) (Table 1).

NTDs disable and disfigure more than they kill. Disability adjusted life years (DALYs) due to NTDs are comprised of 56% by years lost due to disability (YLD) and 44% by years of life lost (YLL), if compared to 7% of YLD and 93% of YLL for malaria, as an example of a known tropical infectious disease.[127] NTDs exhibited a collective DALY burden equivalent to HIV/AIDS, tuberculosis, or malaria.[128]

TABLE 1 Impact of neglected tropical diseases (NTDs), new and continuous threats in the world

Type	Category	Disease	Etiologic agent	Impact in 2018	References
NTDs	Protozoan infections	Human African Trypanosomiasis	*Trypanosoma brucei* spp.	977 new cases	111
		Chagas disease (American Trypanosomiasis)	*Trypanosoma cruzi*	>6 million infected	112
		Leishmaniasis	*Leishmania* spp.	261,000 new cases	111
	Helminth infections	Cysticercosis/ Taeniosis	*Taenia solium*	*Taenia solium* endemic or suspected endemic in at least 76 countries	111
		Dracunculiasis	*Dracunculus medinensis*	28	113
		Echinococcus	*Echinococcus* spp.	>1 million infected	114
		Foodborne Trematodiases	*Clonorchis* spp., *Opisthorchis* spp., *Fasciola* spp. and *Paragonimus* spp.	200,000 illnesses (data from 2015)	115
		Lymphatic filariasis	*Wuchereria bancrofti* and *Brugia* spp.	120 million infected	116
		Onchocercariasis	*Onchocerca volvulus*	154 million people treated	111
		Schistosomiasis	*Schistosoma* spp.	200,000 deaths in 2000, expected to have decreased	117
		Soil-transmitted helminthiases	*Ascaris lumbricoides*, *Trichuris trichiura*, *Necator americanus*, and *Ancylostoma duodenale*	>1.5 billion people infected	118
	Bacterial infections	Buruli ulcer	*Mycobacterium ulcerans*	2708 new cases	111
		Leprosy	*Mycobacterium leprae*	208,613 new cases	111
		Trachoma	*Chlamydia trachomatis*	89 million people treated	111
		Yaws	*Treponema pallidum* subspecies *pertenue*	80,472 suspected; 888 confirmed	119
	Viral infections	Dengue	Four serotypes of dengue virus (DENV)	4.2 million reported cases in 2019	120
		Chikungunya fever	*Togaviridae* viruses	349,936 suspected and 146,914 confirmed cases (data from 2016)	121
		Rabies	*Lyssavirus*	94 deaths in 2017	111
	Fungal infections	Mycetoma, chromoblastomycosis, deep mycosis	Wide variety of common saprotrophs organisms, including bacteria and fungi	8763 (reported in 2013) most likely highly underestimated	122
	Ectoparasitic infections	Scabies	*Sarcoptes scabiei var hominis*	>200 million people	123
		Myiasis	Dipterous species larvae	464 in the last 20 years, severely underreported	124
New threats	Viral infections	Ebola	*Ebolavirus*	Ongoing outbreak in Democratic Republic of Congo, previous outbreak caused 28,610 cases	125
		COVID-19	SARS-CoV-2	16,341,920 confirmed cases of COVID-19 in July 2020	126
Continuous threats	Protozoan infections	Malaria	*Plasmodium* spp.	219 million patients (data from 2017)	111
	Viral infections	VIH/AIDS	Human immunodeficiency viruses (HIV)	3,344,000 patients (in 2019)	111
	Bacterial infections	Tuberculosis	*Mycobacterium tuberculosis*	10 million	111

Currently, the number of people needing NTD interventions is close to 2 billion. In addition, every year more than 1 billion people presently receive treatment against at least one NTD.[129] Ultimately, the focus of improving NTD control, elimination, and eradication should not only be on drug discovery, safeguarding access and delivery but also expanding efforts to develop novel vaccines and rapid diagnostics.

5.1.1 | ML in NTD drug discovery

The estimated cost of drug development rises every few years. The newer in vitro pharmacological screenings have increased costs attributed to manufacturing a medicine from $1 billion to $2.8 billion in less than 20 years.[130] All the same, the biggest challenges of drug discovery, low efficacy and toxicity, remain prevalent.[131] A novel approach to overcome these challenges would be the implementation of ML among each of the unique steps in the drug discovery pathway. Given the existence of available models, they could be used to predict the probability of a compound to be ultimately clinically viable.[131] This would allow the access of smaller companies and researchers into the field of drug development and address more varied pharmaceutical needs. In fact, there are already models that help in the prediction of interactions with the target[89,132] as well as undesirable off-target effects.[89,133]

6 | DATABASES FOR NTD DRUG DISCOVERY

Over the last decade, chemical and biological information available in public databases such as PubChem and ChEMBL[134,135] have become much more accessible and increase coverage of chemical and biological space. Nevertheless, the correct preparation and adaptation of the data including curation is required in order to adequately build and use ML models.[136]

A number of resources are available for chemistry with ML (Table 2), such as software deepchem,[137] chemoinformatics RDKit,[138] MoleculeNet (a benchmark of compounds and data for ML),[139] pytorch software,[140] KNIME (a pipelining tool),[141] and databases ChEMBL, Probes & Drugs,[142] PubChem,[143] SureChEMBL (patented ligands),[145] among others. These tools have also allowed for better reproducibility, transparency, and further use of the models, data, approaches, and projects.[145]

Regarding NTDs, there are several examples of applications of ML including the identification of four in vivo active compounds against *Trypanosoma cruzi*, the eukaryotic parasite causing the NTD Chagas disease.[176] Ekins et al. (2015) analyzed and identified more than 500 molecules with associated target information[176] to create the Collaborative Drug Discovery database.[177]

The Broad dataset (TRYPANOSOME: Broad Primary HTS to identify inhibitors of *T. cruzi* replication) has also been made available in PubChem (bioassay record: AID 2044). Furthermore, a Pathway Genome Data Base (PGDB) was generated for the identification of distinct gene products, enzymatic reactions, pathways, and metabolic compounds in the search for potential targets.

On the other hand, one of the biggest obstacles to overcome in the treatment of NTDs, including leishmaniasis, is the high toxicity and costs of the currently approved drugs along with the sometimes limited activity and high prevalence of drug resistance.[178,179] In order to avoid unwanted effects (toxicity), one option is to target specific proteins of the parasite. Using this strategy, Jamal et al. took a dataset of 292,470 available compounds[180] (bioassay record AID 1721 at PubChem[181]) and analyzed their activity as pyruvate kinase inhibitors, as well as 179 2D-molecular descriptors using four well-known classifier algorithms: Naïve Bayes (NB),[12] Random Forest (RF),[13] J48,[14] and Sequential Minimization Optimization (SMO).[15] After the evaluation of these models, in spite of the similar performance of all models, RF was estimated to be better than the others and was recognized as the best classifier offering a good classification overall.[180] For instance, the RF model showed a BCR (balanced classification rate, the average of sensitivity and specificity) of 83%, whereas the average for the other studied models was 79.3%. RF ROC were also the best, scoring 91.3 against the average of 84.7 scored by the other models.[180] The generated models, manual, and scripts are all available online[182] and may allow progress in the discovery of novel molecules with biological activities against NTDs. A similar study targeting *Schistosoma mansoni* thioredoxin glutathione reductase yielded comparable results, with the RF classifier generating the best balanced classification rate (BCR) of 80.1, and area under the curve (AUC) of 0.87. For comparison, naïve Bayes scored 65 and 0.72, respectively. Furthermore, subsequent independent docking studies showed correlation between the docking scores and the RF

TABLE 2 Databases and resources for machine learning (ML) and data science related to neglected tropical diseases (NTDs)

Name	Description	Reference
MoleculeNet	Benchmark of compounds and data for ML	139
OrphaNet	Portal for rare diseases and orphan drugs, trials, genes, databanks	147
Tox21	Toxicology database	148
EU-ToxRisk	Adverse outcome pathways, read-across, databanks	149
OECD QSAR toolbox	Chemical safety and risk assessment	150
KNIME	Pipelining tool	141
ChEMBL	Compound, assay, and bioactivity database (2 million compounds, >16 million associated biological activities)	134
Probes & Drugs	Libraries and commercial catalogs of bioactive compounds	143
PubChem	Compound, assay, reference, and bioactivity database	144
SureChEMBL	Patented ligands	145
RDKit	Chemoinformatics software	138
Deepchem	ML software	137
Pandas	Data science software	151
Scikit-learn	ML python software	152
ChEMBL-NTD	Subset of the primary screening and medicinal chemistry data for neglected diseases	153
TDR-targets	Data on targets, drugs, and target prioritization in whole genomes	154
DNDi	Collaboration projects with pharma industry for NTD treatment	155
DrugLogit	Drugs and nondrug data according to ATC classifications	64,156
MMV	Medicines for Malaria Venture	157
PathogenBox	Compounds with potential activity against parasites like schistosomiasis and trypanosomiasis (sleeping sickness)	158
PandemicResponseBox	Antiviral, antibacterial, and antifungal leads incl. Zika and Ebola	159
MalariaBox	Compounds with potential activity vs. malaria	160
TriTrypDB	Collection, management, integration and mining of genomic information and other large-scale datasets relevant to tryopanosomatic diseases	161
VEuPathDB	Collection, management, integration and mining of genomic information and other large-scale datasets relevant to infectious disease	162
NCBI-gene	Map, sequence, expression, structure, function, citation, and homology data of genes	163
PANTHER	Tool for classification of proteins (and their genes) to facilitate high-throughput analysis. Proteins have been classified according to family and subfamily, molecular function, biological process, and pathway	164
SwissADME	Tool for the computation of physicochemical descriptors as well as the prediction of ADME parameters, pharmacokinetic properties, drug-like nature and medicinal chemistry friendliness of one or multiple small molecules	165
InterPro	For the functional analysis of proteins by classifying them into families and predicting domains and important sites	166
Pfam	A large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs)	167
SMART	Allows the identification and annotation of genetically mobile domains and the analysis of domain architectures	168
Superfamily	Structural and functional annotation for all proteins and genomes	169
SSGCID	3D atomic structures of proteins and other molecules with an important biological role in human pathogens themselves, or molecules involved in host–pathogen interactions	170

171

TABLE 2 (Continued)

Name	Description	Reference
KEGG: Kyoto encyclopedia of genes and genomes	High-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets	
UniProt	Protein sequence and functional information	[172]
GeneDB	Genome annotation for species that are undergoing manual curation and refinement	[173]
RCSB PDB	Archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies	[174]
ZINC20	Commercially-available compounds for virtual screening	[175,176]

model, supporting the use of this kind of prediction model for the screening of possible active compounds from large datasets.

The most prevalent limitation of current high-throughput screening methods is related to the data produced. It can be relatively straightforward to pinpoint the exact mechanism of action in biochemical assays but compounds identified this way may not be truly active when in contact with the whole cell due to other factors, such as the inability to cross the cell membrane. On the other hand, whole cell or phenotypic screening yield no data concerning the drug target(s) or mechanism(s) of action. Recently, Santa Maria et al.[183] overcame this limitation by examining the relationships between key chemical motifs linked with bioactivity and also linked to binding to known targets. A NB model was proposed since this algorithm is considered to be less sensitive to false negatives than other reported options.[184] Furthermore, this study was able to retrospectively define the mechanism of action for the already known antibiotics against *Escherichia coli* dihydrofolate reductase (DHFR) as well as prospectively identify novel antibacterial agents targeting *Mycobacterium tuberculosis* DHFR (Rv2763c) and were confirmed through docking studies.[183]

ML approaches have also been used in the discovery of new treatments against NTDs with a high prevalence of treatment-resistant parasite strains such as malaria.[185] In 2020, Neves et al. used deep learning strategies for the construction of binary and QSAR models focusing on antiplasmodial activity and cytotoxicity of candidate molecules, as deep learning is well-suited to QSAR modeling.[186] Afterwards, the most promising compounds were filtered according to virtual screening and their properties (antiparasitic activity and toxicity) were validated in vitro against *P. falciparum* and mammalian cells. Furthermore, QSAR models were used in order to predict chemical features contributing to the antiplasmodial activity of the compounds. Thus, six fragments with favorable predicted contributions to the activity were identified and these molecules may find use in the design and optimization of better antiplasmodial compounds. Finally, three new compounds with high activity against sensitive and multidrug-resistant strains of *P. falciparum* and low toxicity in mammalian cells were also found.[185]

ML algorithms can be used not only for drug discovery but also for the improvement of existing treatments. For instance, in the case of multidrug-resistant tuberculosis, treatment with moxifloxacin is recommended though also difficult to adjust due to associated toxicity issues.[187] As levofloxacin is considered safer than moxifloxacin, RF and classification and regression tree (CART) analyses were performed to predict levofloxacin treatment regimens associated with the expected microbiologic cure taking into consideration the pharmacokinetic and pharmacodynamic profile of levofloxacin.[188]

Using ML, drug resistance and metabolic profiles as functional biomarkers may be useful for the accurate prediction of the origin of samples, combining epidemiological and genetic observations. Recently, Casimiro-Soriguer et al. applied a ML innovative approach in which functional profiles of microbiota samples obtained from shotgun sequencing were used as features for predicting the geographic origin of city metagenomics samples.[189]

The emergence of drug resistance is a concern for combating pathologies including NTDs. For example, miltefosine is used as an oral drug against Cutaneous Leishmaniasis. The exact mechanism of action of miltefosine remains unclear. This drug may cause antileishmanial activity by immuno-modulatory effects on the host macrophage. In addition, within the parasite it alters lipid metabolism and membrane lipid composition (mainly of phosphatidylcholine that constitutes more than 30% of the phospholipids on the parasite cell membrane) or DNA fragmentation. It has been reported that miltefosine transporter proteins (P-gp, an ABC transporter and P4ATPase–CDC50 complex) play a critical role in the emergence of drug-resistant *Leishmania major*. The cause of resistance is lower accumulation of drug inside the parasite. Using a ML technique, parasite-specific motifs of these proteins have been identified against which a peptide library was designed.[190] The authors aimed to allosterically modulate miltefosine transporter proteins (P-gp and P4ATPase–CDC50) in *L. major* using small peptides to reduce the drug efflux and increase its uptake to obtain the accumulation of drug within the parasite and this phenomenon may finally induce a reversal in resistance.[190]

Due to the major concern of toxicity issues in drug development and to address this, model cell lines have been traditionally used to evaluate drug cytotoxicity in mammalian cells. NB models can be applied to increase the likelihood of detecting non-cytotoxic compounds. After constructing a dataset with information derived from PubChem, cross-validation, and external tests, results might validate such a novel cytotoxicity Bayesian model as a helpful tool able to foster research in drug development. For example, two distinct substructural descriptors were underlined: FCFP_6 (functional class fingerprints of maximum diameter 6) and ECFP_6 (extended class fingerprints of maximum diameter 6). The ECFP-6 based model was significantly better than the alternative FCFP-6 based model. The ECFP-6 based model showed better ROC-AUC values (77%), specificity (59%), and concordance (68%), when compared to FCFP-6 based model (values of 73%, 31%, and 56%, respectively).[191] Interestingly, similar results were reported by Lane et al., who demonstrated that a Bayesian model using ECFP-6 descriptors at different threshold cutoffs for activity was equivalent to or outperformed other ML models (ROC = 0.90, Matthew's Correlation Coefficient [MCC] = 0.66), in this case for *Mycobacterium tuberculosis* drug discovery.[192]

6.1 | ML in drug repurposing

Computational drug repurposing has increased its utility over the past years as an alternative to the huge costs associated with traditional tools for drug development.[87,193] Nevertheless, applications of different similarity-based ML techniques for the discovery of novel drug-disease associations has been limited due to the need of preexisting data,[194] as well as concerns about bias while model building.[195] In a recent study, Guney demonstrated the drop in performance observed in similarity-based ML when none of the drugs from the training and tests set are repeated and used independently.[196] Since this phenomenon was shown to exist even for the "gold standard," it was then suggested that the usage of unsupervised systems-level drug repurposing approaches would prevent these problems.

The main purported advantages of drug repurposing in drug development are possibly reduced safety concerns and research and development costs since a repurposed drug usually has post-market surveillance data or has been characterized thoroughly during clinical development.[197] On the other hand, Genome-Wide Association Studies (GWASs) based on the genotyping of the complete genome of a huge number of individuals allow to statistically establish a relation between a genetic variation and a disease or disease-related phenotype.[198] In the field of drug repurposing, GWASs can point out potential gene targets and relate them to several diseases including NTDs. Therefore, it may offer new indications for already approved drugs. Furthermore, these studies can even highlight pleiotropic effects of genes (or how a single gene can be implicated in more than one disease with unique phenotypes).[199] Pleiotropic effects of genes are derived from the numerous and different interactions observed in biological networks.[200] Among organisms, it is common to find compensatory mechanisms in most typical signaling pathways. If a disease outcome is not the result of a single perturbation of a single gene, the treatment for such disease has a higher probability to succeed when aiming to disrupt several targets implicated in that same disease.

Consequently, in today's scientific world it is becoming increasingly necessary to add procedures able to assimilate and decipher large amounts of data. In this regard, different studies used these approaches to predict the mortality, discharge diagnosis, length of stay, and readmission into the clinic using deep learning natural language processing (NLP) methods with high accuracy.[201] Likewise, by using the DeepWalk algorithm it was possible to predict new relationships between drugs and targets based on the exploration of the built biological network.[202]

7 | CHALLENGES AND OUTLOOK FOR ML FOR NTDs

Common sensible recommendations to the growing efforts using data science and the best approach to use them have been proposed in 10 rules: (1) Establish data science as a core drug discovery discipline; (2) Engage data scientists ahead of data generation; (3) Enforce FAIR ("Findable, Accessible, Interoperable, and Reusable") play; (4) Build analytics and visualization on top of an integrated data store; (5) Connect distributed data science teams through a strong community; (6) Promote a culture of digital savvy across the organization; (7) Embrace and deploy AI without hyping it;

(8) Complement internal capabilities with strategic partnerships; (9) Allocate sufficient and appropriate resources to data science teams; (10) Invest in attracting and retaining talent.[203]

A related problem is becoming apparent in the number of studies using ML but reusing datasets that may be incomplete or biased and thus increase the difficulty to compare and assess studies or approaches.[23,96] For example, robust, simpler techniques appear to be more transparently published and reproducible in QSAR publications.[204] Moreover, ML models should be stable and not depend too greatly on the random number used, platform, or version of software implemented, but produce reasonably similar results for the same specific inputs and algorithm. In addition, it is expensive, time-consuming, and laborious to gather new activity, toxicity, or diagnosis and treatment data. Robotics and AI-guided synthesis and data generation may help with these, though there is always a need for guidance in order for the decision process not to be trivial (no new knowledge), wasteful (repeating known information), or picking up erroneous bias.

A series of proposals has been made for evaluating generative methods such as GANs or RNNs which use unsupervised ML such that a model learns from a dataset and can then produce data of a similar format: (1) Active molecules used to train the generative model should be made available in electronic form so readers may perform substructure and similarity searches and compare the output molecules with the training set. (2) Papers reporting AI-generated molecules should contain a table showing the training-set molecule most similar to each generated molecule. (3) Journals should use the same criteria for assessing the novelty of AI-generated molecules that are used to assess molecules generated by a team of medicinal chemists.[70] General criticism of ML reports include the lack of full disclosure of algorithms, data, comparison to other ML, comparison to non-ML techniques, interpretability, robustness against spurious results (learning the wrong lesson) and in the face of new data,[205] as well as validation issues such as lack of applicability domain considerations[58,206] and possible scenarios where an ML model may not be provable to solve a problem.[207] Benchmarks have also been proposed for de novo design from deep learning and neural generative models, such as GuacaMOL,[208] for evaluating single and multiobjective tasks, chemical space coverage, ability to generate models, and fidelity to reproduce the property distribution of the training sets. In addition, ML can also suffer from the same problem as other modeling methods to distinguish correlation from causation,[8,209] and on the appropriate training set cover of cases and their diversity in order to make accurate predictions for new cases. Further, tweaking older algorithms may give as good results as newly developed ones, indicating that progress may not be as large as hailed.[210]

Given that data are and will continue to remain the main cornerstone of data science, ML, and AI, it is inevitable that the quality, diversity, accessibility, interpretability, and quantity of data will determine the quality of models, ML, and AI for drug design. Growing in relevance are also data protection and ownership, which can provide as strong a protection for a new treatment as a patent,[211,212] as well as the ability to generate new data according to the distribution of already available data and to balance or guide algorithms toward more relevant, interesting, discoverable chemical space. Indeed, improvements may be more likely to be achieved by generating new data rather than small fiddling with ML algorithms.

As with all other data science and modeling, several existential points for ML/AI may be addressed, such as data sets being unavailable for comparison, or their bias, or being insufficiently large/diverse, or the train/test splits chosen being irrelevant for prospective performance.[213] In addition, prospective validation may still be insufficient to estimate model performance if not enough in number, distribution, and without human intervention, as well as if a baseline (advancement made over existing methods) is not properly chosen, optimized, or compared[213]; also, being models based on data, the quality of the latter defines the upper limit of model performance (data quality must also be reported).[213] Furthermore, the danger exists of irrelevant model endpoints or targets being chosen or not discussed in the context of the actual model application[213] (the endpoints for which there are data may not be relevant for the question or problem to be solved). This may be a problem inherited from experimental data but gets transmitted into the model building and decision making process leading to failure. Finally, there is always "survivor bias" in that retrospective evaluations may provide some guidance, though they can be limited by already known information.[213]

Indeed, COVID-19 has shown the fragility of the world to jointly confront a deadly pandemic, raising concerns about probable future pandemics such as avian flu, or others.[87] A perspective on human-animal health also appears to merit close interdisciplinary cooperation. Many calls from scientists and experts for research support and policies directed toward improving worldwide preparedness for potential (even likely) pandemics were left unheeded. There is still a high probability that another infectious pandemic occurs, including a more transmissible and/or more deadly one.[213]

Explainable AI is also making inroads, with interpretations becoming available for graphical explanation in terms of specific positive and negative functional groups or molecular motifs that are considered relevant by a NN model predicting drug interaction with cytochrome P450; specific atoms and vector weight contributions in graph convolutional methods; instance-based model interpretation; uncertainty estimations; as well as "verbal" explanations of produced steps in an ML algorithm, that is, to synthesize a sentence using natural language that explains the decision performed by the model, simultaneously training generators on large datasets of human-written explanations (though these often need to be hand-coded before-hand), among others.[214]

Also needed is a critical approach to the congruence between in vitro and animal models of disease and the corresponding human disease condition as a fundamental assumption of a lot of biomedical research.[215] If the data generated has little or incomplete relevance to the disease, this fundamental error in experiments will be carried over into data methods.

When dealing with infectious diseases, a great challenge is increasing the multidisciplinary efforts among groups allowing the further control and eradication of NTDs. Interconnectedness may contribute to the dissemination of NTDs, enhancing the infection rates, resistance to treatments, death, and disability associated with these illnesses. Currently, some of the most important challenges are drug and vaccine development and accessibility through collaborative initiatives. In addition, limitations and availability of specific diagnostic systems remain priorities.

On the one hand, confirmatory diagnosis based on specific diagnostic biomarkers will be helpful for the application of a suitable treatment against infections. NTDs are now targeted for elimination in endemic areas and an initiative of mapping their burden should be addressed using adequate tools such as serology. Diagnosis may also become a challenge for clinicians from developed countries who are presently facing an increasing number of such infections.

Furthermore, there is a growing interest to apply in silico methods to develop future treatments against these diseases. In fact, new drugs could be identified using a workflow that integrates chemoinformatic approaches, molecular modeling, and theoretical affinity calculations, as well as validated and relevant in vitro assays. From a chemical perspective, structural and molecular patterns need to be explored taking advantage of the technology to generate more efficient structure–activity-phenotype approaches and focusing on the capacity to predict and propose common targets among these diseases.

The availability of the genome sequence of several pathogens and the subsequent advances in genomic applications and in vivo luminescence-based imaging, might allow improving target-based drug discovery. We need to examine the advantages and limitations of modern tools useful to detect and validate critical genes as drug targets with genomic editing applications. Such functional genomics in combination with target validation through biochemical investigation are valuable for the identification of new drugs.

On the other hand, an interesting strategy for new drugs against NTDs is drug repositioning. Since complex biological systems are hard to model, the generation of false positives is common. This alternative needs to be improved. Moreover, obtaining models to predict ADMETox (Absorption, Distribution, Metabolism, Excretion, and Toxicology) properties has become an important challenge. In this scenario, the training and generation of predictive models could be more powerful and robust through the inclusion of inactive compounds and other "negative data."

There are also challenges regarding vaccine discovery. Indeed, limitations and drawbacks of conventional adjuvants have prompted the development of nano-carriers for vaccine delivery against infectious diseases including NTDs. It is well known that the composition of nano-carriers and mostly the physicochemical properties of nanoparticles have notable potential to overcome issues in vaccine development and to obtain desirable protection against NTDs and other infections.

Finally, during and beyond the COVID-19 pandemic, another important challenge is to foster strategies supporting the prevention and control of NTDs. We propose to maintain the nature of interdisciplinary research in relation to One Health, to promote the co-production of knowledge about One Health and zoonotic diseases, and to explore areas of potential synergies between the COVID-19 pandemic control efforts and NTD control and elimination programs.

8 | CONCLUSIONS AND FUTURE OUTLOOK

ML and AI technologies in drug discovery aim to learn complex design rules from chemists and accelerate decisions and triaging of chemicals by chemists. Therefore, properly implemented they should create a virtuous cycle where chemists can improve ML algorithms and in turn, receive better suggestions for synthesis and testing that cover wider and more relevant chemical space.

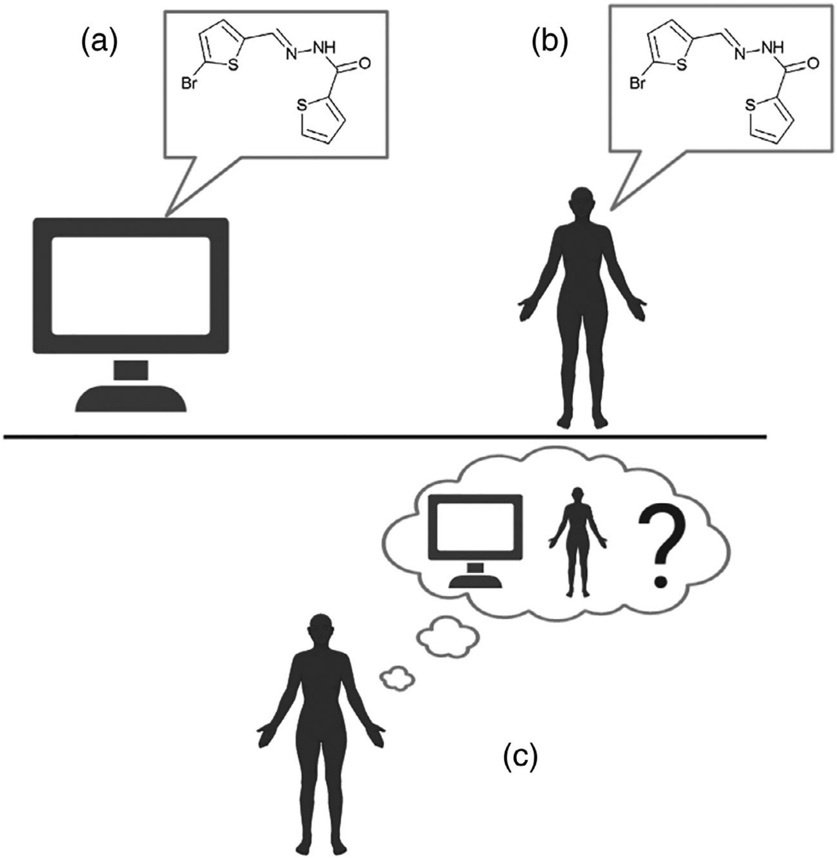

FIGURE 1 A molecular design chemical Turing test, where a machine learning (ML), artificial intelligence (AI), or data science technique (a) proposes a chemical compound that to a chemist's (c) perception and judgment is indistinguishable from another chemist's (b) proposed structure

A Molecular Design Chemical Turing Test can therefore be proposed to be reached in the near future, whereby the predictions or suggestions of compounds output from an ML are not distinguishable from those of a skilled medicinal chemist (Figure 1). This would include considerations such as novelty of the compound, synthetic accessibility, ADME/Tox and reactivity predictions, structural familiarity, chemical beauty or desirability.[216] Also, experiment suggestion and verification may be another application area of ML. Herein lays the greatest challenge for ML methods: provide the just amount of familiarity and reasonable suggestions, while still proposing novelty and design ideas in chemical space that have not been explored before and are worthy of developing through synthesis and testing.

Progress is being achieved in many areas of data science, ML, and AI for biomolecular sciences, drug discovery and design, and also in their application to NTDs. However, challenges, limitations, and avenues for future improvement are also becoming clear. In addition, advances in the field should be carefully assessed and explained to the specialist and lay public in non-sensational terms ("AI creates drug" titles are not helpful since they are misleading and raise incorrect assumptions—a drug should be an officially approved, safe, and guaranteed beneficial compound). The transparency, accessibility, reusability, and availability of procedures and data, together with clear interpretations, applicability domains, and benefit to the medical and chemical questions at hand are key to guarantee that AI and ML approaches will continue to grow and find use in these fields. A strongly active supervision of assumptions and biases in model construction and prediction must always be carried out, since AI and ML tend to very quickly pick up on these and produce erroneous and distorted results. ML, AI, and data science are here to stay and help in drug design and NTDs.

ACKNOWLEDGMENTS

A.T.G.-S. thanks Haridus- ja teadusministeerium for grant IUT34-14. P.A.N. thanks La Caixa Foundation (LCF/PR/PR13/11080005), Fundación Caja Navarra, Gobierno de Navarra-Salud (12/2017), Fundación Roviralta, Ubesol, Departamento de Educación, Gobierno de Navarra, Laser Ebro, Inversiones Garcilaso de la Vega and COST Actions CA18217 and CA18218, EU Project unCoVer (DLV-101016216) for their support. J.P.G. thanks the Spanish Ministry of Education, Culture, and Sports (PhD fellowship FPU17/03304).

AUTHOR CONTRIBUTIONS

José Peña-Guerrero: Data curation; investigation; software; writing-review and editing. **Paul Nguewa:** Data curation; formal analysis; funding acquisition; investigation; methodology; resources; supervision; writing-original draft; writing-review. **Alfonso García-Sosa:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; supervision; validation; visualization; writing-original draft; writing-review and editing.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

ORCID

Alfonso T. García-Sosa https://orcid.org/0000-0003-0542-4446

RELATED WIREs ARTICLES

Formatting biological big data for modern machine learning in drug discovery

Machine learning and artificial neural network accelerated computational discoveries in materials science

FURTHER READING

Miquel D-F, Adrià F-T, Martino B, Patrick A. Formatting biological big data for modern machine learning in drug discovery. WIREs Comput Mol Sci. 2019;9:e1408. https://doi.org/10.1002/wcms.1408.

Jingchao Z, Hengle J, Bo H, Yang H. Machine learning and artificial neural network accelerated computational discoveries in materials science. WIREs Comput Mol Sci. 2020;10:e1450. https://doi.org/10.1002/wcms.1450.

Ramsundar B, Eastman P, Walters P, Pande V. *Deep Learning for the life sciences*. Sebastopol, CA: O'Reilly, 2019;p. 297.

Ending the neglect to attain the Sustainable Development Goals—A road map for neglected tropical diseases 2021–2030. Geneva: World Health Organization; 2020. Licence: CC BY-NC-SA 3.0 IGO.

Inamuddin, Formiga FR, Severino P. *Applications of nanobiotechnology for neglected tropical diseases*. 1st ed. Cambridge: Academic Press, 2021. Paperback ISBN: 9780128211007.

Investing to overcome the global impact of neglected tropical diseases: third WHO report on neglected diseases 2015. 1. Tropical Medicine—trends. 2. Neglected Diseases. 3. Poverty Areas. 4. Universal Coverage—economics. 5. Developing Countries. 6. Annual Reports. I. World Health Organization. ISBN 978 92 4 156486 1.

REFERENCES

1. Schneider P, Walters WP, Plowright AT, et al. Rethinking drug design in the artificial intelligence era. *Nat Rev Drug Discov Nat Res*. 2020;*19*:353–364.
2. Sandfort F, Strieth-Kalthoff F, Kühnemund M, Beecks C, Glorius F. A structure-based platform for predicting chemical reactivity. *Chem*. 2020;*6*(6):1379–1390.
3. Mayr A, Klambauer G, Unterthiner T, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem Sci*. 2018;*9*(24):5441–5451.
4. Mason DJ, Eastman RT, Lewis RPI, Stott IP, Guha R, Bender A. Using machine learning to predict synergistic antimalarial compound combinations with novel structures. *Front Pharmacol*. 2018;*9*(OCT):1096.
5. Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. *npj Comput Mater*. 2019;*5*:1–36.
6. Choo K, Mezzacapo A, Carleo G. Fermionic neural-network states for ab-initio electronic structure. *Nat Commun*. 2020;*11*(1):2368.
7. Cole DJ, Mones L, Csányi G. A machine learning based intramolecular potential for a flexible organic molecule. *Faraday Discuss*. 2020;*224*:247–264.
8. Muratov EN, Bajorath J, Sheridan RP, et al. QSAR without borders. *Chem Soc Rev*. 2020;*49*(11):3525–3564.
9. Bragato M, von Rudorff GF, von Lilienfeld OA. Data enhanced Hammett-equation: reaction barriers in chemical space. *Chem Sci*. 2020;*11*(43):11859–11868.
10. Turing test | Definition & Facts | Britannica [Internet]. https://www.britannica.com/technology/Turing-test. Accessed 13 Nov 2020.
11. Krenn M, Hase F, Nigam A, Friederich P, Aspuru-Guzik A. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. *Mach Learn Sci Technol*. 2020;*1*(4):045024.
12. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Mach Learn*. 1997;*29*:131–163.

13. Breiman L. Random Forests. *Machine Learning*. 2001;*45*(1):5–32. http://dx.doi.org/10.1023/a:1010933404324.
14. Quinlan J. *C4.5. Programs for machine learning*. San Mateo, CA: Morgan Kaufman; 2014.
15. Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995;*203*:273–297. http://dx.doi.org/10.1007/bf00994018.
16. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein–ligand scoring with convolutional neural networks. *J Chem Inf Model*. 2017;*57*(4):942–957.
17. Anastasia Kyrykovych L. Deep neural networks [Internet]. https://www.kdnuggets.com/2020/02/deep-neural-networks.html. Accessed 13 Nov 2020.
18. Elton DC, Boukouvalas Z, Fuge MD, Chung PW. Deep learning for molecular design—a review of the state of the art. *Mol Syst Des Eng*. 2019;*4*(4):828–849.
19. Zhou Z-H. *Ensemble methods: foundations and algorithms*. Boca Raton: Chapman & Hall/CRC, 2012;p. 236.
20. XGBoost Documentation—xgboost 1.3.0-SNAPSHOT documentation [Internet]. https://xgboost.readthedocs.io/en/latest/. Accessed 13 Nov 2020.
21. Liu H, Wang L, Lv M, et al. AlzPlatform: an Alzheimer's disease domain-specific chemogenomics knowledgebase for polypharmacology and target identification research. *J Chem Inf Model*. 2014;*54*(4):1050–1060.
22. Difference between PCA VS t-SNE—GeeksforGeeks [Internet]. https://www.geeksforgeeks.org/difference-between-pca-vs-t-sne/. Accessed 13 Nov 2020.
23. Yang J, Shen C, Huang N. Predicting or pretending: artificial intelligence for protein–ligand interactions lack of sufficiently large and unbiased datasets. *Front Pharmacol*. 2020;*11*:69.
24. García-Sosa AT. Benford's law in medicinal chemistry: implications for drug design. *Future Med Chem*. 2019;*11*(17):2247–2253. https://doi.org/10.4155/fmc-2019-0006.
25. Hanson-Heine MWD, Ashmore AP. Computational chemistry experiments performed directly on a blockchain virtual computer. *Chem Sci*. 2020;*11*(18):4644–4647.
26. Reker D. *Practical considerations for active machine learning in drug discovery*. Drug Discovery Today: Technologies. 2020. http://dx.doi.org/10.1016/j.ddtec.2020.06.001.
27. Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: 34th International Conference on Machine Learning, ICML 2017; 2017. p. 1856–1868.
28. Pan SJ, Yang QA. Survey on transfer learning. *IEEE Trans Knowl Data Eng*. 2009;*22*(10):1345–1359.
29. Unterthiner T, Mayr A, Unter Klambauer G, et al. Deep learning as an opportunity in virtual screening. *Proc Deep Learn Work NIPS*. 2014;*27*:1–9.
30. Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. *Adv Neural Inf Process Syst*. 2017;*30*:4077–4087.
31. Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent Sci*. 2018;*4*(1):120–131.
32. Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent Sci*. 2018;*4*(2):268–276.
33. Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artif Intell*. 2019;*267*:1–38.
34. Kim B, Khanna R, Koyejo O. Examples are not enough, learn to criticize! Criticism for interpretability. *Adv Neural Inf Process Syst*. 2016;*29*:2280–2288.
35. Chapter 2. Interpretability | Interpretable machine learning [Internet]. https://christophm.github.io/interpretable-ml-book/interpretability.html. Accessed 13 Nov 2020.
36. Freeze JG, Kelly HR, Batista VS. Search for catalysts by inverse design: Artificial intelligence, mountain climbers, and alchemists. *Chem Rev*. 2019;*119*:6595–6612.
37. Froemming NS, Henkelman G. Optimizing core–shell nanoparticle catalysts with a genetic algorithm. *J Chem Phys*. 2009;*131*(23):234103.
38. Coley CW, Eyke NS, Jensen KF. Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angewandte Chemie International Edition*. 2020;*59*(51):22858–22893.
39. Coley CW, Eyke NS, Jensen KF. Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angewandte Chemie International Edition*. 2020;*59*(52):23414–23436.
40. Chen H, Engkvist O. Has drug design augmented by artificial intelligence become a reality? *Trends Pharmacol Sci*. 2019;*40*:806–809.
41. Friederich P, Dos Passos Gomes G, De Bin R, Aspuru-Guzik A, Balcells D. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chem Sci*. 2020;*11*(18):4584–4601.
42. MacLeod BP, Parlane FGL, Morrissey TD, et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci Adv*. 2020;*6*(20):eaaz8867.
43. Burger B, Maffettone PM, Gusev VV, et al. A mobile robotic chemist. *Nature*. 2020;*583*(7815):237–241.
44. Smith JS, Nebgen BT, Zubatyuk R, et al. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat Commun*. 2019;*10*(1):2903.
45. Lahey SLJ, Rowley CN. Simulating protein–ligand binding with neural network potentials. *Chem Sci*. 2020;*11*(9):2362–2368.
46. Faber FA, Hutchison L, Huang B, et al. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J Chem Theory Comput*. 2017;*13*(11):5255–5264.

47. Parks CD, Gaieb Z, Chiu M, et al. D3R grand challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J Comput Aided Mol Des*. 2020;*34*(2):99–119.
48. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;*577*(7792): 706–710.
49. Glavatskikh M, Leguy J, Hunault G, Cauchy T, Da Mota B. Dataset's chemical diversity limits the generalizability of machine learning predictions. *J Chem*. 2019;*11*(1):1–15.
50. Plante A, Shore DM, Morra G, Khelashvili G, Weinstein HA. Machine learning approach for the discovery of ligand-specific functional mechanisms of GPCRs. *Molecules*. 2019;*24*(11):2097
51. Wheatley M, Wootten D, Conner MT, et al. Lifting the lid on GPCRs: The role of extracellular loops. *Br J Pharmacol*. 2012;*165*: 1688–1703.
52. Katritch V, Cherezov V, Stevens RC. Structure-function of the G protein-coupled receptor superfamily. *Annu Rev Pharmacol Toxicol Annu Rev*. 2013;*53*:531–556.
53. Wang Y, Lamim Ribeiro JM, Tiwary P. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Curr Opin Struct Biol*. 2020;*61*:139–145.
54. Faber FA, Hutchison L, Huang B, et al. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J Chem Theory Comput*. 2017;*13*(11):5255–5264.
55. Le N-Q-K, Ho Q-T, Ou Y-Y. Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins. *J Comput Chem*. 2017;*38*(23):2000–2006.
56. Do DT, Quynh T, Le T, Quoc N, Le K. Using deep neural networks and biological subwords to detect protein S-sulfenylation sites. *Brief Bioinform*. 2020;*2020*:1–11.
57. Kuenzi BM, Park J, Fong SH, et al. Predicting drug response and synergy using a deep learning model of human Cancer cells. *Cancer Cell*. 2020;*38*(5):672–684.e6.
58. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov*. 2019;*18*:463–477.
59. Myszczynska MA, Ojamies PN, Lacoste AMB, et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat Rev Neurol Nat Res*. 2020;*16*:440–456.
60. Ericksen SS, Wu H, Zhang H, et al. Machine learning consensus scoring improves performance across targets in structure-based virtual screening. *J Chem Inf Model*. 2017;*57*(7):1579–1590.
61. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem*. 2012;*55*(14):6582–6594.
62. Del Rosario Z, Rupp M, Kim Y, Antono E, Ling J. Assessing the frontier: active learning, model accuracy, and multi-objective candidate discovery and optimization. *J Chem Phys*. 2020;*153*(2):024112.
63. Amabilino S, Pogány P, Pickett SD, Green DVS. Guidelines for recurrent neural network transfer learning-based molecular generation of focused libraries. *J Chem Inf Model*. 2020;*60*:5699–5713.
64. Yosipof A, Guedes RC, García-Sosa AT. Data mining and machine learning models for predicting drug likeness and their disease or organ category. *Frontiers in Chemistry*. 2018;*6*:162. http://dx.doi.org/10.3389/fchem.2018.00162.
65. Esposito C, Wang S, Lange UEW, Oellien F, Riniker S. Combining machine learning and molecular dynamics to predict P-glycoprotein substrates. *J Chem Inf Model*. 2020;*60*:4730–4749.
66. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. *Drug Discov Today*. 2018;*23*(6):1241–1250.
67. Zhang J, Mucs D, Norinder U, Svensson F. LightGBM: an effective and scalable algorithm for prediction of chemical toxicity-application to the Tox21 and mutagenicity aata sets. *J Chem Inf Model*. 2019;*59*(10):4150–4158.
68. Struble TJ, Alvarez JC, Brown SP, et al. Current and future roles of artificial intelligence in medicinal chemistry synthesis. *J Med Chem*. 2020;*63*(16):8667–8682.
69. Zhavoronkov A, Ivanenkov YA, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol*. 2019;*37*(9):1038–1040.
70. Walters WP, Murcko M. Assessing the impact of generative AI on medicinal chemistry. *Nat Biotechnol*. 2020;*38*:143–145.
71. Wakefield J. Artificial intelligence-created medicine to be used on humans for first time. BBC News [Internet]; 2020. https://www.bbc.com/news/technology-51315462. Accessed 09 Sep 2020.
72. Smith J. Exscientia's first AI-designed drug enters phase I to treat OCD [Internet]; 2020. https://www.labiotech.eu/ai/exscientia-ocd-ai-sumitomo/. Accessed 09 Sep 2020.
73. Tucs A, Tran DP, Yumoto A, Ito Y, Uzawa T, Tsuda K. Generating ampicillin-level antimicrobial peptides with activity-aware generative adversarial networks. *ACS Omega*. 2020;*5*:22847–22851.
74. Blaschke T, Olivecrona M, Engkvist O, Bajorath J, Chen H. Application of generative autoencoder in de novo molecular design. *Mol Inform*. 2018;*37*(1):1700123
75. Polishchuk P. CReM: chemically reasonable mutations framework for structure generation. *J Chem*. 2020;*12*(28):1–18.
76. Ståhl N, Falkman G, Karlsson A, Mathiason G, Boström J. Deep reinforcement learning for multiparameter optimization in de novo drug design. *J Chem Inf Model*. 2019;*59*(7):3166–3176.
77. Matsuzaka Y, Uesawa Y. DeepSnap-deep learning approach predicts progesterone receptor antagonist activity with high performance. *Front Bioeng Biotechnol*. 2019;*7*:485.

78. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem Rev*. 2019; *119*:10520–10594.
79. Desai B, Dixon K, Farrant E, et al. Rapid discovery of a novel series of Abl kinase inhibitors by application of an integrated microfluidic synthesis and screening platform. *J Med Chem*. 2013;*56*(7):3033–3047.
80. Reker D, Schneider P, Schneider G. Multi-objective active machine learning rapidly improves structure–activity models and reveals new protein–protein interaction inhibitors. *Chem Sci*. 2016;*7*(6):3919–3927.
81. Fujiwara Y, Yamashita Y, Osoda T, et al. Virtual screening system for finding structurally diverse hits by active learning. *J Chem Inf Model*. 2008;*48*(4):930–940.
82. Duros V, Grizou J, Xuan W, et al. Human versus robots in the discovery and crystallization of gigantic Polyoxometalates. *Angew Chem Int Ed*. 2017;*56*(36):10815–10820.
83. Diamond Light Source. Main protease structure and XChem fragment screen. Diamond Light Source [Internet]; 2020. https://www.diamond.ac.uk/covid-19/for-scientists/Main-protease-structure-and-XChem.html. Accessed 09 Sep 2020.
84. Jasial S, Gilberg E, Blaschke T, Bajorath J. Machine learning distinguishes with high accuracy between pan-assay interference compounds that are promiscuous or represent dark chemical matter. *J Med Chem*. 2018;*61*(22):10255–10264.
85. Cruz-Monteagudo M, Medina-Franco JL, Pérez-Castillo Y, Nicolotti O, Cordeiro MNDS, Borges F. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discov Today*. 2014;*19*:1069–1080.
86. Rodríguez-Pérez R, Miljković F, Bajorath J. Assessing the information content of structural and protein–ligand interaction representations for the classification of kinase inhibitor binding modes via machine learning and active learning. *J Chem*. 2020;*12*:36.
87. García-Sosa AT, Sild S, Maran U. Docking and virtual screening using distributed grid technology. *QSAR Comb Sci*. 2009;*28*(8):815–821. https://doi.org/10.1002/qsar.200810174.
88. Lyu J, Wang S, Balius TE, et al. Ultra-large library docking for discovering new chemotypes. *Nature*. 2019;*566*(7743):224–229.
89. Viira B, Selyutina A, García-Sosa AT, et al. Design, discovery, modelling, synthesis, and biological evaluation of novel and small, low toxicity s-triazine derivatives as HIV-1 non-nucleoside reverse transcriptase inhibitors. *Bioorganic Med Chem*. 2016;*24*(11):2519–2529. https://doi.org/10.1016/j.bmc.2016.04.018.
90. Glisic S, Sencanski M, Perovic V, Stevanovic S, García-Sosa AT. Arginase flavonoid anti-leishmanial in silico inhibitors flagged against anti-targets. *Molecules*. 2016;*21*(5):589. http://dx.doi.org/10.3390/molecules21050589.
91. Stevanovic S, Sencanski M, Danel M, et al. Synthesis, in silico, and in vitro evaluation of anti-Leishmanial activity of oxadiazoles and indolizine containing compounds flagged against anti-targets. *Molecules*. 2019;*24*(7):1282. https://doi.org/10.3390/molecules24071282.
92. García-Sosa AT, Hetényi C, Maran U. Drug efficiency indices for improvement of molecular docking scoring functions. *J Comput Chem*. 2010;*31*(1):174–184. https://doi.org/10.1002/jcc.21306.
93. García-Sosa AT, Mancera RL, Dean PM. WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein–ligand complexes. *J Mol Model*. 2003;*9*(3):172–182. http://dx.doi.org/10.1007/s00894-003-0129-x.
94. Jiménez-Luna J, Cuzzolin A, Bolcato G, Sturlese M, Moro S. A deep-learning approach toward rational molecular docking protocol selection. *Molecules*. 2020;*25*(11):2487.
95. Adeshina YO, Deeds EJ, Karanicolas J. Machine learning classification can reduce false positives in structure-based virtual screening. *Proceedings of the National Academy of Sciences*. 2020;*117*31:18477–18488.
96. Chen L, Cruz A, Ramsey S, et al. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS One*. 2019;*14*(8):e0220113.
97. Moret M, Friedrich L, Grisoni F, Merk D, Schneider G. Generative molecular design in low data regimes. *Nat Mach Intell*. 2020;*2*(3): 171–180.
98. Scantlebury J, Brown N, Von DF. BioRxiv CD-, 2020 U. Dataset augmentation allows Deep learning-based virtual screening to better generalize to unseen target classes, and highlight important binding interactions. *J Chem Inf Model*. 2020;*60*(8):3722–3730.
99. Vanden Eynde JJ, Mangoni AA, Rautio J, et al. Breakthroughs in medicinal chemistry: new targets and mechanisms, new drugs, new hopes-6. *Molecules*. 2020;*25*(1):119
100. Whitehouse AJ, Libardo MDJ, Kasbekar M, et al. Targeting of fumarate hydratase from *Mycobacterium tuberculosis* using allosteric inhibitors with a dimeric-binding mode. *J Med Chem*. 2019;*62*(23):10586–10604.
101. Alcaro S, Bolognesi ML, García-Sosa AT, Rapposelli S. Editorial: Multi-target-directed ligands (MTDL) as challenging research tools in drug discovery: from design to pharmacological evaluation. *Front Chem*. 2019;*7*:71. https://doi.org/10.3389/fchem.2019.00071.
102. Garcia-Sosa AT. Designing ligands for leishmania, plasmodium, and aspergillus N-myristoyl transferase with specificity and anti-target-safe virtual libraries. *Curr Comput Aided Drug Des*. 2018;*14*(2):131–141.
103. Mizuno S, Iijima R, Ogishima S, et al. AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease. *BMC Syst Biol*. 2012;*6*:52.
104. Michigan state university investigator receives $2.1m to study existing treatments for select rare diseases [Internet]. TrialSiteNews; 2019. https://www.trialsitenews.com/michigan-state-university-investigator-receives-2-1m-to-study-existing-treatments-for-select-rare-diseases/. Accessed 09 Sep 2020.
105. Nielsen AN, Gratton C, Church JA, et al. Atypical functional connectivity in Tourette syndrome differs between children and adults. *Biol Psychiatry*. 2020;*87*(2):164–173.

106. Gradus JL, Rosellini AJ, Horváth-Puhó E, et al. Prediction of sex-specific suicide risk using machine learning and single-payer health care registry data from Denmark. *JAMA Psychiat*. 2020;*77*(1):25–34.
107. Ekins S, Perlstein EO. Doing it all—how families are reshaping rare disease research. *Pharmaceutical research*. Volume 35. New York: . Springer, 2018.
108. Aung MT, Yu Y, Ferguson KK, et al. Prediction and associations of preterm birth and its subtypes with eicosanoid enzymatic pathways and inflammatory markers. *Sci Rep*. 2019;*9*(1):17049.
109. Le NQK. Fertility-GRU: Identifying fertility-related proteins by incorporating deep-gated recurrent units and original position-specific scoring matrix profiles. *J Proteome Res*. 2019;*18*(9):3503–3511.
110. Word Health Organization. Working to overcome the global impact of neglected tropical diseases. First WHO report on neglected tropical diseases; 2010.
111. Vardell E. Global health observatory data repository. *Med Ref Serv Q*. 2020;*39*(1):67–74.
112. World Health Organization. Chagas disease: Fact sheet [Internet]. Vol. 304, Geneve: Technical Report Series; 2019. p. 1–4. https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis). Accessed 29 Jul 2020.
113. World Health Organization. Dracunculiasis (guinea-worm disease) fact sheet [Internet]. World Health Organization; 2020. https://www.who.int/news-room/fact-sheets/detail/dracunculiasis-(guinea-worm-disease). Accessed 29 Jul 2020.
114. World Health Organization. Echinococcosis fact sheet [Internet]. World Health Organization; 2020. https://www.who.int/news-room/fact-sheets/detail/echinococcosis. Accessed 29 Jul 2020.
115. World Health Organization. Foodborne trematodiases [Internet]. Fact Sheet; 2016. p. 6–11. https://www.who.int/news-room/fact-sheets/detail/foodborne-trematodiases. Accessed 29 Jul 2020.
116. World Health Organization. WHO: lymphatic filariasis epidemiology [Internet]. World Health Organization; 2018. http://www.who.int/lymphatic_filariasis/epidemiology/en/. Accessed 29 Jul 2020.
117. World Health Organization. Schistosomiasis, Fact sheet February 2016 [Internet]. World Health Organization (WHO); 2016. p. 1–5. https://www.who.int/news-room/fact-sheets/detail/schistosomiasis. Accessed 29 Jul 2020.
118. World Health Organization Media Centre. Soil-transmitted helminth infections. Fact sheet N°366 [Internet]. Fact Sheet; 2014. https://www.who.int/news-room/fact-sheets/detail/soil-transmitted-helminth-infections. Accessed 29 Jul 2020.
119. World Health Organization. Yaws: fact sheets. *World Heal Organ*. 2019.
120. World Health Organization. Media centre: dengue and severe dengue fact sheet. *World Heal Organ*. 2016;1–7.
121. World Health Organization. WHO chikungunya fact sheet [Internet]. World Health Organization Media Centre; 2015. p. 1–2. https://www.who.int/news-room/fact-sheets/detail/chikungunya. Accessed 29 Jul 2020.
122. van de Sande WWJ. Global burden of human mycetoma: a systematic review and meta-analysis. *PLoS Negl Trop Dis*. 2013;*7*(11):e2550.
123. World Health Organization. WHO | Scabies and other ectoparasites [Internet]. World Health Organization; 2020. http://www.who.int/neglected_diseases/diseases/scabies-and-other-ectoparasites/en/. Accessed 29 Jul 2020.
124. Bernhardt V, Finkelmeier F, Verhoff MA, Amendt J. Myiasis in humans—a global case report evaluation and literature analysis. *Parasitology research*. 2019;*118*:389–397.
125. World Health Organization. Ebola virus disease: fact sheet No. 103 [Internet]; 2015. http://www.who.int/mediacentre/factsheets/fs103/en/. Accessed 29 Jul 2020. Media Centre. https://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease.
126. World Health Organization. WHO coronavirus disease (COVID-19) dashboard [Internet]. WHO; 2020. https://covid19.who.int/. Accessed 29 Jul 2020.
127. Fitzpatrick C, Nwankwo U, Lenk E, de Vlas SJ, Bundy DAP. An investment case for ending neglected tropical diseases. *Disease control priorities, 3rd ed. (Vol. 6): Major infectious diseases*. Washington: The World Bank, 2017; p. 411–431.
128. Hotez PJ, Molyneux DH, Fenwick A, et al. Control of neglected tropical diseases. *N Engl J Med*. 2007;*357*:1018–1027.
129. Word Health Organization. Global Health Observatory (GHO) data. Neglected tropical diseases.
130. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J Health Econ*. 2016;*47*:20–33.
131. Ekins S, Puhl AC, Zorn KM, et al. Exploiting machine learning for end-to-end drug discovery and development. *Nat Mater*. 2019;*18*:435–441.
132. Reker D, Rodrigues T, Schneider P, Schneider G. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc Natl Acad Sci U S A*. 2014;*111*(11):4067–4072.
133. Lampa S, Alvarsson J, McShane SA, Berg A, Ahlberg E, Spjuth O. Predicting off-target binding profiles with confidence using conformal prediction. *Front Pharmacol*. 2018;*9*:1256
134. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*. 2012;*40*D1:D1100–D1107.
135. Kim S, Thiessen PA, Bolton EE, et al. PubChem substance and compound databases. *Nucleic Acids Res*. 2015;*44*(D):D1202–D1213.
136. Clark AM, Williams AJ, Ekins S. Machines first, humans second: On the importance of algorithmic interpretation of open chemistry data. *J Chem*. 2015;*7*:9
137. DeepChem [Internet]; 2020. https://deepchem.io/docs/index.html. Accessed 31st December 2020.
138. RDKit. Open-source cheminformatics and machine learning [Internet]; 2020. https://rdkit.blogspot.com/. Accessed 09 Sep 2020.
139. Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci*. 2018;*9*(2):513–530.

140. PyTorch. PyTorch is an optimized tensor library for deep learning using GPUs and CPUs. [Internet]; 2020. https://pytorch.org/docs/stable/index.html. Accessed 09 Sep 2020.
141. KNIME [Internet]. https://www.knime.com/about. Accessed 09 Sep 2020.
142. Skuta C, Popr M, Muller T, et al. Probes & drugs portal: an interactive, open data resource for chemical biology. *Nat Methods*. 2017;*14*: 759–760.
143. Kim S, Thiessen PA, Bolton EE, et al. PubChem substance and compound databases. *Nucleic Acids Res*. 2016;*44*(D1):D1202–D1213.
144. Papadatos G, Davies M, Dedman N, et al. SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Res*. 2016;*44*(D1):D1220–D1228.
145. Schaduangrat N, Lampa S, Simeon S, Gleeson MP, Spjuth O, Nantasenamat C. Towards reproducible computational drug discovery. *J Chem*. 2020;*12*:1–30.
146. Orphanet [Internet]. https://www.orpha.net/consor/cgi-bin/index.php. Accessed 16 Nov 2020.
147. Tox21. Overview [Internet]. https://tox21.gov/overview/. Accessed 16 Nov 2020.
148. EU-ToxRisk—EU-ToxRisk—An Integrated European 'Flagship' Programme Driving Mechanism-based Toxicity Testing and Risk Assessment for the 21st century [Internet]. https://www.eu-toxrisk.eu/. Accessed 16 Nov 2020.
149. The OECD QSAR Toolbox—OECD [Internet]. https://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm. Accessed 16 Nov 2020.
150. pandas—Python Data Analysis Library [Internet]. https://pandas.pydata.org/. Accessed 16 Nov 2020.
151. scikit-learn: machine learning in Python—scikit-learn 0.23.2 documentation [Internet]. https://scikit-learn.org/stable/. Accessed 16 Nov 2020.
152. ChEMBL-NTD—ChEMBL-NTD [Internet]. https://chembl.gitbook.io/chembl-ntd/. Accessed 16 Nov 2020.
153. TDR Targets [Internet]. https://tdrtargets.org/. Accessed 16 Nov 2020.
154. Drug discovery | DNDi [Internet]. https://dndi.org/research-development/drug-discovery/. Accessed 16 Nov 2020.
155. García-Sosa AT, Oja M, Hetényi C, Maran U. DrugLogit: logistic discrimination between drugs and nondrugs including disease-specificity by assigning probabilities based on molecular properties. *J Chem Inf Model*. 2012;*52*(8):2165–2180. https://dx.doi.org/10.1021/ci200587h.
156. Medicines for Malaria Venture | Developing antimalarials to save lives [Internet]. https://www.mmv.org/. Accessed 16 Nov 2020.
157. The Pathogen Box | Medicines for Malaria Venture [Internet]. https://www.mmv.org/mmv-open/pathogen-box. Accessed 18 Nov 2020.
158. The Pandemic Response Box | Medicines for Malaria Venture [Internet]. https://www.mmv.org/mmv-open/pandemic-response-box. Accessed 16 Nov 2020.
159. About the Malaria Box | Medicines for Malaria Venture [Internet]. https://www.mmv.org/mmv-open/malaria-box/about-malaria-box. Accessed 16 Nov 2020.
160. TriTrypDB [Internet]. https://tritrypdb.org/tritrypdb/app. Accessed 16 Nov 2020.
161. VEuPathDB [Internet]. https://veupathdb.org/veupathdb/app/. Accessed 16 Nov 2020.
162. Home—Gene—NCBI [Internet]. https://www.ncbi.nlm.nih.gov/gene. Accessed 16 Nov 2020.
163. PANTHER—Gene list analysis [Internet]. http://www.pantherdb.org/. Accessed 16 Nov 2020.
164. SwissADME [Internet]. http://www.swissadme.ch/index.php. Accessed 16 Nov 2020.
165. InterPro [Internet]. http://www.ebi.ac.uk/interpro/. Accessed 16 Nov 2020.
166. Pfam: Home page [Internet]. http://pfam.xfam.org/. Accessed 16 Nov 2020.
167. SMART: Main page [Internet]. http://smart.embl-heidelberg.de/. Accessed 16 Nov 2020.
168. SUPERFAMILY database of structural and functional protein annotations for all completely sequenced organisms [Internet]. http://supfam.org/SUPERFAMILY/index.html. Accessed 16 Nov 2020.
169. SSGCID | SSGCID [Internet]. https://www.ssgcid.org/. Accessed 16 Nov 2020.
170. KEGG: Kyoto Encyclopedia of Genes and Genomes [Internet]. https://www.genome.jp/kegg/. Accessed 16 Nov 2020.
171. UniProt [Internet]. https://www.uniprot.org/. Accessed 16 Nov 2020.
172. GeneDB—Home [Internet]. https://www.genedb.org/. Accessed 16 Nov 2020.
173. RCSB PDB: Homepage [Internet]. https://www.rcsb.org/. Accessed 16 Nov 2020.
174. ZINC [Internet]. http://zinc20.docking.org/. Accessed 16 Nov 2020.
175. Irwin JJ, Tang KG, Young J, et al. ZINC20—a free ultralarge-scale chemical database for ligand discovery. *J Chem Inf Model*. 2020;*60*: 6065–6073.
176. Ekins S, de Siqueira-Neto JL, Mccall LI, et al. Machine learning models and pathway genome data base for trypanosoma cruzi drug discovery. *PLoS Negl Trop Dis*. 2015;*9*(6):e0003878.
177. Collaborative Drug Discovery. *Collaborative drug discovery public*; 2015. https://www.collaborativedrug.com/public-access/. Accessed 28 Dec 2020.
178. World Health Organization. Reports of the World Health Organization 2011; 2011.
179. Croft SL, Coombs GH. Leishmaniasis—current chemotherapy and recent advances in the search for novel drugs. *Trends Parasitol*. 2003;*19*:502–508.
180. Jamal S, Scaria V. Cheminformatic models based on machine learning for pyruvate kinase inhibitors of *Leishmania mexicana*. *BMC Bioinformatics*. 2013;*14*(1):329.

181. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res*. 2009;*37*:623–633.
182. 2C4C Model Repository—Vinod Scaria MBBS, PhD [Internet]. http://vinodscaria.rnabiology.org/2C4C/models. Accessed 11 Sep 2020.
183. Santa Maria JP, Park Y, Yang L, et al. Linking high-throughput screens to identify MoAs and novel inhibitors of *Mycobacterium tuberculosis* dihydrofolate reductase. *ACS Chem Biol*. 2017;*12*(9):2448–2456.
184. Glick M, Jenkins JL, Nettles JH, Hitchings H, Davies JW. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and Laplacian-modified naive Bayesian classifiers. *J Chem Inf Model*. 2006;*46*: 193–200.
185. Neves BJ, Braga RC, Alves VM, et al. Deep learning-driven research for drug discovery: tackling malaria. *PLoS Comput Biol*. 2020;*16*(2): e1007025.
186. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure–activity relationships. *J Chem Inf Model*. 2015;*55*(2):263–274.
187. Komatsu R, Honda M, Holzgrefe HH, et al. Sensitivity of common marmosets to detect drug-induced QT interval prolongation: moxifloxacin case study. *J Pharmacol Toxicol Methods*. 2010;*61*(3):271–276.
188. Deshpande D, Pasipanodya JG, Mpagama SG, et al. Levofloxacin pharmacokinetics/pharmacodynamics, dosing, susceptibility breakpoints, and artificial intelligence in the treatment of multidrug-resistant tuberculosis. *Clin Infect Dis*. 2018;*67*(Suppl 3):S293–S302.
189. Casimiro-Soriguer CS, Loucera C, Perez Florido J, López-López D, Dopazo J. Antibiotic resistance and metabolic profiles as functional biomarkers that accurately predict the geographic origin of city metagenomics samples. *Biol Direct*. 2019;*14*(1):15.
190. Kabra R, Ingale P, Singh S. Computationally designed synthetic peptides for transporter proteins imparts allostericity in Miltefosine resistant *L. major*. *Biochem J*. 2020;*477*(10):2007–2026.
191. Perryman AL, Patel JS, Russo R, et al. Naïve Bayesian models for Vero cell cytotoxicity. *Pharm Res*. 2018;*35*:170.
192. Lane T, Russo DP, Zorn KM, et al. Comparing and validating machine learning models for *Mycobacterium tuberculosis* drug discovery. *Mol Pharm*. 2018;*15*(10):4346–4360.
193. Dinić J, Efferth T, García-Sosa AT, et al. Repurposing old drugs to fight multidrug resistant cancers. *Drug Resist Updat*. 2020;*52*: 100713100713. https://doi.org/10.1016/j.drup.2020.100713.
194. Hodos RA, Kidd BA, Shameer K, Readhead BP, Dudley JT. In silico methods for drug repurposing and pharmacology. *Wiley Interdiscip Rev Syst Biol Med*. 2016;*8*(3):186–210.
195. Vilar S, Uriarte E, Santana L, et al. Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nat Protocols*. 2014;*9*: 2147–2163.
196. Guney E. Reproducible drug repurposing: when similarity does not suffice. *Pacific symposium on biocomputing*. 2017;*22*:132–143.
197. Nabirotchkin S, Peluffo AE, Rinaudo P, Yu J, Hajj R, Cohen D. Next-generation drug repurposing using human genetics and network biology. *Curr Opin Pharmacol*. 2020;*51*:78–92.
198. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet*. 2017; *101*:5–22.
199. Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet*. 2016;*48*(7):709–717.
200. Barabási AL, Oltvai ZN. Understanding the cell's functional organization. *Nat Rev Genet*. 2004;*5*:101–113.
201. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *npj Digit Med*. 2018;*1*(1):18.
202. Zong N, Kim H, Ngo V, Harismendy O. Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations. *Bioinformatics*. 2017;*33*(15):2337–2344.
203. Ferrero E, Brachat S, Jenkins JL, et al. Ten simple rules to power drug discovery with data science. *PLoS Comput Biol*. 2020;*16*(8): e1008126.
204. Piir G, Kahn I, García-Sosa AT, Sild S, Ahte P, Maran U. Best practices for QSAR model reporting: physical and chemical properties, ecotoxicity, environmental fate, human health, and toxicokinetics endpoints. *Environmental health perspectives*. 2018;*126*:126001. https://dx.doi.org/10.1289/EHP3264.
205. Lowe D. Another AI-generated drug? | In the pipeline [Internet]; 2020. https://blogs.sciencemag.org/pipeline/archives/2020/01/31/another-ai-generated-drug. Accessed 09 Sep 2020.
206. Brown N, Ertl P, Lewis R, Luksch T, Reker D, Schneider N. Artificial intelligence in chemistry and drug design. *J Comput Aided Mol Des*. 2020;*34*:709–715.
207. Reyzin L. Unprovability comes to machine learning. *Nature*. 2019;*565*(7738):166–167.
208. Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: benchmarking models for de novo molecular design. *J Chem Inf Model*. 2019;*59*(3):1096–1108.
209. Muratov EN, Bajorath J, Sheridan RP, et al. QSAR without borders. *Chem Soc Rev*. 2020;*49*:3525–3564.
210. Hutson M. Core progress in AI has stalled in some fields. *Science*. 2020;*368*:927.
211. Bakhoum M, Gallego B, Mackenrodt M, Surblytė-Namavičienė G. *Personal data in competition, consumer protection and intellectual property law*. Berlin: Springer, 2018.
212. Banterle F. *The interface between data protection and IP law: the case of trade secrets and the database sui generis right in marketing operations, and the ownership of raw data in big data analysis*. Berlin, Heidelberg: Springer, 2018;p. 411–443.

213. How to lie with computational predictive models in drug discovery—DrugDiscovery.NET—AI in drug discovery [Internet]. http://www.drugdiscovery.net/2020/10/13/how-to-lie-with-computational-predictive-models-in-drug-discovery/. Accessed 16 Nov 2020.

214. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. *Nat Mach Learn*. 2020;*2*:573–584.

215. Horrobin DF. Modern biomedical research: An internally self-consistent universe with little contact with medical reality? *Nat Rev Drug Discov*. 2003;*2*(2):151–154.

216. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. *Nat Chem*. 2012;*4*(2):90–98.

How to cite this article: Peña-Guerrero J, Nguewa PA, García-Sosa AT. Machine learning, artificial intelligence, and data science breaking into drug design and neglected diseases. *WIREs Comput Mol Sci*. 2021;e1513. https://doi.org/10.1002/wcms.1513