Chemical structure and machine learning: leveraging the power of deep neural networks to aid the development of long-acting drugs for HIV and HIV coinfections

Camille Lake, PhD

Emerging Leaders in Data Science Fellowship Rotation Project

Office of Data Science and Emerging Technologies, NIAID

Supervisors: Marina Protopopova, PhD and Mohamed Nasr, PhD

Table of Contents

Introduction

The human immunodeficiency virus (HIV) continues to have a significant health impact throughout the world. Currently, 38 million people worldwide live with HIV, and 1-2 million new infections are recorded each year[1]. Opportunistic infections are the major cause of morbidity and mortality among those living with HIV. *Mycobacterium tuberculosis* and hepatitis B virus (HBV) are two of the pathogens most encountered in co-infections with HIV. Tuberculosis is responsible for roughly 26% of all HIV-related deaths worldwide, while about 10% of all individuals with HIV are also infected with HBV[2]. Current therapeutic regimens for HIV focus on reducing the frequency and severity of such opportunistic co-infections.

Despite decades of effort, there are currently no therapies which eradicate HIV from an individual, nor a preventative vaccine[1]. Current antiretroviral therapies (ART) focus on reducing HIV viral loads to undetectable levels, preventing immune suppression. Most ART regimens are administered once daily as oral preparations. Individual adherence to regular treatment is critical to successful management, but consistent daily use over a lifetime can prove challenging for patients to sustain. Challenges to treatment are amplified by lack of healthcare resources and infrastructure, issues that remain especially prominent in those regions where HIV is most prevalent[2]. The World Health Organization (WHO) estimates that as much as two-thirds of the global HIV burden remains within Africa[1], while in the West, HIV is most prevalent in areas with the highest burdens of poverty[3,4]. Stigma, medication side-effects, lack of assistance and forgetfulness were among the top reasons for lack of adherence to ART in sub-Saharan Africa[5], with similar concerns impacting adherence rates in the U.S.[6]. There is a clear need for continued development of ART that simplifies dosing to improve treatment success at the population level.

The recent introduction of long-acting ART regimens has demonstrated the potential to manipulate drug formulation in ways that may simplify adherence to treatment protocols[7]. A novel long-acting HIV treatment recently introduced combines cabotegravir and rilpivirine into an extended-release injectable suspension formulation[8,9]. Both modeling predictions and clinical trials have demonstrated that the long-acting cabotegravir/rilpivirine injection is an effective measure to increase adherence to ART regimens[10,11]. A number of other promising long-acting drug candidates are currently in the developmental stages, as summarized in Table 1[12].

Over the past decade, the utilization of machine learning applications have played an increasingly critical role in drug discovery and development[13]. Pharmaceutical research has embraced deep learning methods[14] such as deep neural networks (DNNs) and convolutional neural networks (CNNs) as a means to optimize the drug discovery process[13]. The application of such machine learning techniques to development of long-acting ART formulations may offer a valuable approach for identifying new and more effective HIV treatment regimens.

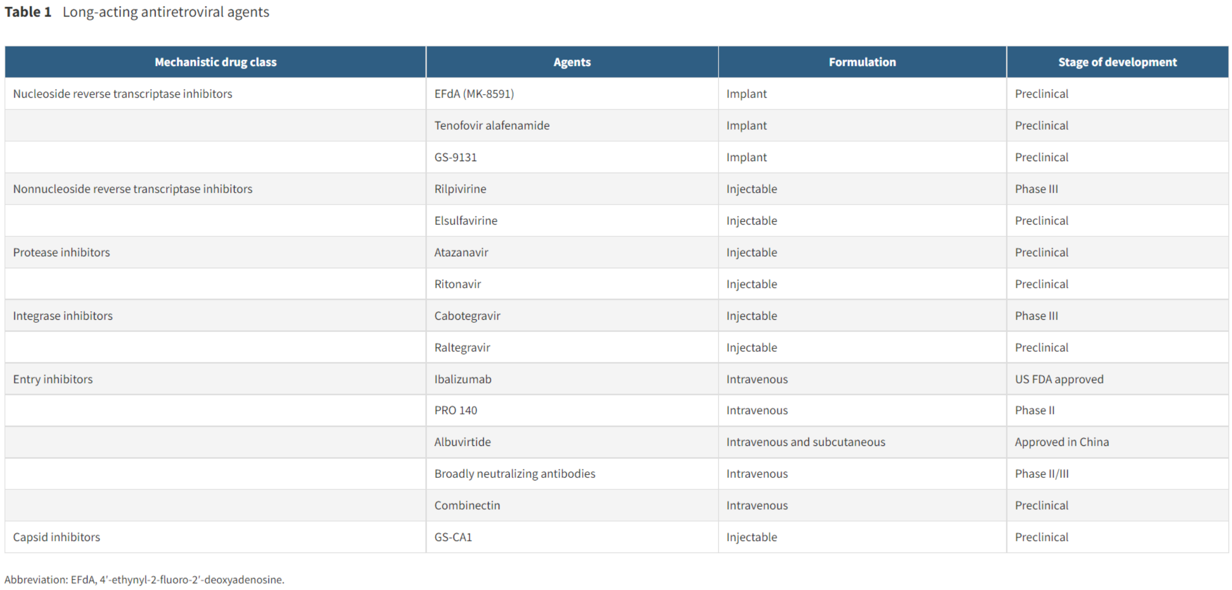


Table 1 Long-acting antiretroviral agents

Mechanistic drug class	Agents	Formulation	Stage of development
Nucleoside reverse transcriptase inhibitors	EFdA (MK-8591)	Implant	Preclinical
	Tenofovir alafenamide	Implant	Preclinical
	GS-9131	Implant	Preclinical
Nonnucleoside reverse transcriptase inhibitors	Rilpivirine	Injectable	Phase III
	Elsulfavirine	Injectable	Preclinical
Protease inhibitors	Atazanavir	Injectable	Preclinical
	Ritonavir	Injectable	Preclinical
Integrase inhibitors	Cabotegravir	Injectable	Phase III
	Raltegravir	Injectable	Preclinical
Entry inhibitors	Ibalizumab	Intravenous	US FDA approved
	PRO 140	Intravenous	Phase II
	Albuvirtide	Intravenous and subcutaneous	Approved in China
	Broadly neutralizing antibodies	Intravenous	Phase II/III
	Combinectin	Intravenous	Preclinical
Capsid inhibitors	GS-CA1	Injectable	Preclinical

Abbreviation: EFdA, 4′-ethynyl-2-fluoro-2′-deoxyadenosine.

Table 1 A summary of long-acting antiretroviral agents currently under development, taken from Gulick et al[12].

Rotation Objective and Aims

Project objective

Train a deep neural network to predict pharmacokinetic parameters based on molecular properties and drug structure, as a means of prioritizing optimal compounds for the repurposing process of long-acting formulations.

The long-term objective of this rotation project is twofold:

1. To gain experience with the interface of deep neural networks and pharmaceutical compounds for career development
2. To facilitate the distribution of this information for researchers working on long-acting formulations, thereby supporting the HIV research community

Aims

Aim 1: Determine the most relevant pharmacokinetic parameter(s) necessary for long-acting formulation development.

Aim 2: Develop a machine learning model (likely a deep neural network) to predict the relevant pharmacokinetic parameter determined in Aim 1.

Aim 3: Develop a plan to integrate the relevant pharmacokinetic parameter(s) and possibly the model into ChemDB to provide wider access of these to the research community.

Approach & Progress to Date

Aim 1: Determine the most relevant pharmacokinetic parameter necessary for long-acting formulation development.

Approach:

- Perform literature and textbook review of key pharmacokinetic parameters
- Discuss topic with subject matter experts

Progress to Date: Facilitated meeting with rotation mentors Drs. Marina Protopopova and Mohamed Nasr, as well as Drs. Andrew Owen, Charles Flexner, and Tao You to discuss the best parameter to pursue as the prediction value of the neural network. It was generally agreed upon by the subject matter experts and confirmed by the literature that clearance rate would be the most relevant parameter to long-acting formulation development.

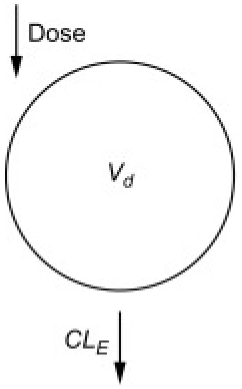


Figure 2 Total drug elimination clearance (CL_E) is defined as the removal of drugs eliminated by first-order kinetics and is directly related to the drug volume of distribution (V_d). Taken from Atkinson et al[15].

Aim 2: Develop a machine learning model to predict clearance.

Approach:

- Research most relevant machine learning models relevant to drug discovery and development
- Determine necessary components for success of model, including necessary datasets and information format

Progress to Date: A comprehensive literature search suggested that a Directed Message Passing Neural Network (D-MPNN) called ChemProp[16] would be the most relevant model to train. Drug structure in SMILES format as well as relevant pharmacokinetic parameters will be used to train and validate the model (see below for more information).

SMILES: Simplified Molecular Input Line Entry System; a string-based computer-readable format for 3-dimensional drug structure[17].

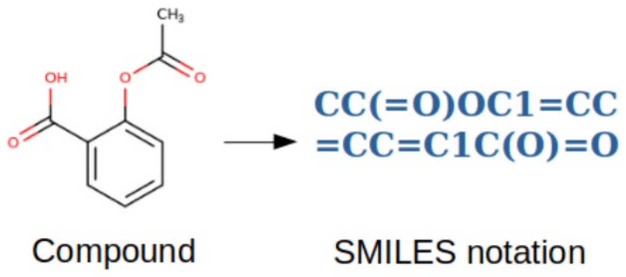


Figure 3 An example chemical compound and its corresponding SMILES string. Taken from Pham et al.[18]

MPNNs: Neural networks which predict drug properties based on chemical "fingerprints": molecular descriptors of functional groups and bonds[19,20]. Because these networks function by passing "messages" across the defined vector space of the reconstructed chemical compound, they are termed directed "message passing" neural networks (D-MPNNs)[21].

ChemProp: A molecular property prediction D-MPNN which has been successfully trained to predict novel antibiotics[22]. ChemProp is outfitted with an additional algorithm which seeks to identify features within the supplied chemical structures responsible for the decisions that affect the outcomes[16]. As a result, this algorithm is the ideal choice for a project which is centered around property prediction and structure optimization for long-acting potential.

Datasets and Information to Collect: Multiple databases were assessed for their ability to provide the data necessary for the training of the model, including PubChem[23] and its supporting databases. One such database, DrugBank[24], provides the most comprehensive dataset including pharmacokinetic information on a wide variety of drugs. The SMILES information will be manually annotated (using the PubChem database) into the dataset in order to provide ChemProp with all relevant information for making clearance predictions.

Aim 3: Develop a plan to integrate the relevant pharmacokinetic parameter(s) and possibly the model into ChemDB to provide wider access of these to the research community.

Approach:

- Work with Dr. Mohamed Nasr to determine which parameters would be most likely to help the larger research community
- Determine the process for integration of DrugBank's API with ChemDB

Progress to Date: discussion on-going.

Update Regarding Project Direction / Aims

During the course of this project, new directions and courses were taken in order to maximize the output and provide additional learning experiences for the student. In general, the above aims were met, but the following additions were made:

1. Though it was generally agreed that clearance was the most apt PK parameter to study, *all PK parameters (volume of distribution, clearance, and half-life) were explored,* both singly and in combination, in order to determine which created the most accurate model.
2. Additional ChemProp utilities, including the interpret script, were used in order to explore the substructures contributing to the decision-making process of the trained algorithms.
3. For the sake of completeness and thoroughness of the project, the datasets for training and testing were kept separate from the prediction dataset. More explicitly, the machine learning models were trained and tested on a dataset of drugs across indications *but were used to predict PK parameters of a separate set of drugs with indications for HIV/HIV co-morbidities*. It was generally thought that this separation and utilization of the HIV dataset would demonstrate the utility of the model and its ability to predict on the dataset of interest.

The above aims were used as guideposts for the course of this project and were generally followed and met. Below is the final project that the student engaged in with the ultimate goal of exploring deep neural networks within the field of pharmacology, while also providing a proof-of-concept layout for the Division of AIDS / Therapeutic Research Program to increase transparency of compounds and speed the development of long-acting agents within this field.

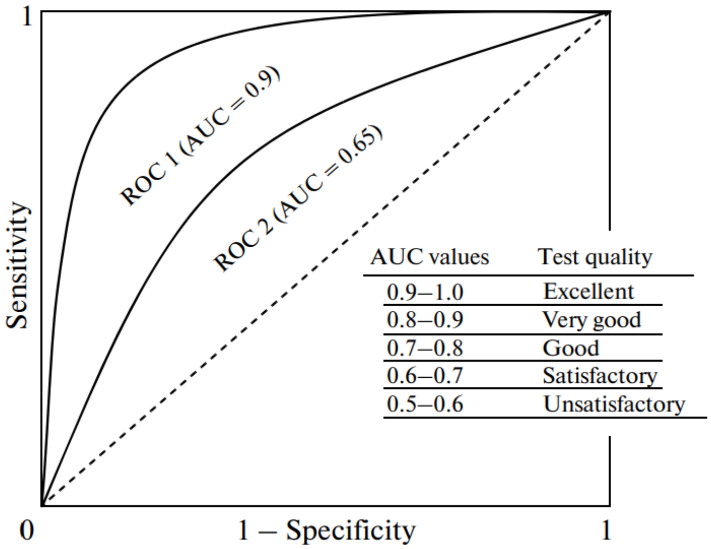
Results

To create the machine learning algorithms necessary for predicting the chosen pharmacokinetic parameters, training, testing, and prediction datasets were compiled. For this project, the training and testing dataset included a wide variety of drugs across a spectrum of indications, which was then split into train/test/validation datasets. A separate collection of drugs indicated for HIV, Tuberculosis, or Hepatitis B was used as the prediction set. More details are included below.

- *Train/Test/Validation dataset:* drugs were collected from DrugBank[24], and the dataset was appended to include relevant information on SMILES, molecular weight, volume of distribution, clearance, half-life, and route of delivery (Supplemental Table 1). The drugs were small molecules (Fig. 5D) with normally distributed pharmacokinetic parameters (Fig. 5A-C). Drugs included a wide variety of indications. The majority were indicated for oral delivery, with the remaining subset indicated for intravenous or other administration (Fig. 5E).
- *Prediction dataset:* drugs with an indication for HIV[25], tuberculosis[26], or Hepatitis B[27] were collected from DrugBank[24], as well as the sources cited after each indication. Information on SMILES and relevant pharmacokinetic parameters (if available) were also collected and can be found in Supplemental Figure 2. The drugs in this dataset were also generally small molecules (Fig. 6A) and were predominantly indicated for HIV and tuberculosis, though some in the set were indicated for Hepatitis B or were multi-indication (Fig. 6B).

ChemProp can be trained either in classification or regression (i.e., binary values of 0 and 1, or in actual raw values of the data, respectively). Because the train/test/validate model was small (only 806 molecules total), classification was opted for in the training. The median values of each PK parameter served as the boundary for binary classification, in which “1” was interpreted to be more favorable and “0” value was interpreted to be less favorable. For volume of distribution, any drug above 1.25 ml/min*kg (median of the train/test/validate test set) was set to 1, whereas any drug with a volume of distribution below this value was set to 0. For half-life, drugs with half-lives above 6.7 hours was set to 1, where any below this value were set to 0. Lower clearance values are generally considered more optimal; therefore, clearance values below 3.645 L/kg (median) were set to 1, and vice versa above this value. For more information please reference the Methods section.

Following hyperparameter optimization (included in the ChemProp package), the AUC/ROC prediction scores were output on the training and testing dataset (Fig. 7A). In general, favorable AUC/ROC curves can be interpreted as shown below:



AUC values	Test quality
0.9–1.0	Excellent
0.8–0.9	Very good
0.7–0.8	Good
0.6–0.7	Satisfactory
0.5–0.6	Unsatisfactory

Figure 4 An example of ROC curves with excellent (AUC = 0.9) and satisfactory (AUC = 0.65) outcomes. Taken from Trifonova et al. [28]

In general, clearance and half-life did not perform well under the current conditions for binary classification prediction (0.59 for both); however, volume of distribution performed well (0.78) (Fig. 7A), which could likely be improved with an increased dataset size. In addition to the PK parameters being used to create separate ChemProp algorithms, the scores were combined in two ways: the first was the keep all classification entries as separate columns in the train/test dataset. This method led to increased scores across the board for each PK parameter (Fig. 7B), suggesting that the model utilizes the three scores collectively to improve each one; this is perhaps not surprising given the interconnected nature of the PK parameters used in this project (see Figure 2). The second method of score combination included "melting" the three separate binary classifications into a single column score. This was achieved by creating score combinations: any drug with a total combined score of 2+ for the three PK parameters was deemed "ideal" and was assigned a 1; any drug with a combined score of 0 or 1 was deemed "less ideal" and assigned a 0. This yielded an AUC/ROC of 0.75 (Fig. 7A).

In order to determine how close these scores were to actual values in the literature, the predicted classifications of each drug were plotted against actual values (in this case, only volume of distribution was investigated, as this PK parameter had the highest classification AUC/ROC). In general, those predicted to have a "lower" volume of distribution (bin 0) had lower overall clinical volume of distribution values, and vice versa (Fig. 7C). A Fisher's exact test (using a 2x2 contingency table) also echoed these results (Fig. 7D). In both cases, a more robust dataset would improve the outcomes; this is discussed in more detail in the discussion.

ChemProp has a separate interpret function which uses a Monte Carlo tree search algorithm to interpret the predicted outcomes of the dataset. This algorithm points to the substructures and functional groups which contributed the most to determining the classification that the trained classifier chose, allowing a glimpse into the decision-making process of this directed message passing neural network (D-MPNN). The interpret function was run on the merged variable dataset, in which all scores were combined into a single “ideal” (2+) versus “less ideal” (1 or 0 respectively) score. Two highlighted drugs are included as examples in Fig. 8A and B, with zoomed in substructures which were determined to be the most important part of the entire molecule to decide that these would be “ideal” drug candidates. To quantify and explore these substructures further, RDKit was used to search for unique functional groups (a complete list of which can be found in Supplemental Figure 3). The most common functional groups among the substructures of those “ideal” drug candidates can be found in Fig. 8C.

Discussion & Conclusion

Long-acting drugs have had a remarkable impact on a multitude of communities throughout the world, across a wide range of drug indications. While pharmaceutical companies typically modify drug formulation in order to achieve long-acting activity, utilizing the underlying pharmacokinetic parameters to predict which drugs might be optimal candidates prior to long-acting formulation development could maximize productivity and reduce time and expenditures related to trial-and-error methods of long-acting formulation development. Using a machine learning platform called ChemProp, we were able to predict the key pharmacokinetic parameters of volume of distribution, clearance, and half-life on a dataset of small molecules, and extend these algorithms to a dataset of drugs indicated for HIV and associated co-morbidities. While there were significant drawbacks of this research which caution interpretation and utilization of the models used in this report, the overarching concept of using deep neural networks to predict these PK parameters to further drug research and decision-making suggests that this framework could be a useful asset for drug-development pipelines in the future.

The drugs utilized in the training and testing section of this study were taken from a dataset provided by DrugBank[24], in which the final version consisted of 806 small molecules with volume of distribution (L/kg), clearance rates (ml/min*kg), and half-life (hours) included (see Supp. Fig. 1 for the full dataset). As mentioned previously, all were small molecules across a range of indication, with the majority being orally administered (Fig. 5). ChemProp was trained and tested on this dataset, in which all raw values were converted to binary values with the median score being the cutoff, with varying levels of success (as interpreted by the AUC/ROC scores, Fig. 7). Intriguingly, the volume of distribution had the overall highest AUC/ROC score when used to train and test a model alone (0.78) and in tandem with the other pharmacokinetic parameters (0.82) following hyperparameter optimization. This potentially suggests some underlying scientific basis for the higher classification score, such as consistency of substructures among the molecules with the highest scores. This is perhaps not surprising, given the existence of Lipinski scores and other "rules" which are associated with the success of molecules in oral delivery systems based on their structure and other related parameters[29]. The implication of this is that long-acting drug formulation could be improved by prioritizing drugs with predicted higher volumes of distribution based on their structure, as data in Fig. 7D-E suggest. It is important to keep in mind, however, that working to improve clearance and half-life scores through the addition of more molecules to the training and testing dataset, or making changes to the cutoff point for classification, would both likely lead to score improvements.

Although the classification model trained and tested only on volume of distribution data had the most success, including all of the scores together (as separate entities) improved all three of the PK parameter scores (Fig. 7B). This suggests that the scores are interrelated, which is unsurprising given that predicted half-life measurements are directly proportional to volume of distribution and inversely proportion to clearance rates[30]. Additionally, another "summation" algorithm was also tested by computing a proxy combination score, in which drugs were considered "optimal" if they had two or more favorable scores across the three PK parameter classifications. This combination score yielded an AUC/ROC score of 0.75 (Fig. 7A). Overall, the scores of the models seem to point to the utilization of all available PK parameters if available in order to maximize the accuracy of the model; however, further research should be conducted to determine which method is the most beneficial and useful for determining optimal candidates for long-acting drug formulation.

In addition to understanding if a model could be built to predict pharmacokinetic parameters from drug structure, the ChemProp interpret script was also used to determine if certain functional groups or substructures could be identified among the HIV drug set that aided in the decision-making process of the binary classifier. For simplicity, only the model trained on the combination score (merged into a single variable, as detailed above) was used to assess these substructures (Fig. 8). Although certain functional groups, such as the presence of amines (Nitrogen in general) and halides, seemed to be prominent in this decision-making, heavy caution is advised in overinterpreting the value of these data. First, the decisions are a direct result of the data that was used to train and test the classifier, and therefore, these functional groups are directly related to this dataset (which may not have anything to do with the HIV prediction dataset). Secondly, the paucity of data in the training data for the model likely directly influences its decisions. We therefore regard the data in Figure 8 as a proof of concept that substructures can be directly inferred from the algorithm itself and could be utilized in the future to potentially aid in the structural modification of HIV drugs; however, this data is not complete enough to make these inferences at present.

Overall, in this report, it was demonstrated that ChemProp could be used to predict important pharmacokinetic parameters of HIV, Hepatitis B, and Tuberculosis drugs. The model trained singly in predicting volume of distribution had the best overall AUC/ROC; however, training the models on combinations of the PK parameters (either as separate entities or as a combined single score) generally led to good prediction scores as well, suggesting that the utilization of multiple PK parameters may be a useful framework to integrate into drug development pipelines in the future. As the datasets are relatively small in this study, and overgeneralization is cautioned, the findings in this report stand as a proof-of-concept that pharmacokinetic property prediction can be done using deep neural networks, which may eventually help advance the field towards creating long-lasting drug regimens that significantly improve overall health outcomes.

Methods

Datasets

Train/test/validate Dataset

Datasets used in this report were provided by DrugBank[24], using an academic license application. Datasets were modified so that all entries were in the same units, as specified in the below table. Common rules for amending the original dataset from DrugBank are listed below, with specific rules for each PK parameter are outlined in the table below.

Rules that applied to all PK parameters: in cases where the kilograms were not provided in the original dataset, the average human weight of 62 kg was used to calculate the kg into the units if applicable. Only human data was kept from the original dataset (i.e., no animal data). If multiple values were provided, healthy adult information was prioritized. In the cases where a mean was provided, the mean was accepted as the input value; in the cases where only a range was provided, the smallest number in that range was accepted as the input value. If multiple values are provided from multiple drug doses, the smallest dose of the drug was chosen as the input value.

PK Parameter	Unit	Notes for dataset correction
Volume of Distribution	Liters/kilogram (L/kg)	If have both central and peripheral values, using peripheral
Clearance	Milliliters/minute*kilogram (mL/min*kg)	Order to prioritize clearance measurement: total clearance, plasma clearance, renal clearance
		Order to prioritize clearance differentiated by drug administration: oral, intravenous
Half-life	Hours (hrs)	If given multiple half-lives due to kinetics, choosing the terminal half-life

The dataset used in training and testing can be found in Supplemental Figure 1.

Prediction Dataset

The HIV/Hepatitis B/Tuberculosis drug dataset was collected from a multitude of sources. The HIV drugs were collected from HIVinfo.nih.gov[25], the Hepatitis B drugs were collected from HepB.org[27], and the Tuberculosis drugs from MedIndia.net[26]. All drugs were assessed on PubChem[23] for SMILES and DrugBank for relevant pharmacokinetic parameters if available.

Software Packages

ChemProp

ChemProp is a message passing neural network package for property prediction of chemical molecules and is available on GitHub (https://github.com/chemprop/chemprop). Full links to documentation, associated publications, and use cases are available on the GitHub.

In this report, all ChemProp algorithms were utilized on the command line, as per instructions in the authors' documentation. Commands run included training/testing, hyperparameter optimization, interpretation, and prediction, all using standard, default parameters recommended in the documentation. For a complete list of commands used, please refer to the personal GitHub of the author, where a repository housing all code used in this report is available (https://github.com/mad-scientist-in-training/PK_Project). Specifically for commands run in command line, please see conda_hx.txt. For all other code run in this project, please refer to the Chemical Structure Prediction Jupyter Notebook.

RDKit

RDKit is a suite of cheminformatics and machine learning software packages and tools available on GitHub (https://github.com/rdkit/rdkit). In this report, the Chem package from RDKit was used to toggle data between SMILES, SMARTS, and string formats where necessary.

Chemistry Drawer

ChemDraw is a package used to visualize molecular structures from SMILES or SMARTS format available on GitHub (https://github.com/dylanwal/chemistry_drawer). In this report, all molecular structures generated (including those in Figure 8) were created using this package.

Visualization & Statistics

Python visualization packages (Matplotlib and Seaborn)

Matplotlib (v3.5.2) and Seaborn (v0.11.2) Python libraries were used to generate graphical representations found within this report; for more information please see the above documentation on the author's GitHub page.

GraphPad Prism

GraphPad Prism (version 9) was used to generate graphics and run statistical analyses of the data included in this report. Specifically, Prism was used to run the Fisher's Exact Test shown in Figure 7D.

Adobe Illustrator

Adobe Illustrator 2022 was used as the primary visual medium for figure collection and presentation throughout this report.

Figures

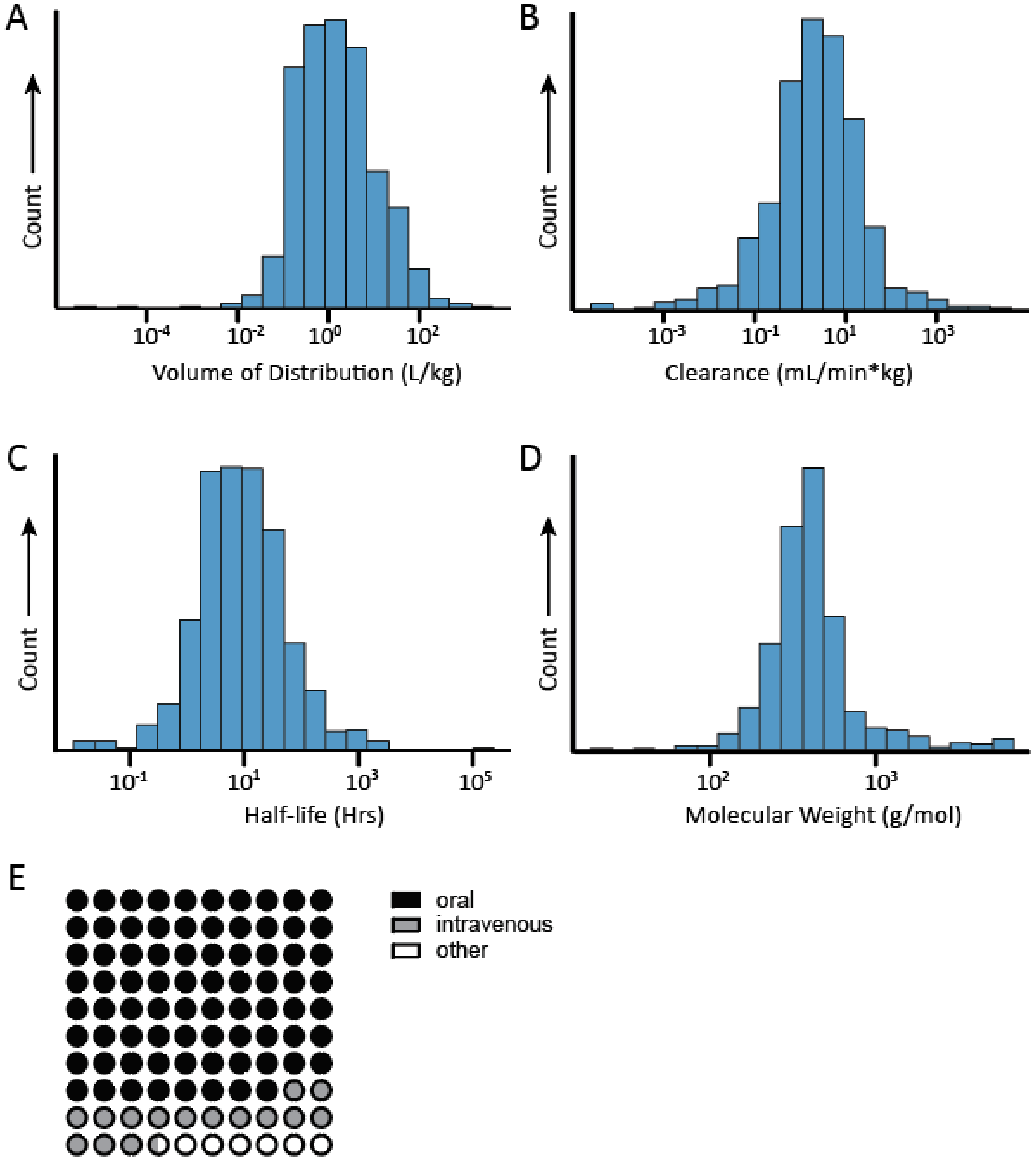
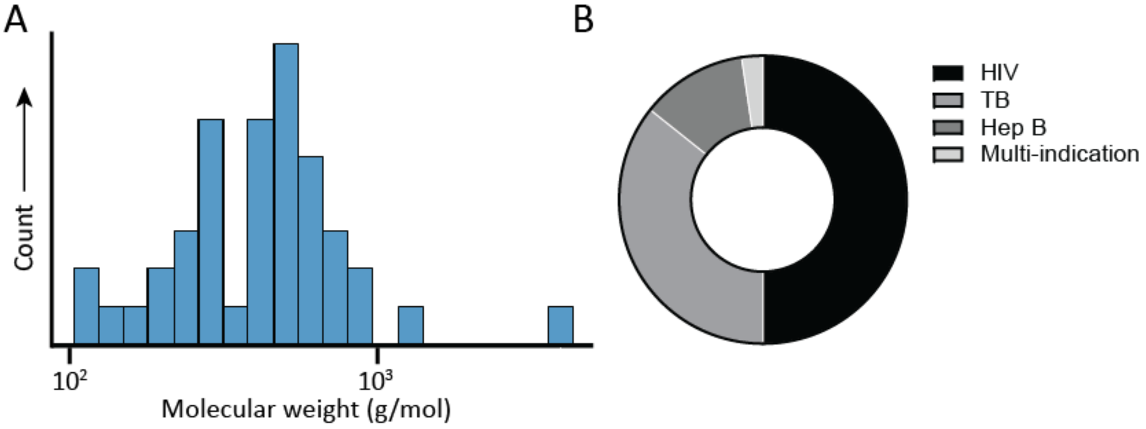


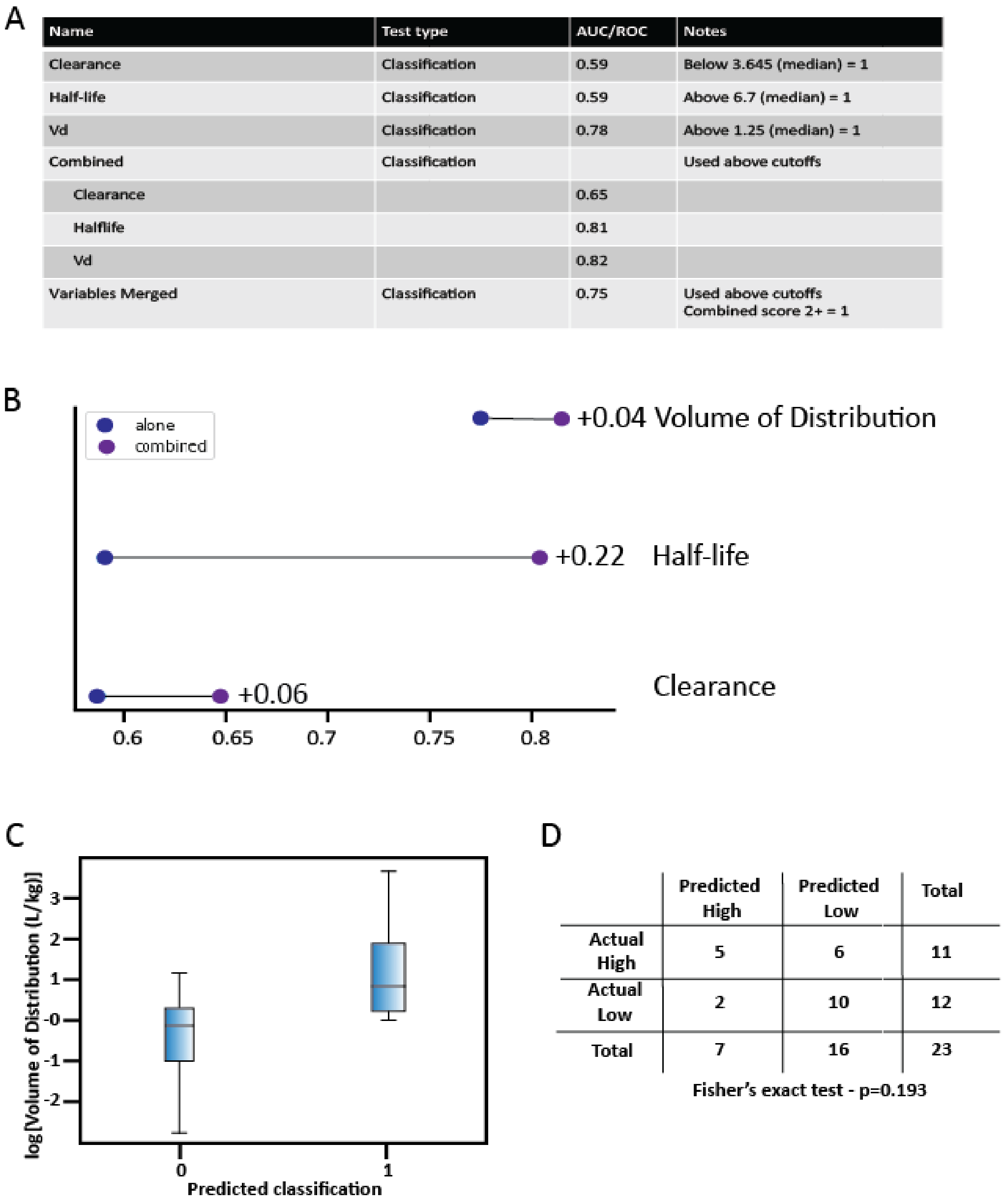
Figure 5: Drugs used in the training and testing dataset of ChemProp. Small molecule drugs were collected from DrugBank[24], and used to train and test ChemProp. In sum, 806 drugs were included in the training and testing dataset across a multitude of indications, structures, and pharmacokinetic values. (5A) The median volume of distribution was 1.25 L/kg, with an interquartile range of 0.36 – 4.9 L/kg. (5B) The median clearance was 3.645 ml/min*kg, with an interquartile range of 1.2 – 11.7 ml/min*kg. (5C) The median half-life was 6.7 hours, with an interquartile range of 2.3 – 19 hours. (5D) The molecular weights ranged between 18 – 7183 g/mol, with an average molecular weight of 556.8 g/mol. (5E) The majority of drugs included in this study were delivered orally (78%), followed by intravenous delivery (15%). A small fraction of drugs were delivered by other methods (7%).



C

Pyrazinamide	Isoniazid	Nevirapine	Dexamethasone	Kanamycin	Cycloserine	Ethambutol
Zidovudine	Lamivudine	Adefovir dipivoxil	Emtricitabine	Efavirenz	Terizidone	Fosamprenavir
Atazanavir	Telbivudine	Acebrophylline	Etravirine	Darunavir	Ritonavir	Abacavir
Prothionamide	Ethionamide	Capreomycin	Maraviroc	Bedaquiline	Tenofovir disoproxil	Tenofovir Disoproxil Fumarate
Rilpivirine	Delamanid	Tenofovir alafenamide	Fostemsavir	Enfuvirtide	Cobicistat	Raltegravir
Cabotegravir	Dolutegravir	Doravirine	Entecavir	Rifampin	Rifapentine	Tipranavir

Figure 6: Drugs in the prediction dataset. A total of 42 drugs were included in the hold-out prediction dataset. (6A) Drugs used in the hold-out dataset ranged in molecular weight from 102 – 4492 g/mol, with an average molecular weight of 550 g/mol. (6B) Drugs in this dataset had indications for HIV (50%), Tuberculosis (36%), Hepatitis B (12%), or were multi-indicated for two or more of the previous diseases (2%). (6C) The full list of drugs used in this dataset.



A

Name	Test type	AUC/ROC	Notes
Clearance	Classification	0.59	Below 3.645 (median) = 1
Half-life	Classification	0.59	Above 6.7 (median) = 1
Vd	Classification	0.78	Above 1.25 (median) = 1
Combined	Classification		Used above cutoffs
Clearance		0.65	
Halflife		0.81	
Vd		0.82	
Variables Merged	Classification	0.75	Used above cutoffs Combined score 2+ = 1

	Predicted High	Predicted Low	Total
Actual High	5	6	11
Actual Low	2	10	12
Total	7	16	23

Fisher's exact test - p=0.193

Figure 7: Results of classification model training, testing, and predicting. ChemProp was used to create classification models to predict "high" versus "low" binary classes of the above pharmacokinetic parameters, where the cut-off for each parameter was the median value of the dataset. (7A) Each PK parameter were used separately and in combination to train various ChemProp models with the shown AUC/ROC scores. (7B) The results of improved separate scores when the PK parameters were used in a combined classification model. (7C) Volume of distribution binary score predictions (x-axis) were plotted against the documented experimental values of those same drugs from DrugBank (y-axis). (7D) A 2x2 contingency table of the data plotted in 7C.

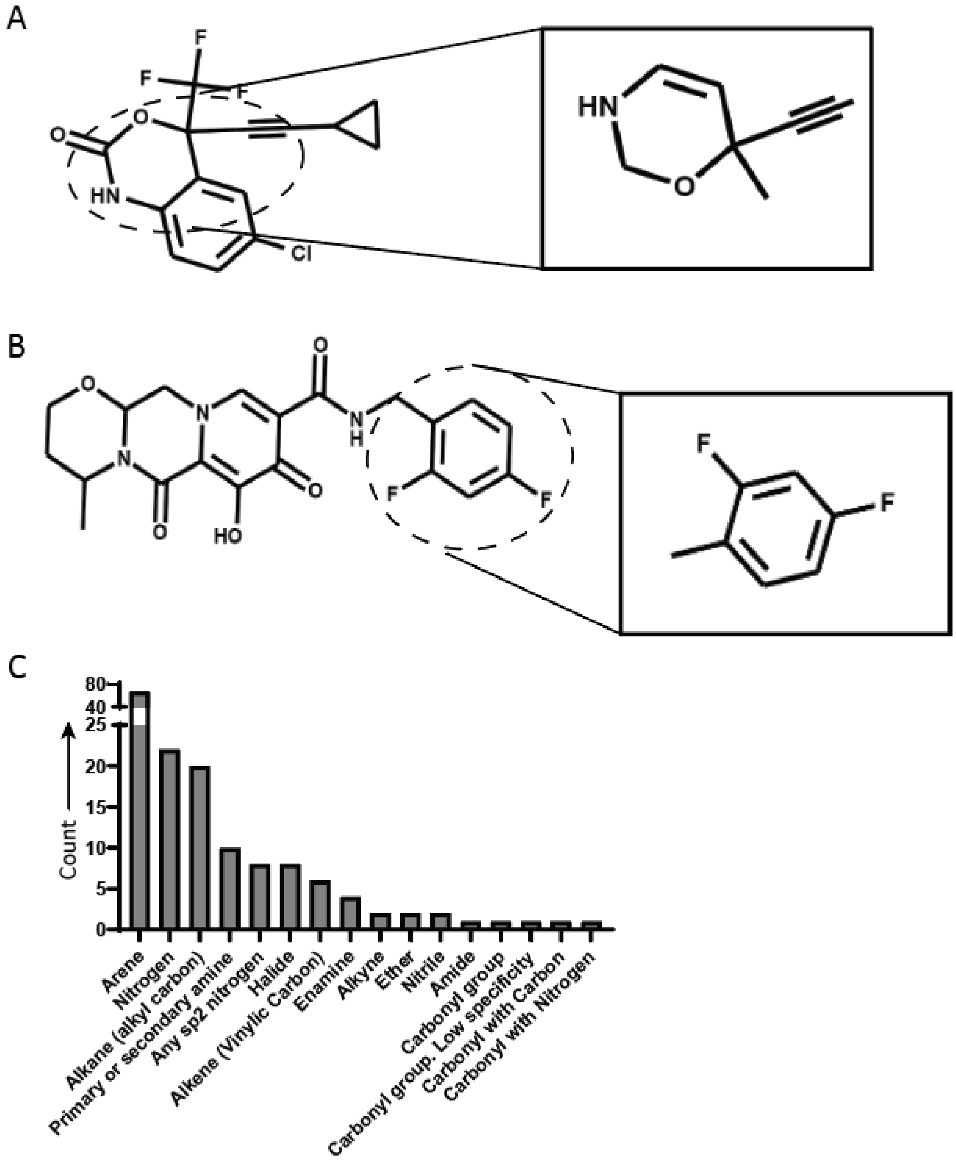


Figure 8: Substructure analysis of ChemProp combined score output. ChemProp interpret script was used to determine which substructures were used most heavily in the decision-making of the binary classification of the combined model. (8A-B) Example drugs with highlighted substructures that were key in the model deciding that these drugs had 2+ ideal PK parameters. (8C) A breakdown of the most common substructures and functional groups in the combined model.

Future Directions: Potential integration of this framework into ChemDB database

The third aim of this report states: *Develop a plan to integrate the relevant pharmacokinetic parameter(s) and possibly the model into ChemDB to provide wider access of these to the research community.* This section will be devoted to outlining an actionable plan to realize this aim, including the steps necessary and challenges which may arise during this process.

An overarching goal of this project was to provide wider access to pharmacokinetic information to the general research community primarily through developing plans for additions to ChemDB[31]. ChemDB is a database developed and maintained by the National Institute of Allergy and Infectious Diseases (NIAID), whose general purpose is to provide information regarding compounds with indications for HIV, HIV enzymes, and associated HIV co-morbidities. Many useful parameters are included for the research community on this website, including a drug structure manipulation GUI, information on drug properties such as Lipinski score, and related literature regarding the chemical compound. Additional information on drug parameters researched in this report would be helpful to include to the wider research community on this site, so that those in drug development, monitoring, or clinical trials might have more streamlined access to this information for their own projects. A major bottleneck to this strategy is that much of this information is often withheld from the general public and is considered proprietary information from the pharmaceutical companies with financial interests in the drug product. To address this, a two-fold approach to integrate PK parameter information into ChemDB is proposed:

1. For compounds with published PK parameters (on sites like PubChem and DrugBank): integrating this information directly into ChemDB in order to maximize the spread of information to the wider research community.
2. For compounds without published PK parameters: developing the ChemProp framework (such as that used in this report) and integrating it into ChemDB, in order to provide estimates of the PK parameters to the wider research community.

Part 1: Integration of existing, published PK parameters information into ChemDB

Objective: provide access to published data on volume of distribution, clearance rates, and half-life to the general research community using ChemDB.

Plan: work with federal agencies (National Library of Medicine: PubChem) or private companies (DrugBank) with published drug information to integrate this into ChemDB.

Challenges:

- Working with external companies like DrugBank may be heavily regulated by federal rules and regulations
- Developing the integration between these different external information sources into ChemDB may prove challenging to the website developers and maintainers

Benefits:

- May aid and facilitate research in the long-acting drug development community
- Increased transparency and access to information may boost confidence in federal activities regarding long-acting formulation development

Part 2: Development and integration of ChemProp framework into ChemDB

Objective: provide estimates of key pharmacokinetic parameters to the research community, with the goal of aiding in the decision-making process of long-acting formulation drug development

Plan: Integrate ChemProp model (or other deep neural network) into ChemDB with the express goal of providing estimates of key PK parameters. This could either integrate a classification prediction (i.e. compound is likely to have clearance rate above X value) or regression prediction (i.e. compound is likely to have approximate clearance rate of X mL/min*kg).

Challenges:

- In order to achieve better prediction accuracy, would need more data for model training/testing/validation
- Would need to decide whether classification or regression would better serve the research community
- May prove a challenge to the developers to integrate this model into ChemDB

Benefits:

- Only SMILES information from compounds is needed to create the model, which is already provided on ChemDB
- Provide information to those outside the proprietary bounds of pharmaceutical company access which may boost research in long-acting drug area

In conclusion, this report has demonstrated that deep neural networks such as ChemProp can be used to predict molecular properties from structural information alone. Even using a small dataset, ChemProp demonstrated its utility by being able to predict volume of distribution classifications with up to 78% accuracy. With the addition of more robust and accurate data, future models can be trained with higher precision and accuracy than this pilot study. With the overarching goal of increased transparency in the scientific community, we hope that this framework can be used in the future to provide key information for those developing drugs with long-acting potential across a wide spectrum of indications, and particularly within the area of HIV and HIV co-morbidities research.

References

1. HIV/AIDS. https://www.who.int/news-room/fact-sheets/detail/hiv-aids.
2. Chang, C. C. *et al.* HIV and co-infections. *Immunol Rev* **254**, 114–142 (2013).
3. Economically Disadvantaged | HIV by Group | HIV/AIDS | CDC. https://www.cdc.gov/hiv/group/poverty.html (2022).
4. Report on the global AIDS epidemic. https://www.unaids.org/en/resources/documents/2008/20081107_jc1510_2008globalreport_en.pdf.
5. Ammon, N., Mason, S. & Corkery, J. M. Factors impacting antiretroviral therapy adherence among human immunodeficiency virus–positive adolescents in Sub-Saharan Africa: a systematic review. *Public Health* **157**, 20–31 (2018).
6. Masters, M. C., Krueger, K. M., Williams, J. L., Morrison, L. & Cohn, S. E. Beyond one pill, once daily: current challenges of antiretroviral therapy management in the United States. *Expert Rev Clin Pharmacol* **12**, 1129–1143 (2019).
7. Flexner, C., Owen, A., Siccardi, M. & Swindells, S. Long-acting drugs and formulations for the treatment and prevention of HIV infection. *International Journal of Antimicrobial Agents* **57**, 106220 (2021).
8. Long-Acting Cabotegravir and Rilpivirine after Oral Induction for HIV-1 Infection | NEJM. https://www.nejm.org/doi/full/10.1056/NEJMoa1909512.
9. Long-Acting Cabotegravir and Rilpivirine for Maintenance of HIV-1 Suppression | NEJM. https://www.nejm.org/doi/full/10.1056/NEJMoa1904398.
10. Scarsi, K. K. & Swindells, S. The Promise of Improved Adherence With Long-Acting Antiretroviral Therapy: What Are the Data? *J Int Assoc Provid AIDS Care* **20**, 23259582211009012 (2021).
11. Phillips, A. N. *et al.* The potential role of long-acting injectable cabotegravir–rilpivirine in the treatment of HIV in sub-Saharan Africa: a modelling analysis. *The Lancet Global Health* **9**, e620–e627 (2021).
12. Long-Acting HIV Drugs for Treatment and Prevention | Annual Review of Medicine. https://www.annualreviews.org/doi/10.1146/annurev-med-041217-013717.
13. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* **18**, 463–477 (2019).
14. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
15. Atkinson, A. J. Chapter 2 - Clinical pharmacokinetics. in *Atkinson's Principles of Clinical Pharmacology (Fourth Edition)* (eds. Huang, S.-M., Lertora, J. J. L., Vicini, P. & Atkinson, A. J.) 11–26 (Academic Press, 2022). doi:10.1016/B978-0-12-819869-8.00021-5.
16. Interpretation — chemprop 1.5.2 documentation. https://chemprop.readthedocs.io/en/latest/interpret.html.
17. Appendix F SMILES Notation Tutorial. 4 (2012).
18. Pham, H. & Le, T. Attention-based Multi-Input Deep Learning Architecture for Biological Activity Prediction: An Application in EGFR Inhibitors. in 1–9 (2019). doi:10.1109/KSE.2019.8919265.
19. Mauri, A., Consonni, V., Pavan, M. & Todeschini, R. DRAGON software: An easy approach to molecular descriptor calculations. *MATCH Communications in Mathematical and in Computer Chemistry* **56**, 237–248 (2006).
20. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

21. Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
22. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688-702.e13 (2020).
23. PubChem. PubChem. https://pubchem.ncbi.nlm.nih.gov/.
24. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2018).
25. FDA-Approved HIV Medicines | NIH. https://hivinfo.nih.gov/understanding-hiv/fact-sheets/fda-approved-hiv-medicines.
26. List of drugs/medicine used for Tuberculosis (Tuberculosis). *Medindia* https://www.medindia.net/drugs/medical-condition/tuberculosis.htm.
27. Hepatitis B Foundation: Approved Drugs for Adults. https://www.hepb.org/treatment-and-management/treatment/approved-drugs-for-adults/.
28. (PDF) Metabolic profiling of human blood. https://www.researchgate.net/publication/276079439_Metabolic_profiling_of_human_blood.
29. Lipinski's Rule of Five - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/pharmacology-toxicology-and-pharmaceutical-science/lipinskis-rule-of-five.
30. Greenblatt, D. J. Elimination Half-Life of Drugs: Value and Limitations. *Annual Review of Medicine* **36**, 421–427 (1985).
31. Division of AIDS Anti-HIV/OI/TB Therapeutics Database - Simple Search Page. https://chemdb.niaid.nih.gov/.