International Conference on Intelligent Technologies & Science - 2022
(ICITS-2022)

Implementation of Machine Learning in Drug Discovery and Drug Design

Arvind Kumar[1], Sushil Kumar Maurya[2]
Computer science and Engineering Department,
Bansal Institute of Engineering and Technology, Lucknow
Arvindrajbhar323@gmail.com

Abstract: We have developed a simple medical application relevant machine learning model for enzyme inhibition parameter search for normal and impaired kidney function for two drugs. The machine learning model has been developed in the MATLAB platform to accelerate model reuse by other drugs discovery application. With the speed of the model solution, the real-life pharmaceutical applications, and the GUI, the machine learning model can be readily used in educational contexts as a dynamic, interactive visualization tool for learning about biomedical and chemical application of computer science and engineering. The model and results of this work will enable further studies on the impacts of inhibition of enzymes in disease states such as kidney infection.

Keywords: Artificial intelligence, Drug discovery, Machine learning, Target validation.

1. Introduction:

Drugs are key therapeutic treatments for a wide variety of medical conditions and are an essential building block of a functioning health system (World Health Organization, 2010). However, there are many medical needs, both existing and emergent, that are currently unmet by existing medicines (Kaplan et al., 2013). The ability to rapidly and effectively address unmet medical needs as and when they occur has been further highlighted by the recent worldwide emergency caused by the current coronavirus pandemic (COVID-19, Rosa et al., 2020). Developing new therapeutics is an extremely challenging, multi-stage process, involving many disciplines and typically requiring many years to complete. On average each new therapeutic is estimated to cost $1.5-3 billion, depending on how this is calculated, (Avorn, 2015; DiMasi et al., 2016) and takes over ten years (Paul et al., 2010). On average, the FDA has approved 31 novel drugs per year from 2008-16 (U.S. Food and Drug Administration, 2018a). These figures are not improving and, as such, current practices have been called unsustainable (Moors et al., 2014; Ernst & Young, 2017). Much of the cost of drug discovery arises from the high chance of failure, with the investment of sufficient time and financial resource far from a guarantee of success. A recent study found that only 13.8% of all drug development programs eventually lead to approval while the overall success rate for drugs that treat rare diseases, also known as 'orphan drugs', was as low as 6.2% (Wong et al., 2018). The high cost and low productivity in drug development is a long-standing problem, for which a solution is critically important (Myers and Baker, 2001). Computer-aided drug design (CADD) is seen as having the potential to accelerate this process and reduce the expense of developing new therapeutics (Ou-Yang et al., 2012). However, despite broad adoption of computational methodologies across the entire drug discovery workflow, costs have continued to increase (DiMasi et al., 2003; Avorn, 2015; DiMasi et al., 2016) with sustained low productivity (Khanna, 2012). New techniques and approaches are still sorely needed to revolutionise drug discovery. Recently, there has been renewed interest in the use of artificial intelligence across a broad range of fields, driven by the rise of deep learning. While many of the core principles of deep learning were introduced decades ago (e.g. Rosenblatt, 1958; Fukushima, 1980; Rumelhart et al., 1986), it was not until 2012 that the power and effectiveness of such techniques was demonstrated, in what is now known as the "ImageNet moment". In the annual ImageNet Large Scale Visual Recognition Challenge, Krizhevsky et al. (2012) performed 41% better than the next best competitor by adopting a deep neural network. It is widely acknowledged that this breakthrough was made possible by the combination of unprecedented availability of labelled data and computational power. This has led to learningbased systems matching, and indeed often surpassing, humans at image recognition (He et al., 2015), single-player games (Mnih et al., 2015), and two-player games including Go (Silver et al., 2016; Silver et al., 2017), Chess (Silver et al., 2018), and StarCraft II (Vinyals et al., 2019).

These advances quickly caught the attention of the field of cheminformatics, with several early promising results reported. In 2013, deep neural networks were the best performing models in the Merck molecular activity challenge (Ma et al., 2015) while a similar result was obtained in the Tox21 toxicity data challenge in 2015 (Mayr et al., 2016). Learning-based algorithms have a long history in drug discovery. Early quantitative structure activity relationship (QSAR) models were first described in the early 1960s (Hansch et al., 1962) and have become commonplace (Salt et al., 1992). However, traditional machine learning and classical statistical approaches typically require the explicit featurisation of the target input, such as molecules or protein-ligand complexes, in the form of a one-dimensional vector (Klambauer et al., 2019). This requirement has resulted in the development of hundreds of descriptors for molecular property prediction alone (e.g. Deng et al., 2004; Zhang et al., 2006; Durrant and McCammon, 2011). However, an advantage of deep learning methods that is seen as key to their success is the ability to eliminate the need for abstraction and allow a

much broader variety of data types be learnt from directly (Klambauer et al., 2019).

Finally, the QSAR models discussed above are typically bespoke models, constructed within the context of a specific drug discovery project based on a small amount of data. Thus, while useful, they do not have general applicability, and often do not extend beyond a specific chemical series. Success in other domains (e.g. ImageNet, Deng et al., 2009) has shown that a key requirement for generalpurpose models is sufficient data (Halevy et al., 2009; Sun et al., 2017). Over the past decade, there has been a rapid increase in the amount of publicly available molecular activity and biochemical data (e.g. Kim et al., 2015; Papadatos et al., 2015), as well as structural data (Berman et al., 2000; Burley et al., 2019), largely due to increased focus and the emergence of new experimental techniques (e.g. high-throughput screening, Inglese et al., 2007). An example of the benefits of the availability of such data that would not have been possible otherwise are the recent successes in the field of protein-structure prediction, culminating in the performance of AlphaFold (Senior et al., 2020) and AlphaFold 2 (Jumper et al., 2020) in CASP 13 and 14, respectively (Kryshtafovych et al., 2019).

2. Methodology:

Angiotensin II is a hormone that regulates the blood pressure by constricting the blood vessels. A sustained high level of Ang II leads to hypertension, which increases the work required for the heart to pump blood through the body and causes numerous health problems. Angiotensin converting enzyme (ACE) inhibitors are a class of pharmaceuticals that inhibit the production of Ang II from angiotensin I (Ang I), thereby reducing the Ang II concentration and thus lowering the blood pressure. The conversion of Ang I to Ang II is a crucial step in the biochemical reaction network called the renin-angiotensin system (RAS) (Fig. 1). Ang II concentration has a negative feedback effect on the production of renin, which catalyzes the upstream production of Ang I from angiotensinogen (AGT). Here, we consider the subset of the RAS reaction network that includes the hormones and the enzymes affected directly by ACE inhibition (Fig. 1).

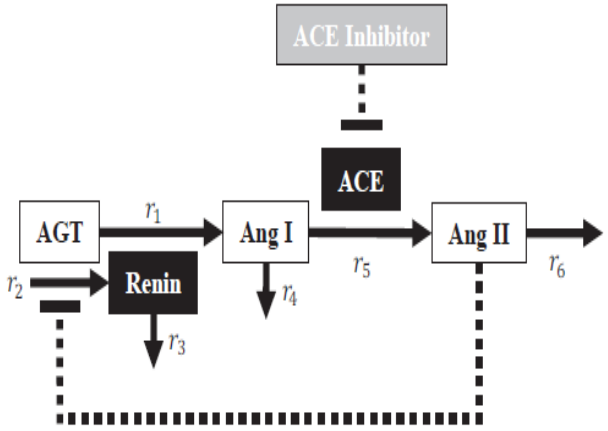

Fig 1: The reaction network for the renin-angiotensin system.

ACE inhibitors are known to aid patients with hypertension, congestive heart failure, and chronic kidney disease (CKD) (Balfour and Goa, 1991; Brown and Vaughan, 1998; Corbo et al., 2016). CKD is a severe complication of diabetes and a significant factor leading to morbidity and mortality in both type I and type II diabetic patients. CKD is the primary cause for end-stage renal disease and can lead to kidney failure if Ang II is not regulated. It is important to slow the rate of progression of CKD before irreversible kidney damage occurs, and ACE inhibitors have been shown to be effective at doing so (Asher and Murray, 1991; Hoyer et al., 1993; Brown and Vaughan, 1998; Hsu et al., 2014; Yamout et al., 2014). As both diabetes and chronic kidney disease are comorbidities of hypertension, ACE inhibitors are particularly attractive pharmaceuticals for targeting multiple indications simultaneously. Many ACE inhibitor pharmaceuticals are on the market. The ACE inhibitors benazepril and cilazapril are selected for this study because both are renoprotective and used to treat CKD and hypertension (Hoyer et al., 1993; Niu et al., 2014). Also, studies have been conducted to collect data on benazepril and cilazapril drug dose and effects on the RAS for hypertensive human patients with normal renal function (NRF) and impaired renal function (IRF) (Shionoiri et al., 1988, 1992; Kloke et al., 1996). The model and parameter estimation techniques in the present work could be readily adapted for other ACE inhibitor drugs if pharmacokinetic and pharmacodynamic experimental stud ies have been performed to characterize the actions of those drugs. Note that benazepril and cilazapril are both extensively bioactivated to diacid chemical forms, as is common for most ACE inhibitors (Hoyer et al., 1993; Toutain and Lefebvre, 2004; LeBlanc et al., 2006). Therefore, all parameters and calculations involving either drug only use the pharmacologically active diacid form of the corresponding drug.

3. Machine Learning

Machine learning attempts to learn patterns directly from data without explicit functional pre-specification for use in prediction, decision making, or other out comes of interest (Mitchell, 1997; Murphy, 2012). These methodologies are often classified into several paradigms: supervised learning, unsupervised learning, and reinforcement learning. However, these paradigms are not mutually exclusive, and there are many connections between them.

In supervised learning we are interested in fitting a function f : $X \rightarrow Y$ using a data set, D, of n labelled observations

$$D = \{(x_i, y_i), i = 1 \ldots n\}$$

where $x_i \in X$ and $y_i \in Y$. Typical applications include regression and classification tasks.

In unsupervised learning, we do not have access to labels and thus our data set, D, consists of only observations of the source domain X, reducing to

$$D = \{x_i, i = 1 \ldots n\}$$

In this paradigm, the goal is to find some notion of internal structure or common featurisation. Clustering (Lloyd, 1982),

anomaly detection (Hodge and Austin, 2004), and dimensionality reduction techniques (Maaten et al., 2007) are common unsupervised methodologies.

Finally, reinforcement learning aims to learn an optimal policy for an agent in an environment, given some notion of reward. While many formulations exist, a basic and natural one is

$$D = \{(si, ai, ri), i = 1 \ldots n\}$$

where $s_i \epsilon S$ is the state of the system of environment, ai 2 A is the action taken by the agent, and ri 2 R is the reward given for taking action ai in state si. For the work presented in this thesis we mostly are concerned with the supervised and unsupervised paradigms, with a particular emphasis on deep learning-based models and their applications to drug discovery. In the remainder of this section, we will introduce two broad categories of machine learning algorithms that are applicable within any of the paradigms outlined above. The first, convolutional neural networks (CNNs), led to the "ImageNet moment" discussed previously.

4. Result and Discussion:

In the experimental data used to fit the parameters, the dose for benazepril was 5 mg and that for cilazapril was 1.25 mg. The following parameters for the PD model were estimated for four cases (benazepril and cilazapril for both normal function (NRF) and for impaired function (IRF): Vmax/KM, kR, kf, kAII, and f. The values reported for the fitted parameters were the median of the best-fit parameter sets along with the minima and maxima of the ranges of parameters (Table 5 for benazepril and Table 6 for cilazapril). The best-fit parameter sets were those with WSSR within 1% of that for the single best set (74, 93, 7, and 93 out of 101 multistart parameter sets for benazepril NRF, benazepril IRF, cilazapril NRF, and cilazapril IRF, respectively). The 95% prediction confidence intervals were determined using kernel density estimation with the best-fit parameter sets. The simulation results for the fitted parameters for all four cases are shown for output variables CAI, CAII, and PRA, respectively. Data obtained from (Shionoiri et al., 1992, 1988) are also shown in the figures.

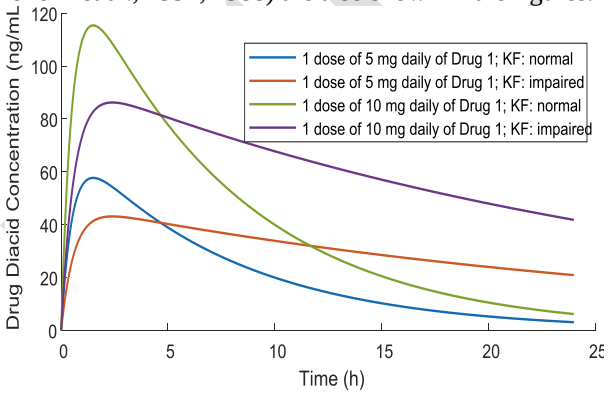

Fig 2: Benazepril validation results: diacid form of benazepril concentration versus time after a single dose for 5 mg and 10 mg doses for NRF and IRF.

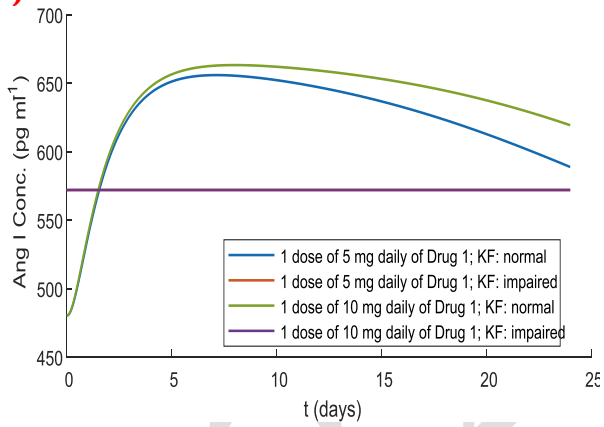

Fig 3: Benazepril validation results for doses of benazepril for NRF: Ang I concentration

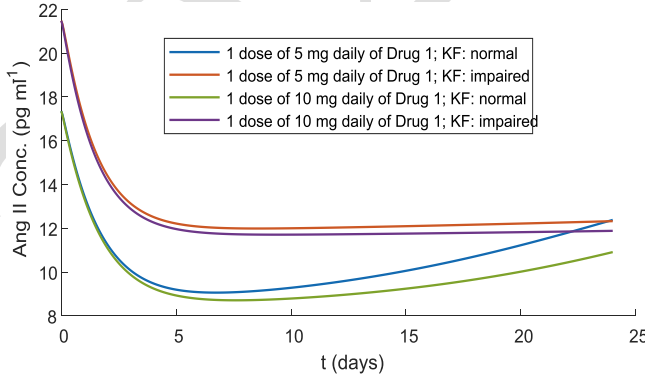

Fig 4: Benazepril validation results for doses of benazepril for Ang II concentration as functions of time

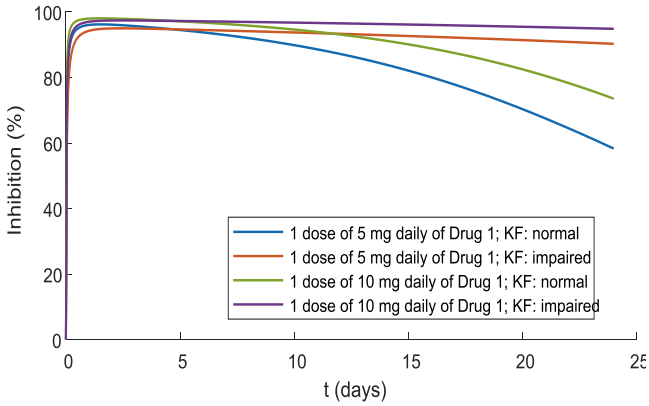

Fig 5: Benazepril validation results for doses of benazepril for Ang II concentration as functions of time

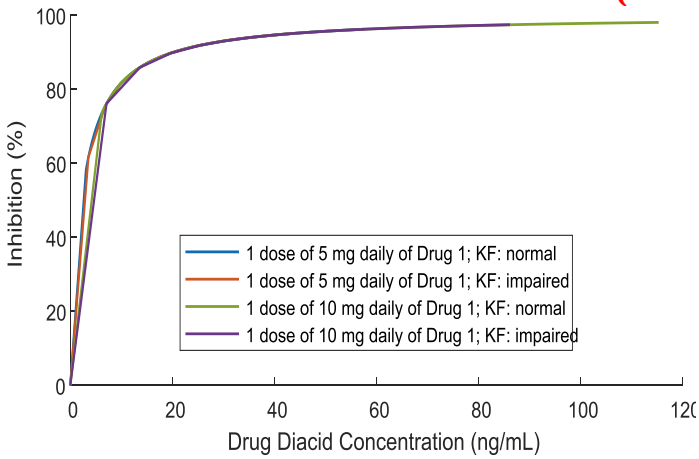

Fig 6: Validation results for doses of benazepril for inhibition as a function of drug diacid concentration

5. Conclusion:

We have developed a simple physiologically relevant PK/PD model for ACE inhibition parameterized for normal and impaired renal function for two drugs. The model has also been packaged as a MATLAB app to accelerate model reuse by other researchers and/or educators. With the speed of the model solution, the real-life pharmaceutical applications, and the MATLAB GUI, the model can be readily used in educational contexts as a dynamic, interactive visualization tool for learning about biomedical and chemical engineering. The model and results of this work will enable further studies on the impacts of ACE inhibition in disease states such as CKD. In the future, the predictions of systemic hormone levels for Ang I and Ang II could be fed into clinically relevant multiscale models of local tissue effects in microvasculature complications of CKD, diabetes, or hype.

References:

[1] Collins FS, Varmus H. A new initiative on precision medicine. New England J Med 2015;372(9):793–5.

[2] Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486(7403):346–52.

[3] Romond EH, Perez EA, Bryant J, Suman VJ, Geyer Jr CE, Davidson NE, Tan-Chiu E, Martino S, Paik S, Kaufman PA, et al. Trastuzumab plus adjuvant chemotherapy for operable her2-positive breast cancer. N Engl J Med 2005;353(16):1673–84.

[4] Blanco JL, Porto-Pazos AB, Pazos A, Fernandez-Lozano C. Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection. Sci Rep 2018;8(1):1–11.

[5] Munteanu CR, Fernández-Blanco E, Seoane JA, Izquierdo-Novo P, Angel Rodriguez-Fernandez J, Maria Prieto-Gonzalez J, Rabunal JR, Pazos A. Drug discovery and design for complex diseases through qsar computational methods. Current Pharmaceutical Des 2010;16(24):2640–55.

[6] García I, Munteanu CR, Fall Y, Gómez G, Uriarte E, González-Díaz H. Qsar and complex network study of the chiral hmgr inhibitor structural diversity. Bioorganic Med Chem 2009;17(1):165–75.

[7] Liu Y, Tang S, Fernandez-Lozano C, Munteanu CR, Pazos A, Yu Y-Z, Tan Z, González-Díaz H. Experimental study and random forest prediction model of microbiome cell surface hydrophobicity. Expert Syst Appl 2017;72:306–16.

[8] Riera-Fernández P, Munteanu CR, Dorado J, Martin-Romalde R, Duardo- Sanchez A, Gonzalez-Diaz H. From chemical graphs in computer-aided drug design to general markov-galvez indices of drug-target, proteome, drugparasitic disease, technological, and social-legal networks. Current Computeraided Drug Des 2011;7(4):315–37.

[9] Shirvani P, Fassihi A. Molecular modelling study on pyrrolo [2, 3-b] pyridine derivatives as c-met kinase inhibitors, a combined approach using molecular docking, 3d-qsar modelling and molecular dynamics simulation. Mol Simul 2020:1–16.

[10] B. Suay-Garcia, J.I. Bueso-Bordils, A. Falcó, M.T. Pérez-Gracia, G. Antón-Fos, P. Alemán-López, Quantitative structure–activity relationship methods in the discovery and development of antibacterials, Wiley Interdisciplinary Reviews: Computational Molecular Science e1472.

[11] Fernandez-Lozano C, Gestal M, Munteanu CR, Dorado J, Pazos A. A methodology for the design of experiments in computational intelligence with multiple regression models. PeerJ 2016;4:e2721.

[12] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al., Drugbank 5.0: a major update to the drugbank database for 2018, Nucleic acids research 46 (D1) (2018) D1074– D1082..

[13] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, et al. Pubchem 2019 update: improved access to chemical data. Nucleic Acids Res 2019;47(D1):D1102–9.

[14] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, et al. Chembl: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40(D1): D1100–7.