FEATURIZATION OF DRUG COMPOUNDS AND TARGET PROTEINS FOR DRUG-TARGET INTERACTION PREDICTION

[1]Ebenezer Nanor, [2]Victor K. Agbesi, [3]Wei-Ping Wu, [4]Brighter Agyemang

School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, P. R. C.
Email: {[1]sireben21, [2]victoragbesivik, [4]brighteragyemang}@gmail.com, [3]wei-ping.wu@uestc.edu.cn

DOI: 10.29322/IJSRP.10.02.2020.p9813
http://dx.doi.org/10.29322/IJSRP.10.02.2020.p9813

Abstract- In recent times, computational methods for Drug-Target Interaction (DTI) prediction have become pervasive due to the advances in computational resources and techniques. Computational methods, particularly Machine Learning (ML) based methods, are very efficient DTI predictors. They help greatly in reducing the search space of drug candidates, consequently reducing the cost and time for drug development. Featurization is an important task in machine learning-based DTI prediction methods. Descriptors of drug compounds and target proteins are computed with available Cheminformatics software packages and fed as inputs to ML models for DTI prediction. The process of featurization poses a challenge to non-experts with little or no knowledge of molecular fingerprints, descriptors, and computational toolkits in the domain of Cheminformatics. In this study, we aim to provide reference and tutorial insights into featurization of drug compounds and proteins for DTI prediction. Firstly, we present an overview of DTI prediction encompassing definitions of key terms, various categories of computational methods, and bioactivity repositories. Subsequently, we discuss feature extraction, drug compound descriptors, protein descriptors, and Cheminformatics toolkits. There is a demonstration of how Cheminformatics toolkits can be used for featurization and a review of recent works that constructed compound and protein feature vectors for DTI prediction. Lastly, we give a summary of the study and highlight the challenges of featurization. This is the maiden work that discusses featurization of compounds and proteins for DTI prediction and gives practical examples of how it is done.

Index Terms- Drug discovery, Drug repurposing, Drug-target interaction, Featurization, Machine learning

INTRODUCTION

The discovery and development of chemical synthetic drugs is challenging, risky, and cumbersome as it involves a number of delicate stages, namely, target identification and selection; target validation; lead compound identification; hit lead optimization; clinical evaluation. Statistics collected from the 50 leading pharmaceutical companies over a decade and reported in [1] shows that only 19% of developed drugs are approved onto the market. Because of the high attrition risk associated with the development of new drugs, pharmaceutical companies and researchers have become more interested in identifying other uses or indications for drugs that are already on the market. Moreover, post-market studies of commercial drugs reveals the propensities of drugs to interact with multiple targets. These challenging circumstances have brought about drug repurposing- a new direction in drug development. Drug repurposing tends to find new targets for an existing drug by screening the drug against targets to determine interactivity. As an example, thalidomide manufactured and used in countries like Germany and England as a medication for morning sickness

has now been repurposed for treatment of leprosy [2]. Sometimes, a drug that has failed clinical evaluation can also be repurposed for the treatment of a different kind of disease. Generally, this form of drug repurposing/ repositioning is referred to as drug rescue, and it is normally considered in cases of life-threatening diseases. It is worth mentioning that drug repurposing cannot always be a resort to finding a suitable drug for a disease. At times, there will be no existing drug that would counteract the activities of the disease-causing target. In situations like this, the complete process of drug discovery and development needs to be initiated. It is obvious that the primary task in repurposing and discovery of drugs is Drug-Target Interaction (DTI) prediction.

DTI prediction involves screening known drug compounds against large pools of targets in the quest of novel targets that bind effectively with any drug compound to develop a therapeutic effect. Reversely, DTI prediction activity screens drug compounds against a defined target to find drug compounds that develop desired therapeutic effect upon binding with the target. Although DTI prediction finds new indications for approved drugs, it sometimes brings to light certain off-targets interactions which cause undesirable side effects when drugs are used [3]. Figure 1 depicts a drug's interaction with intended targets and an off-target.

In this study, we provide a comprehensive overview of DTI prediction with narrowed focus on how drug compounds and target proteins are featurized for the purpose of feature-based DTI predictions. The organization of the study is as follows. Section 1 contains introduction, defines relevant basic terminologies related to DTI prediction and discusses DTI prediction in detail. Furthermore, it tabularizes extant databases, together with their characteristics. Section 2 provides an insight into extraction of features for feature-based machine learning DTI prediction method. We enumerate and throw more light on representations and descriptors that exist for drug compounds and target proteins in this section. Section 3 highlights available software packages used for generating descriptors and mentions the dependencies of these packages. Section 4 contains the main idea of this study, i.e., it describes vividly how compound and protein features are computed using examples of the software packages mentioned in Section 3 of this paper. Furthermore, it looks at existing DTI prediction works that constructed and used features of drug compounds, target proteins, and drug-target pairs as inputs to their models. Section 5 summarizes and concludes the paper by looking at the challenges associated with learning featurization and the application of it for DTI predictions. In the section, an observation about the use of complex methods for featurization is made.

The basic terminologies in DTI prediction as considered in this study are discussed below:

Drug: It is any substance which alters the normal functioning of the body mentally, physically or emotionally [4], [5]. In pharmacology, a drug is a chemical substance that creates an organic impact [6] when directed to a living life form. Drugs work better by interacting with multiple targets simultaneously.

Target: A biomolecule in the body whose activity is modulated by a drug compound is known as a target. A drug would interact with the active site of a target to elicit some therapeutic effect. Enzyme(s) is a typical target for therapeutic mediation and so many well-studied examples exist. Outdated or traditional chemical enzyme targets include phosphatases, phosphodiesterase's, proteases, and kinases. Proteins, G-Protein–Coupled Receptors (GPCRs), and ion channels are other examples of drug targets.

To drug discovery, the detection of interaction between drug compounds and targets plays a key role. Drugs usually interact to perform their roles with one or more targets. Drug-target interaction refers to the binding of a drug to a particular location in a target that results in a change(s) in the functions of the target [7]. Intuitively, DTI prediction helps in finding drug compounds and targets that interact well with each other for validating experiments to be carried out on them. It is performed in the third stage of the drug discovery and development process (i.e., lead identification stage) after the identification and validation of a target of interest. Several compounds are screened against validated target to ascertain potential compounds that will exhibit good therapeutic effect upon their interaction with the target. Nowadays, research in DTI prediction has been considered very serious to find the right medication for the variety of incurable diseases that are in existence [8], [9]. Apart from the development of novel drugs, DTI prediction has other useful applications in poly-pharmacology, drug resistance, side-effect prediction, drug repositioning [10], protein subcellular location prediction [11], disease-related miRNA prediction [12], and protein-protein interaction prediction [13].

Before the last two decades, traditional methods like in vitro experiments were relied on to detect DTIs. These methods are overly expensive, time consuming, labor intensive, and can only be considered for small-scale development of drugs. Advances in computational resources (e.g. high performing processors and GPUs), techniques, and bioactivity data collections have provided in silico alternatives to in vitro methods [14]. A well-used in silico (computational) method for DTI prediction in recent times is Virtual Screening (VS). The past 4 years have seen lots of Machine Learning (ML) and Deep Learning (DL) techniques being used in virtual screening to improve the efficiency and precision of DTI models.

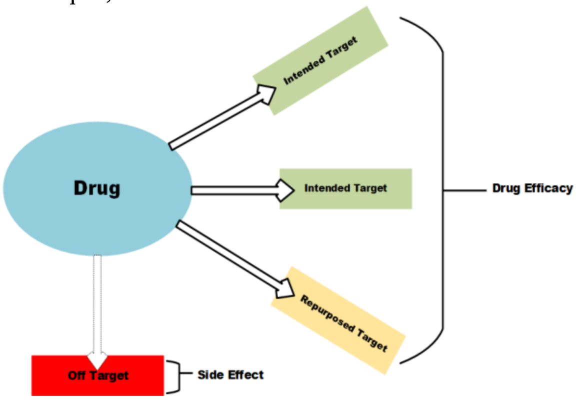

Figure 1: Schematic representation of Drug Target Interaction showing drug repurposing and side effect development

Generally speaking, in vitro experiments and in silico methods play complementary role to each other in DTI predictions. In order to facilitate the development of a drug, in silico methods can be used to quickly make suggestions of potential compounds that interact with a given target of interest, thereby reducing the scope of compound search. In vitro experiments are then performed to verify and validate selected compounds before they are being optimized. So clearly, computational methods cannot take the place of in vitro experiments completely.

Computational methods provide a "virtual shortcut" to identifying lead molecules that bind well with a pre-specified target protein, thereby reducing the time and cost of developing drugs. Furthermore, they allow both a greater understanding of complex biological interactions and essential biological processes, as well as speeding up the discovery of new drugs and improving human medicine. The computational approaches to predicting DTIs can be grouped into three major categories:

1. Ligand-based methods are non-structure-based virtual screening methods. A ligand is an electron(s) donating atom or molecule. Ligand-based methods make use of information such as molecular properties of a ligand to predict activity/interaction based on its similarity or dissimilarity to previously known active ligands. This is unlike structure based VS methods, which use structure information of both target and ligand [15] in their predictions. Some examples of ligand-based methods include ligand-based pharmacophores,

quantitative structure-activity relationships (QSAR), and molecular descriptors. Ligand-based screening methods have the advantage of identifying lead molecules from a family of active ligands based on a set of pharmacophore components. However, they perform poorly when there is insufficient number of identified ligands.

2. Docking methods are structure-based CADD methods. They depend on information about target structures to simulate and predict the binding affinities of compounds to the targets. A sizeable collection of ligands is docked into the approximated binding site of interest of a target protein and several poses of binding are evaluated using an energy scoring function. The poses are then ranked based on their binding energy scores. Poses with smaller scores are considered the best and are selected for further experimental verification. Some scoring functions used are Knowledge-based scoring function [16], [17], Force-field-based scoring function [18], Empirical scoring function [19], [20], and Consensus-based scoring function [21], which is a combination of two or more other scoring functions. Figure 2 is a schematic illustration of molecular docking. The method of docking a ligand into the protein's binding site involves studies about both the protein and its ligand's structure and chemistry.

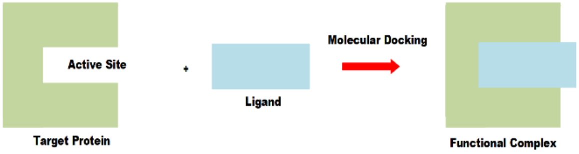

Figure 2: Schematic illustration of molecular docking

Docking methods are becoming less applicable and effective because there are so many proteins with unknown 3D structures. Moreover, membrane proteins [22] such as GPCRs and ion channels have complex structures, therefore it is impossible to obtain their 3D representations. In addition, docking simulations are cumbersome tasks; they take much time to be performed [8]. Some popular docking tools include Glide [23], Fred [24], AutoDock3 [25], AutoDock Vina [26], GOLD [27] and FlexX[28].

3. Chemogenomic methods are the latest computational approaches for predicting DTIs. They combine information about compounds (chemical) and targets (genomic) into a unified space to perform prediction. The purpose for the introduction of chemogenomic methods in DTI prediction is to address the challenges inherent in ligand-based and docking methods, as already highlighted in this section. In addition, chemogenomic methods allow for simultaneous screening of multiple compounds and targets. Current research and advanced tools in CADD have made provision for a wide variety of data and representations of compounds and targets. Chemogenomic methods [29] can further be categorized into network-based methods, graph-based methods, and machine learning-based methods. Figure 3 shows a hierarchical structure of computational methods for DTI prediction. The focus of our study will permit us to discuss machine learning-based methods only. [30] and [31] provide detailed information about network-based methods and graph-based methods, respectively.

Machine learning methods develop the most accurate results in the prediction of associations between compound-target pairs. In general, machine learning approaches can be grouped into supervised learning, semi-supervised learning, and unsupervised learning. Supervised learning methods for DTI predictions harness data sets consisting of positive and negative samples of compound-target interaction pairs. Samples that have been experimentally confirmed are regarded as positives, otherwise negatives. These positive and negative samples exist in an unbalance proportion in all bioactivity databases. Supervised learning algorithms applied in current works can be categorized into similarity-based approach [32], feature-based approach [7], and hybrid-based approach [33].

Feature-based approaches provide a numerical or quantized way to represent compounds, target molecules, and their pairings for DTI prediction tasks. Feature descriptors for drug compounds and targets are computationally transformed independently into feature vectors by the use of Cheminformatics tools or software packages. Basically, the generated feature vectors are combined programmatically by tensor product calculation or by using the concatenation function to form drug-target pair vector, which is then served as input to a machine learning model to learn and predict possible interactions between the pairs. All feature-vector inputs are required to be of a fixed length. Feature-based methods are very effective for task-specific learning as relevant features can be extracted and selected to ensure accurate predictive results. They can disclose prominent features of compounds and targets that contributed to effective DTIs, thereby making their results more interpretable. Despite these advantages, feature-based machine learning methods are disadvantaged in terms of model complexity tradeoffs, feature selection, class imbalance, high dimensionality of input vectors, and many more. Class imbalance is a disadvantage peculiar to feature-based classification methods.

1.1. Data Resource

Availability of data is critical to the development of machine learning-based models. Generally, models are trained on collected data samples and tested on new data samples to determine their performance in a given task.

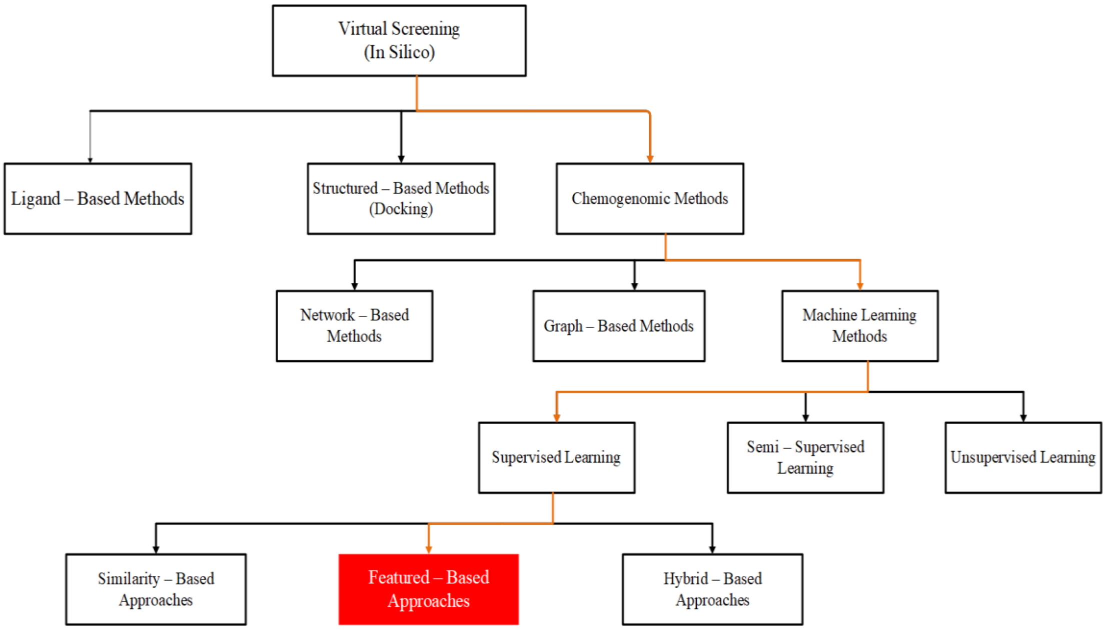
Virtual Screening (In Silico)
Ligand – Based Methods
Structured – Based Methods (Docking)
Chemogenomic Methods
Network – Based Methods
Graph – Based Methods
Machine Learning Methods
Supervised Learning
Semi – Supervised Learning
Unsupervised Learning
Similarity – Based Approaches
Featured – Based Approaches
Hybrid – Based Approaches

Figure 3: A hierarchical structure of recent computational methods in DTI prediction

Table 1. Bioactivity database: Name, link and description

Name	Link	Description
PubChem [34]	https://pubchem.ncbi.nlm.nih.gov	A chemical structure knowledge base which stores information about compounds, their features and activities.
ChEMBL [35]	http://www.ebi.ac.uk/chembl	An open access large-scale compound and bioactivity database of nearly 15 million experimentally derived bioactivities. It also includes specialized databases for specific diseases such as Neglected Tropical Disease (NTD) archive and malaria.
DrugBank [36]	http://www.drugbank.ca	A resource of approved and experimental drugs along with their targets
STITCH [37]	http://stitch.embl.de	A database of experimental and predicted interactions between chemicals and target proteins
KEGG [38]	http://www.kegg.jp	A knowledge base for understanding of biological systems from molecular level information of genes and genomes

PDB [39]	http://www.wwpdb.org/	An archive of curated and annotated 3D structural data of proteins, nucleic acids, and complex assemblies.
UniProtKB	http://www.uniprot.org	An online resource for retrieving information about target proteins.
BindingDB [40]	http://www.bindingdb.org/bind/index.jsp	An online database with 7225 protein targets; 621,060 compounds; 1,391,403 binding data. It also stores protein-ligand crystal structures and their binding affinities.
BioGRID [41]	https://thebiogrid.org/	A database of approved protein-protein interactions, chemical interactions, and genetic interactions.
BRENDA [42]	http://www.brenda-enzymes.org/	The main database for enzyme and enzyme-ligand data.
SIDER [43]	http://sideeffects.embl.de/	A database with information about commercial drugs and their side effects.
Pfam [44]	http://pfam.xfam.org/	A repository providing detailed information about curated protein families.
SuperTarget [45]	http://insilico.charite.de/supertarget/	A large database of drugs, proteins, drug-target interactions, protein-protein interactions, side effects.
SuperPred [46]	http://prediction.charite.de/	A repository of compound-target interactions.
MATADOR [47]	http://matador.embl.de/	An archive of automatic curated and manually annotated chemical compound-protein interactions and binding affinities.
TTD [48]	http://bidd.nus. edu.sg/group/cjttd	Therapeutic Target Database containing approved, clinical trial and validated targets and drugs. It also contains gene expression data.
GO database [49]	http://geneontology.org/page/go database	A structured database of GO ontologies and their respective annotated genes and gene products.
ASDCD [50]	http://asdcd.amss.ac.cn/	A database of reported synergistic antifungal drug combinations, indications, targets, etc.
DCDB [51]	http://www.cls.zju.edu.cn/dcdb/	An archive of known drug combinations.
Binding MOAD [52]	http://www.bindingmoad.org/	A database of high resolution protein-ligand structures and their binding data. It is known as the Mother of All Databases.

For DTI prediction, there exist various repositories with different kinds of information about compounds, targets, and their associations. These information are leveraged by predictive models to effectively predict new and unknown drug compound-target interactions. It is imperative to mention that databases such as Drugbank [36], STITCH [37], SuperTarget [45] should be considered when DTI prediction is modeled as a classification task. BindingDB [40] may be considered for regression, because it provides binding affinity measurement values between drug compound-target pairs. In Table 1, we listed quite a number of extant databases, their accessible links and sketchy descriptions.

FEATURE EXTRACTION

Feature extraction plays a vital role in leveraging feature-based machine learning approaches to perform predictions of possible interactions present among compound-target pairs. Feature extraction, also known as feature engineering, describes the mechanism for transforming arbitrary data like images, texts, etc., into feature vectors suitable for machine learning tasks. The process of feature extraction is difficult and time-consuming. Expert knowledge is often required to extract features that will give a better representation of data to be fed as input to a computer. Feature extraction for DTI prediction and other drug discovery related tasks relies on descriptors of compounds and targets. Descriptors enable the identification of compounds and targets stored in bioactivity databases by depicting their intrinsic chemical and physical properties. For example, a drug compound may be featurized based on a subset of descriptors available in its side effects descriptive data, where 0 or 1 respectively represents the absence or presence of the selected descriptors. In the process of feature extraction, valuable information about compounds and targets are loss when less and inappropriate descriptors are considered. There exist different types of drug compounds and targets, but this section will focus on descriptors for small molecules (compounds) and protein molecules since they form the majority of drug compounds and targets, respectively.

2.1. Small molecules descriptors

Existing two dimensional structures of small molecules are represented as SMILES [53], [54] and stored in the bioactivity databases for further computations. SMILES, which stands for "Simplified Molecular-Input Line-Entry System" is a string of texts used by chemists to concisely represent the substructures (atoms and bonds) of a molecule.

For instance, "OCCc1c(C)[n+](cs1)Cc2cnc(C)nc2N" is a SMILE notation for thiamine molecule. SMILES can easily be converted to molecular graphs, feature vectors or other representations by open source toolkits such as RDKit to be used as input for machine learning. In addition to SMILES, InChI (International Chemical Identifier) [54] and InChIKey line notations have also been released as representations for small molecules. InChI and its hash key version, InChIKey, are maintained by InChI Trust (http://www.inchi-trust.org) to allow easy search of chemical compounds in major search engines [15]. Mol2, SDF, CML are other formats for representing small molecules.

Given the various formats (SMILES, InChI, InChIKey, Mol2, SDF, and CML) for representing the chemical structures of molecules, researchers can compute their substructure descriptors or molecular descriptors. Molecular descriptors are numerical or feature vectors generated as a result of logic and mathematical operations performed on symbolic representations of small molecules. They are generated based on the structural, geometrical and physiochemical properties of molecules. A lot of different types of molecular descriptors exist in literature, but the most popular ones are fingerprints. Fingerprints are predefined features of drug compounds. They encode the properties: structural fragments, connectivity pathways, chemical bonds and functional groups of compounds as binary vectors, where 1 represents the presence of a particular property and 0 represents the absence of it. There are several types of fingerprints; each type represents different properties of compounds. Cited as an example, circular fingerprints like Extended Connectivity Fingerprints (ECFPs) represent the structural aspects of compounds; substructure keys-based like MACCs [55] represents pre-specified sub-structures in compounds. Other types of fingerprints are Estate fingerprints, Hybridization fingerprints, Lingo fingerprints, Pubchem fingerprints, Signature fingerprints, etc. Some variants of ECFP include ECFP_4 and ECFP_6. ECFPs are fast to compute [56]. They make it easy to compare molecules for similarity based on matching elements of the molecules' fingerprints. The main classes of molecular descriptors and their properties are discussed in Table 2.

Graph Convolutions offer an alternative to predefined features. Graph Convolution featurization is performed by deep learning models. As it is known, deep learning models try to learn features of input data themselves through training. Thus, deep learning approaches circumvent the process of feature engineering/extraction. Given a molecular graph as input, the convolutional neural network learns important features from the molecular graph to perform the task at hand. Other variations of graph convolutional networks are Message Passing Neural Networks (MPNN), Weave models, Deep Tensor Neural Networks (DTNN), etc. The greatest limitation of Graph convolutional networks is that they fail to make predictions in tasks that involve molecular conformations. They base their calculations exclusively on molecular graphs.

2.2. Protein descriptors

Proteins are compounds made up of hydrogen, oxygen, carbon, and nitrogen being arranged as chains of amino acids. The inclusion of nitrogen element in the composition of proteins uniquely distinguishes them from fats and carbohydrates. Amino acid compositions [57], dipeptide compositions [58], normalized Moreau-Broto autocorrelation [59], [60], Moran autocorrelation [61], Geary autocorrelation [62], and the CTD (Composition, Transition and Distribution) of structural and physicochemical properties [63]–[68] are some examples of protein descriptors. The individual descriptors can be combined in varying ways to form different sets of descriptors for different purposes such as predicting protein–protein interactions [69], protein functional classes [70], and protein structural classes [71]. Quasi sequence order [72] is a descriptor group formed by the combination of the weighted sums of physicochemical coupling correlations and amino acid descriptors. In studies [57], [70], there also exist Pseudo Amino Acid composition (PseAA) constituted by weighted sums of physicochemical square correlations and amino acid compositions. Ong et al. [73] combined all the individual and combined sets of protein descriptors mentioned in this section in their research, which aim to determine the most effective descriptor when it comes to predicting the families of proteins based on their functions. Their work gave the insight that right selection of descriptor sets for machine learning activities, generally, influences the performance of predictive models positively. Protein descriptors can be categorized as sequence-based and structure-based descriptors. Sequence-based target protein descriptors use protein amino acid sequence obtainable from UniProt database (http:/www.uniprot.org) [74], whiles structure-based protein descriptors use three dimensional protein atomic structures, which can be retrieved from the Protein Databank (PDB) (http://www.rcsb. org) [75] and other repositories.

LIBRARIES AND TOOLS FOR FEATURIZATION

Several software packages have been made openly available and accessible for computation of drug compound and target protein descriptors, among other Cheminformatics computations. They do not necessarily construct same dimensional feature vectors when used to featurize compounds and proteins, because they compute varied descriptor blocks of these molecules. Toolkits such as Rcpi [76] and PyDPI [77] can perform computation of both compounds and protein target descriptors (Table 6). Table 4 provides information on compound featurization tools and their computational dependencies. The table also provides access links to these resources. Similar information for protein featurization tools is contained in Table 5. Most of these tools or software do not have Graphical User Interface (GUI), therefore knowledge in programming is essential for their usage. Few ones that provide graphical interface are PaDEL [78], CDK Descriptor Calculator [79], Dragon [80], PROFEAT [81], ProtDCal [82], and ProtParam [83].

FEATURIZATION OF COMPOUNDS AND PROTEINS IN PRACTICE

In Section 1 and 3, the various databases and bioservices required to featurize chemical compounds and target proteins for DTI prediction have been extensively discussed. This section outlines the method of how some of these bioservices, namely, ProPy and PaDEL-descriptor can be used to featurize (compute descriptors) protein sequences and compound SMILES respectively.

4.1. Featurizing Target Proteins with ProPy

ProPy, which stands for Protein in Python, is a python software package used to compute the physicochemical and structural features of proteins from amino acid sequence. Users can specify the types of physicochemical and structural properties needed for featurization. ProPy is freely available and runs on Windows and Linux. The various groups of protein features and their corresponding numbers of descriptors that ProPy computes have been presented in Table 3.
We discuss below the procedures to compute or construct protein feature vectors using ProPy.

NB: A Python IDE is required for this exercise.
1. Download protein sequence data from UniProt (http://www.uniprot.org) or other related database.
2. Write a Python script to extract the protein IDs and sequences from the protein sequence data.
3. For easy featurization to be done, in a Python IDE, the python package – pandas can be used to handle the extracted IDs and sequences.
4. Import the installed propy library, and from propy import Pypro.
5. In a for loop, Pypro's GetProsDes method (GetProDes()) can then be used to get the descriptors of the protein sequences.
6. Compute the protein feature vectors using the desired descriptors and save them. Dimension of the feature vectors depends on the descriptors considered for the featurization. As an example, to compute feature vectors based on Protein Sequence Composition (PSC) descriptors (that is, Amino acid composition, Dipeptide composition, Tripeptide composition), use the functions: GetAAComp(), GetDPComp(), GetTPComp(). In the end, the constructed feature vectors will be of 8420 dimension.
7. The dimensionality of the protein feature vector can be reduced by selecting subset of the features, using whatever technique appropriate.

4.2. Featurizing Compounds with PaDEL-descriptor

PaDEL-descriptor is a stand-alone application written in Java. As already hinted, it is one of the few software packages that provides graphical interface to users for featurization. Users may create a configuration of the options and types of fingerprints and descriptors they want to calculate and save this configuration to a file for future featurization purpose. PaDEL-descriptor also provides a command line interface, which enables it to run in computer clusters. It runs on Windows, MAC OS, and Linux operating systems; and supports over 90 different compound molecular file formats. Some fingerprints generated by PaDEL include MACCS fingerprints; KlekotaRoth fingerprintsCount; Extented fingerprints; KlekotaRoth fingerprints; Substructure fingerprintsCount; AtomPairs2D fingerprintsCount; GraphOnly fingerprints; AtomPairs2D fingerprints; Estate fingerprints; Pubchem fingerprints; and Substructure fingerprints. The following steps below describe how PaDEL-descriptor can be used to featurize chemical compounds for machine learning-based DTI prediction.

1. Download PaDEL-Descriptor package from http://www.yapcwsoft.com/dd/padeldescriptor/, unzip and launch it using the executable jar file.
2. Download SMILES or any other file formats of compound molecules from any of the omic databases. PaDEL supports over 90 different molecular file formats.
3. At PaDEL's GUI, upload the downloaded molecular structural file and specify location and a file to save computed descriptors to.
4. Select from the available descriptors the ones to compute. The option "Fingerprints" may be selected if one wants to calculate fingerprints. Likewise the option "1D & 2D" and other options may be selected. Specification of the type(s) of fingerprints to compute can be made by clicking on the "Fingerprints" tab.
5. Click on the Start button to get the compound molecules featurized.

4.3. DTI related works that implemented featurization technique

Generally, the concept of transforming data of drug compounds and target proteins obtained from bioactivity databases into feature vectors for DTI predictions has been applied by many researchers. Yu et al. [84] featurized drugs and target proteins retrieved from DrugBank in their research to systematically predict multiple drug-target interactions from chemical, genomic, and pharmacological data. The drugs were featurized by computing a total of 1664 of their descriptors using the DRAGON software. But only 1080 of these descriptors were used for the actual work. Some of the descriptors computed include various topological, constitutional, molecular properties among other descriptors. Featurization of the proteins was done using the PROFEAT webserver. A feature vector of 1080 dimension was constructed as a result of computing the Moran autocorrelation descriptors, Dipeptide descriptors, and several other descriptors of the proteins. The constructed feature vectors of the drugs and proteins were combined and fed as input to Random Forest and SVM models for the prediction task. In study [3], drug-target pairs for DTI prediction were formed by concatenating corresponding feature vectors of drugs and targets. The drug features were generated by the use of Rcpi to compute constitutional, molecular properties, topological and geometrical descriptors of the drugs. On the other hand, PROFEAT web server was used to generate protein features by calculating descriptors related to dipeptide composition; amino acid composition; composition, transition and distribution; amphiphilic pseudo-amino acid composition; autocorrelation; quasi-sequence-order; and total amino acid properties. [85] used substructure fingerprints to construct a 881 dimensional drug feature vector of 0s and 1s.

Table 2. Drug compound descriptors: class, properties and fingerprints

Descriptor Class	Properties	Fingerprints
0D descriptors	Molecular weight Atom Number Atom – type count Other basic descriptors such as number of heavy atoms	
1D descriptors	Functional groups List of structural fragments Substituent atoms	
2D descriptors	Topological descriptors Graph invariants Graph-based substructures Connectivity bonds	1. Substructure keys based (e.g. MACCS) 2. Path based (e.g. Day Light and FP2) 3. Circular (e.g. ECFPs)
3D descriptors	Steric properties Geometrical molecular descriptors Surface area Volume Binding site properties 3D-based graph invariants	1. Geometrical (e.g. triangular descriptors) 2. Pharmacophore (e.g. hydrogen bond, hydrophobicity, charge and aromacity)
Non-structure-based molecular descriptors	Substring occurrence in SMILES Text-based molecular fingerprints ATC code annotations	1. LINGO descriptors

Table 3. List of protein feature groups, features and number of descriptors computed by ProPy

Feature groups	Features	No. of descriptors
Amino acid composition	Amino acid composition Dipeptide composition Tripeptide composition	20 400 8000
Autocorrelation	Normalized Moreau – Broto autocorrelation Moran autocorrelation Geary autocorrelation Geary autocorrelation	240 240 240
Composition, transition and distribution	Composition Transition Distribution	21 21 105
Quasi-sequence order	Sequence-order-coupling number Quasi-sequence-order descriptors	60 100
Pseudo-amino acid composition	Type I pseudo-amino acid composition Type II pseudo-amino acid composition	50 50

Table 4. List of protein feature groups, features and number of descriptors computed by ProPy

Feature groups	Features	No. of descriptors
Amino acid composition	Amino acid composition	20
	Dipeptide composition	400
	Tripeptide composition	8000
Autocorrelation	Normalized Moreau – Broto autocorrelation	240
	Moran autocorrelation Geary autocorrelation	240
	Geary autocorrelation	240
Composition, transition and distribution	Composition	21
	Transition	21
	Distribution	105
Quasi-sequence order	Sequence-order-coupling number	60
	Quasi-sequence-order descriptors	100
Pseudo-amino acid composition	Type I pseudo-amino acid composition	50
	Type II pseudo-amino acid composition	50

Table 5. Tools to compute features for drug compounds

Tool and library	Programming languages	Link
ChemCPP [89]	C++	chemcpp.sourceforge.net/
RDKit [90]	Python; wrappers for Java and C#	http://www.rdkit.org/
OpenBabel [91]	C++, Perl, Python interfaces	http://openbabel.org/
PaDEL [78]	Java	www.yapcwsoft.com/dd/padeldescriptor/
OpenEye Toolkit [92]	C++; Wrappers for Python, Java, and .NET	https://docs.eyesopen.com/toolkits/python/index.html
CDK Descriptor [79]	Java	www.rguha.net/code/java/cdkdesc.html
Dragon [80]	Stand-alone application	https://chm.kode-solutions.net/products_dragon.php
BlueDesc [93]	Java	https://omictools.com/bluedesc-tool
DayLight Tookit [94]	C, Fortran, Wrappers for Java/ C++	https://www.daylight.com/products/index.html

ChemDes [95]	Web Server	www.scbdd.com/chemdes/
Rchemcpp [96]	R	http://shiny.bioinf.jku.at/Analoging/
ChemoPy [97]	Python	http://code.google.com/p/pychem/downloads/list
ChemmineR [98]	R	https://www.bioconductor.org/packages/release/bioc/html/ChemmineR.html
Indigo [99]	C++, Java, Python, Wrapper for .NET	https://lifescience.opensource.epam.com/indigo/
jCompoundMapper [100]	Java	http://jcompoundmapper.sourceforge.net

Table 6. Tools to compute features for target proteins

Tool and library	**Programming Languages**	**Link**
Protr/ProtrWeb [101]	R, Web server	https://cran.r-project.org/web/packages/protr/vignettes/protr.html
SPiCE [102]	Web sever	http://helix.ewi.tudelft.nl/spice
PROFEAT [81]	Web server	http://bidd2.nus.edu.sg/cgi-bin/profeat2016/protein/profnew.cgi
Camb [103]	C++, Java, Python, R	https://omictools.com/camb-tool
ProtDCal [82]	Java	http://bioinf.sce.carleton.ca/
ProPy [104]	Python	http://code.google.com/p/protpy/
Pse-in-One [105]	Web Server	http:// bioinformatics.hitsz.edu.cn/Pse-in-One/
POSSUM [106]	Web server, Perl, Python	http://possum.erc.monash.edu/
ProtParam [83]	Web server	https://web.expasy.org/protparam/
ProFET [107]	Python	https://www.mybiosoftware.com/profet-protein-feature-engineering-toolkit.html
KeBABS [108]	R	http://www.bioinf.jku.at/software/kebabs/

Table 7. Tools to compute features of both drug compounds and target proteins

Tool and library	Programming language	Link
Rcpi [76]	R	http://bioconductor.org/packages/ release/bioc/html/Rcpi.html
PyDPI [77]	Python	https://sourceforge.net/projects/pydpicao/

The substructure fingerprint that was computed to featurize the drug was the PUBCHEM_CACTVS_SUBGRAPHKEYS' property available in the PubChem database. Authors in [86] came out with a classifier-based approach for recognizing chemo genomic characteristics that are involved in networks of drug-target interaction. Data sets on the drug-target interaction were collected from the DrugBank database and used for the experimental research. Pharmaceutical chemical information was taken from the PubChem dataset; target genomic information from UniProt and PFAM datasets were retrieved. The feature vectors of the drug compounds were encoded using information from the PubChem dataset. 881 dimensional binary vector was constructed, with each component representing the presence or absence of a substructure of PubChem. Equally, 876 dimensional protein characteristic vector was built, where each component represented the presence or absence of PFAM domain. The protein feature set was designed using only the protein sequences.

In recent times, researchers have adopted more complex methods to featurize drugs and targets for their works. To completely extract the interaction-related features, researchers in [87] adopted an innovative method of multi-scale protein sequence representation to extract feature vectors from sequences using binary coding schemes. Usually, an original sequence of polypeptides will contain multiple continuous segments of residues. The authors in [88] used multi-scale descriptors to quantify and concatenate each continuous local region by implementing decomposition technique to collect unique feature vectors of the protein sequence. Based on the actual situation, this approach was able to transform the protein sequences into multi-scale feature vectors which spanned several length levels. Feng et al. [109] proposed a framework known as Protein And Drug Molecule interaction prediction (PADME) to predict binding affinities between compounds and proteins. PADME was developed based on Deep Neural Networks. The authors represented the compounds by Molecular Graph Convolution (MGC) [110] and combined them with protein descriptors to train their model. Performance of the model was measured on Davis, KIBA, Metz, and ToxCast datasets. Study [110] developed a deep learning model with graph CNNs for a DTI related task. Graph convolutions were employed as feature vectors for the compounds in the training of the model. [111] also leveraged a deep learning model to predict the binding strength of protein-compound interactions. They employed a CNN model to learn the representations of the proteins and compounds respectively from raw protein sequences and SMILES of compounds.

CONCLUSION

In this study, we aimed to provide references and insight into how featurization of drug compounds and target proteins is done with available cheminformatics toolkits and libraries for DTI prediction and other related works. In the beginning, we gave a comprehensive introduction of DTI prediction by discussing drugs, drug targets, different categories and subcategories of computational methods for DTI prediction. An examination of existing bioactivity databases from which data sets of drugs and targets can be obtained is presented. In subsequent parts of the paper, we elaborated on feature extraction, and the compound and protein descriptors involved. We highlighted the Cheminformatics software packages, which are used to generate descriptors of drug compounds and proteins and also to construct their feature vectors. Also, we demonstrated the use of Cheminformatics tools for drug compound and protein featurization. A review has been presented on state-of-the-art machine learning and deep learning works in which featurization was performed on the data sets used.

Featurization of data sets for machine learning-based DTI prediction methods is very challenging, particularly to non-experts in the Cheminformatics domain. Majority of the cheminformatics software toolkits do not provide Graphical User Interface for their utilization. Users have to write codes to accomplish the task of featurization. This becomes a seemingly unsurmountable problem for non-experts without good programming skills and deep knowledge of molecular fingerprints, descriptors and, in general, Cheminformatics. In addition, the manuals of these tools do not provide a well-documented information on how they can be used for featurization, so as to serve as a guide for non-cheminformatics experts. Another challenge of featurization is that, features generated for drug compounds and target proteins often have constant and/or missing values among them. Also, some of the feature values are disproportional. This demands for extra tasks of feature selection and data normalization. Feature selection may result in the loss of vital information about compounds and targets. In our study, we observed that learning complex featurization of drug compounds and target proteins with deep learning models is now becoming pervasive in the domain of DTI prediction. We tend to look at this new development in our future work.

ACKNOWLEDGEMENTS

This work was supported by SipingSoft Co. Ltd; Chengdu, P.R. China.

REFERENCES

[1] J. A. Dimasi, L. Feldman, A. Seckler, and A. Wilson, "Trends in risks associated with new drug development: Success rates for investigational drugs," *Clin. Pharmacol. Ther.*, vol. 87, no. 3, pp. 272–277, 2010.

[2] T. T. Ashburn and K. B. Thor, "Drug repositioning: Identifying and developing new uses for existing drugs," *Nat. Rev. Drug Discov.*, vol. 3, no. 8, pp. 673–683, 2004.

[3] A. Ezzat, M. Wu, X. L. Li, and C. K. Kwoh, "Drug-target interaction prediction via class imbalance-aware ensemble learning," *BMC Bioinformatics*, vol. 17, no. Suppl 19, 2016.

[4] S. Lipperman-Kreda, J. P. Lee, C. Morrison, and B. Freisthler, "Availability of tobacco products associated with use of marijuana cigars (blunts)," *Drug Alcohol Depend.*, 2014.

[5] H. Kalant, "A critique of cannabis legalization proposals in Canada," *International Journal of Drug Policy*. 2016.

[6] Drugs, "Drug Policy and the Public Good: a summary of the book.," *Addiction*, 2010.

[7] K. Sachdev and M. K. Gupta, "A comprehensive review of feature based methods for drug target interaction prediction," *J. Biomed. Inform.*, vol. 93, no. March, p. 103159, 2019.

[8] T. Huang *et al.*, "Functional association between influenza A (H1N1) virus and human," *Biochem. Biophys. Res. Commun.*, vol. 390, no. 4, pp. 1111–1113, 2009.

[9] T. Huang, K. Tu, Y. Shyr, C. C. Wei, L. Xie, and Y. X. Li, "The prediction of interferon treatment effects based on time series microarray gene expression profiles," *J. Transl. Med.*, vol. 6, pp. 1–9, 2008.

[10] A. Masoudi-Nejad, Z. Mousavian, and J. H. Bozorgmehr, "Drug-target and disease networks: polypharmacology in the post-genomic era," *Silico Pharmacol.*, vol. 1, no. 1, pp. 2–5, 2013.

[11] X. Z. Zhen Wang, Quan Zou, Yi Jiang, Ying Ju, "Review of Protein Subcellular Localization Prediction," *Curr. Bioinform.*, vol. 9, no. 3, p. 12, 2014.

[12] X. Zeng, L. Liu, L. Lu, and Q. Zou, "Prediction of potential disease-associated microRNAs using structural perturbation method," *Bioinformatics*, vol. 34, no. 14, pp. 2425–2432, 2018.

[13] J. Zeng, D. Li, Y. Wu, Q. Zou, and X. Liu, "An Empirical Study of Features Fusion Techniques for Protein-Protein Interaction Prediction," *Curr. Bioinform.*, vol. 11, no. 1, pp. 4–12, 2015.

[14] A. S. Rifaioglu, H. Atas, M. J. Martin, R. Cetin-Atalay, V. Atalay, and T. Doğan, "Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases," *Briefings in Bioinformatics*. 2018.

[15] G. R. Sliwoski, J. Meiler, and E. W. Lowe, "Computational Methods in Drug Discovery Prediction of protein structure and ensembles from limited experimental data View project Antibody modeling, Antibody design and Antigen-Antibody interactions View project," *Comput. Methods Drug Discov.*, vol. 66, no. 1, pp. 334–95, 2014.

[16] I. Muegge and Y. C. Martin, "A general and fast scoring function for protein-ligand interactions: A simplified potential approach," *J. Med. Chem.*, vol. 42, no. 5, pp. 791–804, 1999.

[17] H. F. G. Velec, H. Gohlke, and G. Klebe, "DrugScoreCSD-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction," *J. Med. Chem.*, vol. 48, no. 20, pp. 6296–6303, 2005.

[18] T. J. A. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz, "DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases," *J. Comput. Aided. Mol. Des.*, vol. 15, no. 5, pp. 411–428, 2001.

[19] N. Schneider, G. Lange, S. Hindle, R. Klein, and M. Rarey, "A consistent description of HYdrogen bond and DEhydration energies in protein-ligand complexes: Methods behind the HYDE scoring function," *J. Comput. Aided. Mol. Des.*, vol. 27, no. 1, pp. 15–29, 2013.

[20] R. Wang, L. Liu, L. Lai, and Y. Tang, "SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex," *J. Mol. Model.*, vol. 4, no. 12, pp. 379–394, 1998.

[21] R. Wang, L. Lai, and S. Wang, "Further development and validation of empirical scoring functions for structure-based binding affinity prediction," *J. Comput. Aided. Mol. Des.*, vol. 16, no. 1, pp. 11–26, 2002.

[22] M. A. Yildirim, K. Il Goh, M. E. Cusick, A. L. Barabási, and M. Vidal, "Drug-target network," *Nat. Biotechnol.*, vol. 25, no. 10, pp. 1119–1126, 2007.

[23] R. A. Friesner *et al.*, "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy," *J. Med. Chem.*, 2004.

[24] M. McGann, "FRED pose prediction and virtual screening accuracy," *J. Chem. Inf. Model.*, 2011.

[25] G. M. Morris *et al.*, "Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function," *J. Comput. Chem.*, 1998.

[26] O. Trott and A. J. Olson, "Software news and update AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *J. Comput. Chem.*, 2010.

[27] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, and R. D. Taylor, "Improved protein-ligand docking using GOLD," *Proteins Struct. Funct. Genet.*, 2003.

[28] B. Kramer, M. Rarey, and T. Lengauer, "Evaluation of the FlexX incremental construction algorithm for protein-ligand docking," *Proteins Struct. Funct. Genet.*, 1999.

[29] Z. Mousavian and A. Masoudi-Nejad, "Drug-target interaction prediction via chemogenomic space: Learning-based methods," *Expert Opin. Drug Metab. Toxicol.*, vol. 10, no. 9, pp. 1273–1287, 2014.

[30] Z. Wu, W. Li, G. Liu, and Y. Tang, "Network-based methods for prediction of drug-target interactions," *Front. Pharmacol.*, vol. 9, no. OCT, pp. 1–14, 2018.

[31] W. Ba, P. Fulfillment, and R. For, "Novel Methods for Drug-Target Interaction Prediction using Graph Mining," 2016.

[32] H. Ding, I. Takigawa, H. Mamitsuka, and S. Zhu, "Similarity-basedmachine learning methods for predicting drug-target interactions: A brief review," *Brief. Bioinform.*, vol. 15, no. 5, pp. 734–747, 2013.

[33] R. Sawada, H. Iwata, S. Mizutani, and Y. Yamanishi, "Target-Based Drug Repositioning Using Large-Scale Chemical-Protein Interactome Data," *J. Chem. Inf. Model.*, 2015.

[34] S. Kim *et al.*, "PubChem substance and compound databases," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1202–D1213, 2016.

[35] A. Gaulton *et al.*, "The ChEMBL database in 2017," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D945–D954, Jan. 2017.

[36] D. S. Wishart *et al.*, "DrugBank 5.0: A major update to the DrugBank database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, Jan. 2018.

[37] "STITCH: chemical association networks." [Online]. Available: http://stitch.embl.de/. [Accessed: 23-Dec-2019].

[38] "KEGG: Kyoto Encyclopedia of Genes and Genomes." [Online]. Available: https://www.genome.jp/kegg/. [Accessed: 23-Dec-2019].

[39] S. K. Burley *et al.*, "Protein Data Bank: The single global archive for 3D macromolecular structure data," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D520–D528, Jan. 2019.

[40] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, "BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1045–D1053, 2016.

[41] "BioGRID | Database of Protein, Chemical, and Genetic Interactions." [Online]. Available: https://thebiogrid.org/. [Accessed: 23-Dec-2019].

[42] I. Schomburg *et al.*, "BRENDA in 2013: Integrated reactions, kinetic data, enzyme function data, improved disease classification: New options and contents in BRENDA," *Nucleic Acids Res.*, 2013.

[43] "SIDER Side Effect Resource." [Online]. Available: http://sideeffects.embl.de/. [Accessed: 23-Dec-2019].

[44] S. El-Gebali *et al.*, "The Pfam protein families database in 2019," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D427–D432, Jan. 2019.

[45] "SuperTarget." [Online]. Available: http://insilico.charite.de/supertarget/. [Accessed: 23-Dec-2019].

[46] "SuperPred webserver." [Online]. Available: http://prediction.charite.de/. [Accessed: 23-Dec-2019].

[47] "MATADOR." [Online]. Available: http://matador.embl.de/. [Accessed: 23-Dec-2019].

[48] Y. Wang *et al.*, "Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics," *Nucleic Acids Res.*, 2019.

[49] "Gene Ontology Resource." [Online]. Available: http://geneontology.org/. [Accessed: 23-Dec-2019].

[50] "Home Page." [Online]. Available: http://asdcd.amss.ac.cn/. [Accessed: 23-Dec-2019].

[51] Y. Liu, Q. Wei, G. Yu, W. Gai, Y. Li, and X. Chen, "DCDB 2.0: A major update of the drug combination database," *Database*, vol. 2014, pp. 1–6, 2014.

[52] A. Ahmed, R. D. Smith, J. J. Clark, J. B. D. Jr, and H. A. Carlson, "Recent improvements to Binding MOAD: A resource for protein-ligand Binding affinities and structures," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D465–D469, Jan. 2015.

[53] D. Weininger, "SMILES, a Chemical Language and Information System.," *J. Chem. Inf. Comput. Sci.*, 1988.

[54] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi, "InChI, the IUPAC International Chemical Identifier," *J. Cheminform.*, 2015.

[55] "Tutorial:Fingerprints - Open Babel." [Online]. Available: https://openbabel.org/wiki/Tutorial:Fingerprints. [Accessed: 23-Dec-2019].

[56] B. Ramsundar, P. Eastman, P. Walters, and V. Pande, *Deep Learning for the Life Sciences*. 2019.

[57] K. C. Chou and Y. D. Cai, "Prediction of membrane protein types by incorporating amphipathic effects," *J. Chem. Inf. Model.*, 2005.

[58] Q. Bin Gao, Z. Z. Wang, C. Yan, and Y. H. Du, "Prediction of protein subcellular location using a combined feature of sequence," *FEBS Lett.*, 2005.

[59] Z. P. Feng and C. T. Zhang, "Prediction of membrane protein types based on the hydrophobic index of amino acids," *J. Protein Chem.*, 2000.

[60] X. Y. Xia, M. Ge, Z. X. Wang, and X. M. Pan, "Accurate prediction of protein structural class," *PLoS One*, 2012.

[61] D. S. Horne, "Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities," *Biopolymers*, 1988.

[62] R. R. Sokal and B. A. Thomson, "Population structure inferred by local spatial autocorrelation: An example from an Amerindian tribal population," *Am. J. Phys. Anthropol.*, 2006.

[63] C. Z. Cai, L. Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen, "SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence," *Nucleic Acids Res.*, 2003.

[64] L. Y. Han, C. Z. Cai, S. L. Lo, M. C. M. Chung, and Y. Z. Chen, "Prediction of RNA-binding proteins from primary sequence by a support vector machine approach," *RNA*, 2004.

[65] S. L. Lo, C. Z. Cai, Y. Z. Chen, and M. C. M. Chung, "Effect of training datasets on support vector machine prediction of protein-protein interactions," in *Proteomics*, 2005.

[66] J. Cui *et al.*, "Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties," *Mol. Immunol.*, 2007.

[67] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim, "Prediction of protein folding class using global description of amino acid sequence," *Proc. Natl. Acad. Sci. U. S. A.*, 1995.

[68] H. H. Lin, L. Y. Han, C. Z. Cai, Z. L. Ji, and Y. Z. Chen, "Prediction of transporter family from protein sequence by support vector machine approach," *Proteins Struct. Funct. Genet.*, vol. 62, no. 1, pp. 218–231, 2006.

[69] J. Shen *et al.*, "Predicting protein-protein interactions based only on sequences information," *Proc. Natl. Acad. Sci. U. S. A.*, 2007.

[70] K. C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins Struct. Funct. Genet.*, 2001.

[71] C. Chen, X. Zhou, Y. Tian, X. Zou, and P. Cai, "Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network," *Anal. Biochem.*, 2006.

[72] K. C. Chou and Y. D. Cai, "Prediction of protein subcellular locations by GO-FunD-PseAA predictor," *Biochem. Biophys. Res. Commun.*, 2004.

[73] S. A. K. Ong, H. H. Lin, Y. Z. Chen, Z. R. Li, and Z. Cao, "Efficacy of different protein descriptors in predicting protein functional families," *BMC Bioinformatics*, vol. 8, pp. 1–14, 2007.

[74] A. Bateman *et al.*, "UniProt: The universal protein knowledgebase," *Nucleic Acids Res.*, 2017.

[75] J. Desaphy, G. Bret, D. Rognan, and E. Kellenberger, "Sc-PDB: A 3D-database of ligandable binding sites-10 years on," *Nucleic Acids Res.*, 2015.

[76] D. S. Cao, N. Xiao, Q. S. Xu, and A. F. Chen, "Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions," *Bioinformatics*, vol. 31, no. 2, pp. 279–281, 2015.

[77] D. S. Cao, Y. Z. Liang, J. Yan, G. S. Tan, Q. S. Xu, and S. Liu, "PyDPI: Freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies," *J. Chem. Inf. Model.*, 2013.

[78] C. W. Yap, "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints," *J. Comput.*

Chem., 2011.

[79] E. L. Willighagen *et al.*, “The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching,” *J. Cheminform.*, 2017.

[80] A. Mauri, V. Consonni, M. Pavan, and R. Todeschini, “DRAGON software: An easy approach to molecular descriptor calculations,” *Match*, 2006.

[81] H. B. Rao, F. Zhu, G. B. Yang, Z. R. Li, and Y. Z. Chen, “Update of PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence,” *Nucleic Acids Res.*, 2011.

[82] Y. B. Ruiz-Blanco, W. Paz, J. Green, and Y. Marrero-Ponce, “ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins,” *BMC Bioinformatics*, 2015.

[83] E. Gasteiger *et al.*, “Protein Identification and Analysis Tools on the ExPASy Server,” in *The Proteomics Protocols Handbook*, 2005.

[84] H. Yu *et al.*, “A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data,” *PLoS One*, 2012.

[85] J. Shen, F. Cheng, Y. Xu, W. Li, and Y. Tang, “Estimation of ADME properties with substructure pattern recognition,” *J. Chem. Inf. Model.*, 2010.

[86] Y. Tabei, E. Pauwels, V. Stoven, K. Takemoto, and Y. Yamanishi, “Identification of chemogenomic features from drug-target interaction networks using interpretable classifiers,” *Bioinformatics*, 2012.

[87] Z. H. You, K. C. C. Chan, and P. Hu, “Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest,” *PLoS One*, 2015.

[88] M. Wen *et al.*, “Deep-Learning-Based Drug-Target Interaction Prediction,” *J. Proteome Res.*, vol. 16, no. 4, pp. 1401–1409, 2017.

[89] “ChemCpp: CHEMCPP.” [Online]. Available: http://chemcpp.sourceforge.net/html/index.html. [Accessed: 23-Dec-2019].

[90] “RDKit.” [Online]. Available: https://www.rdkit.org/. [Accessed: 23-Dec-2019].

[91] “Open Babel.” [Online]. Available: http://openbabel.org/wiki/Main_Page. [Accessed: 23-Dec-2019].

[92] “Toolkit Development Platform | Cheminformatics + Modeling TKs.” [Online]. Available: https://www.eyesopen.com/toolkit-development. [Accessed: 23-Dec-2019].

[93] “BlueDesc - omicX.” [Online]. Available: https://omictools.com/bluedesc-tool. [Accessed: 23-Dec-2019].

[94] “Daylight>Products.” [Online]. Available: https://daylight.com/products/toolkit.html. [Accessed: 23-Dec-2019].

[95] J. Dong *et al.*, “ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation,” *J. Cheminform.*, vol. 7, no. 1, Dec. 2015.

[96] G. Klambauer, M. Wischenbart, M. Mahr, T. Unterthiner, A. Mayr, and S. Hochreiter, “Rchemcpp: A web service for structural analoging in ChEMBL, Drugbank and the Connectivity Map,” *Bioinformatics*, vol. 31, no. 20, pp. 3392–3394, Mar. 2015.

[97] D. S. Cao, Q. S. Xu, Q. N. Hu, and Y. Z. Liang, “ChemoPy: Freely available python package for computational biology and chemoinformatics,” *Bioinformatics*, vol. 29, no. 8, pp. 1092–1094, Apr. 2013.

[98] Y. Cao, A. Charisi, L. C. Cheng, T. Jiang, and T. Girke, “ChemmineR: A compound mining framework for R,” *Bioinformatics*, vol. 24, no. 15, pp. 1733–1734, Aug. 2008.

[99] “Indigo Toolkit.” [Online]. Available: https://lifescience.opensource.epam.com/indigo/. [Accessed: 23-Dec-2019].

[100] G. Hinselmann, L. Rosenbaum, A. Jahn, N. Fechner, and A. Zell, “jCompoundMapper: An open source Java library and command-line tool for chemical fingerprints,” *J. Cheminform.*, vol. 3, no. 1, p. 3, Dec. 2011.

[101] N. Xiao, D. S. Cao, M. F. Zhu, and Q. S. Xu, “Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences,” in *Bioinformatics*, 2015, vol. 31, no. 11, pp. 1857–1859.

[102] “GitHub - basvandenberg/spice: SPiCe - Sequence-based Protein Classification.” [Online]. Available: https://github.com/basvandenberg/spice. [Accessed: 23-Dec-2019].

[103] D. S. Murrell *et al.*, “Chemically Aware Model Builder (camb): an R package for property and bioactivity modelling of small molecules,” *J. Cheminform.*, vol. 7, no. 1, p. 45, Dec. 2015.

[104] D. Cao, Q. Xu, and Y. Liang, “Systems biology propy : a tool to generate various modes of Chou ’ s PseAAC,” vol. 29, no. 7, pp. 960–962, 2013.

[105] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K. C. Chou, “Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences,” *Nucleic Acids Res.*, vol. 43, no. W1, pp. W65–W71, 2015.

[106] “POSSUM | Home Page.” [Online]. Available: http://possum.erc.monash.edu/. [Accessed: 23-Dec-2019].

[107] “ProFET – Protein Feature Engineering Toolkit | My Biosoftware - Bioinformatics Softwares Blog.” [Online]. Available: https://www.mybiosoftware.com/profet-protein-feature-engineering-toolkit.html. [Accessed: 23-Dec-2019].

[108] J. Palme, S. Hochreiter, and U. Bodenhofer, “KeBABS: An R package for kernel-based analysis of biological sequences,” *Bioinformatics*, vol. 31, no. 15, pp. 2574–2576, 2015.

[109] Q. Feng, E. Dueva, A. Cherkasov, and M. Ester, “PADME: A Deep Learning-based Framework for Drug-Target Interaction Prediction,” pp. 1–21.

[110] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, “Molecular graph convolutions: moving beyond fingerprints,” *J. Comput. Aided. Mol. Des.*, 2016.

[111] H. Öztürk, A. Özgür, and E. Ozkirimli, “DeepDTA: Deep drug-target binding affinity prediction,” *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.