LINGO Profiles Fingerprint and Association Rule Mining for drug-target interaction prediction

Muhammad Jaziem Mohamed Javeed, Azwaar Khan Azlim Khan, Nurul Hashimah Ahamed Hassain Malim*

School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang 11800, Malaysia

*** Corresponding author:** Nurul Hashimah Ahamed Hassain Malim, nurulhashimah@usm.my

ARTICLE INFO

Received: 27 June 2023
Accepted: 18 July 2023
Available online: 22 August 2023

http:10.59400/cai.v1i1.99

ABSTRACT: The prediction of drug-target interactions (DTIs) using machine learning techniques together with the proper representation of compounds can speed up the time-consuming experimental work in predicting DTIs especially when a large dataset is used. Hence, in this paper, we have proposed a new molecular descriptor based on LINGO Profiles known as LINGO Profiles Fingerprint (LPFP). LPFP is used together with machine learning to predict DTIs on a ChEMBL dataset. Dimensionality reduction using Association Rule Mining (ARM) is also introduced to overcome the high dimensionality suffered by LPFP. LPFP managed to reach an equal accuracy reading to the state-of-the-art descriptor called ECFP4 (Δ0.18%), but it suffers in the time taken (Δ27 mins) due to the dimensionality problem mentioned. Hence, three new smaller size LPFPs (s = 60%, s = 70%, s = 80%) were constructed by only extracting the important fragments using ARM and then a benchmark analysis with the original LPFP and ECFP4 fingerprints was done. This study not only solved the dimensionality problem, but also managed to excel in both the accuracy and time taken when predicting DTIs. An increase in the accuracy of over 250 times faster than the original LPFP was observed after the benchmark analysis is performed. Furthermore, an accuracy of over 80% was achieved in three new activity classes that are acquired from ChEMBL, further proving the promising performance of ARM which has made it favourable for LPFPs to be used in DTI prediction and in other drug discovery problems.

KEYWORDS: LINGO Profiles Fingerprint (LPFP); Association Rule Mining (ARM); machine learning; dimensionality reduction; ECFP4; drug-target interactions

1. Introduction

During the past centuries, effort to discover new types of drugs was initiated because there were diseases or clinical conditions which required curable drugs to tackle the immediate menace as there was a shortage in the market for these medical products[1,2]. Due to the conventional methods of drug discovery as explained in the research of Hann and Green[3] are costly and time-consuming, the use of in silico methods to predict the interactions between drugs and target proteins provide a crucial leap for drug repositioning, as it can remarkably reduce wet-laboratory work and lower the cost of the experimental discovery of new drug-target interactions (DTIs)[4,5]. Tools such as machine learning created from the cheminformatics domain provide a huge potential to go further in the drug design and discovery realm, as they serve the integration of information in several levels to enhance the reliability of data outcomes[4–7]. Generally, machine learning is used to develop classification models and these were used to distinguish active and inactive pairs of drug-like small molecule ligands and their respective target proteins with a higher accuracy[8]. The most common machine learning methods used in drug discovery are typically binary classifiers such as the Naïve Bayes classifiers used by Glick et al.[9] to predict DTIs, Artificial Neural Networks (ANNs) as depicted by Wen et al.[10] and Prado-Prado et al.[11], Support Vector Machines (SVM) by Nasution et al.[12] and Rodríguez-Pérez et al.[13], and Random Forests (RF) as observed by Riddick et al.[14] and Shi et al.[15].

In the works mentioned above[9–15], the use of the molecular descriptor, ECFP4 or Extended Connectivity Fingerprints of bond diameter 4 is very common. When the machine learning methods are evaluated, the superiority of ECFP4 can be seen clearly. In the SVM and Naïve Bayes classifiers, accuracy greater than 80% were recorded[9,12,13] while in ANN, the models developed with the use of ECFP4 managed to outperform other classification techniques such as Decision Trees and Random Forests[10]. The nature of ECFP4, which contains features that are significant to any type of compounds, allow the prediction of a drug to a protein target being made correctly. The convincing performance of ECFP4 makes other researchers in this field to use it as their main compound representation especially for the purpose of DTI prediction. However, in most chemical databases the compounds are usually represented in the SMILES format, or Simplified Molecular Input Line Entry System which is a string that represents a compound structure in the form of text[16]. SMILES is the most extensively used representation in drug discovery as it is easier to use and comprehend compared to other representations. The SMILES notation can be further converted into molecular descriptors such as ECFP4 for the purpose of drug-target interaction prediction, as briefly discussed above.

LINGO is also another representation of chemical compounds that was introduced in the study of Vidal et al.[16]. LINGO is a simple text-based method that calculates the molecular similarities and predicts the structure-related properties using the SMILES representation of a compound[17]. The LINGO method breaks down the canonical SMILES string into a set of substrings of defined length of four characters in which they are called LINGO Profiles or LINGOs[17]. Recently, LINGO Profiles have come into view of the drug discovery community as it provides the simplicity when retrieving the molecules from a database[16]. Hence, in this paper, we will be proposing a new molecular representation based on LINGO Profiles called LINGO Profiles Fingerprint (LPFP). LPFP is a binary vector of 0 s and 1 s whereby each position in the vector indicates the absence (0) or presence (1) of features predetermined in the design of the fingerprint of a specific compound.

Nonetheless, not much work has utilized LINGO Profiles in the prediction of DTIs, mainly due to the superiority of ECFP4. However, based on our previous work[18], we found out that LINGO Profiles gave a comparable performance to ECFP4 in the virtual screening experiment. Hence, we foresee the potential of LINGO Profiles to give a good accuracy in DTI prediction as well. The only drawback of the LINGO Profiles is that the number of LINGO Profiles generated depends on the length of the SMILES representation of a compound. Usually the longer the SMILES string, more LINGO Profiles are generated. Thus, the amount gets bigger when a greater number of compounds is used which will then burden the machine learning model even more as it had to process a huge number of fragments to predict DTIs. Therefore, it is often beneficial to reduce the dimension of the data that is being fed into the learning algorithms such as adopting the Association Rule Mining (ARM) technique not only for reasons of computational efficiency but also to improve the accuracy of the analysis when high dimensional data is being used, as depicted in the studies of Siswanto et al.[19], Li et al.[20] and Malavika et al.[22]. Furthermore, dimensionality reduction can also help to deduce important facts and to discover new findings that might be useful for the drug discovery community in future[22–25].

Hence, in this paper we will benchmark our proposed molecular descriptor called LINGO Profiles Fingerprint (LPFP) with ECFP4 when performing DTI prediction and evaluate the performance of LPFP in terms of the accuracy and time taken as to how efficient LPFP works with the use of the Association Rule Mining (ARM) technique to reduce the dimension of LPFP to only include significant features for efficient prediction of drug-target interactions.

2. Proposed methodology

The proposed methodology is depicted in **Figure 1**.

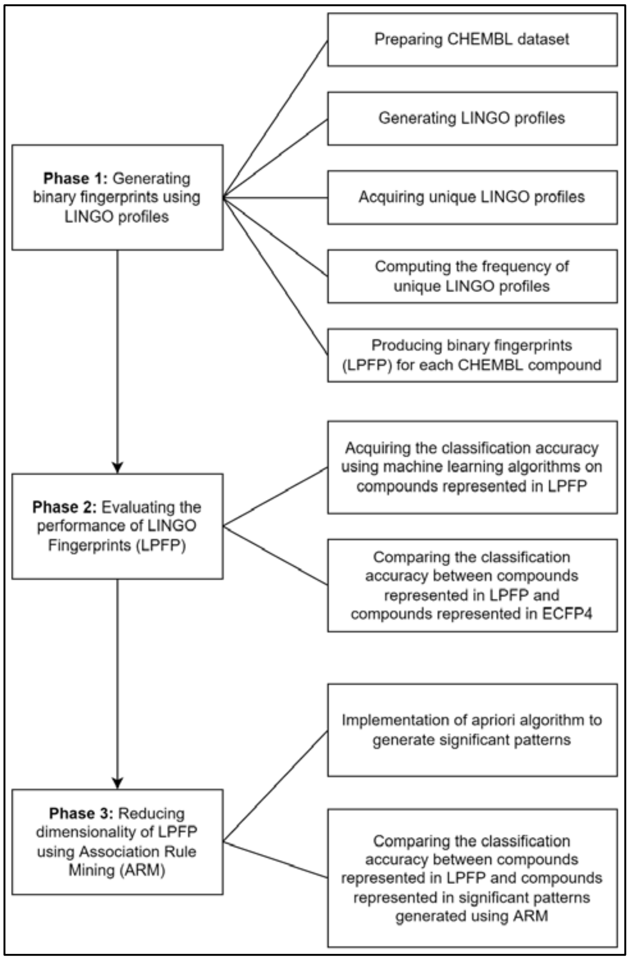

Phase 1: Generating binary fingerprints using LINGO profiles
Preparing CHEMBL dataset
Generating LINGO profiles
Acquiring unique LINGO profiles
Computing the frequency of unique LINGO profiles
Producing binary fingerprints (LPFP) for each CHEMBL compound
Phase 2: Evaluating the performance of LINGO Fingerprints (LPFP)
Acquiring the classification accuracy using machine learning algorithms on compounds represented in LPFP
Comparing the classification accuracy between compounds represented in LPFP and compounds represented in ECFP4
Phase 3: Reducing dimensionality of LPFP using Association Rule Mining (ARM)
Implementation of apriori algorithm to generate significant patterns
Comparing the classification accuracy between compounds represented in LPFP and compounds represented in significant patterns generated using ARM

Figure 1. The overall framework of the three main phases involved in all the experiments conducted in this study in generating LPFPs and using ARM to reduce the dimensionality of the fingerprints.

2.1. Phase 1: Generating Binary Fingerprints (LPFP) using LINGO Profiles

The general idea of creating binary fingerprints using LINGO Profiles is to encode the structural information of a chemical compound into bit strings and then splitting the molecule into multiple fragments. If a fragment is present in the chemical compound, the corresponding bit will be assigned to '1' and if a fragment is absent, the bit will be assigned to '0' as depicted in **Figure 2**. The occurrence of a fragment determines the bit, not the number of times the same fragment occurs in the structure. For instance, a fragment will be set only to bit '1' no matter if it is present once or ten times. Therefore, the number of different types of fragments decide the number of bits set in the binary fingerprints. The fragment here is referring to one LINGO Profile. From the LINGO Profiles generated, the duplicated ones are first removed. Then, we sorted the fragments based on the number of occurrences in a decreasing order. This is important as the fragment with a higher frequency sits on the lower bit position while the fragment with a lower frequency sits on the higher bit position of the respective fingerprint. The process of generating LPFPs are discussed in detail in the subsections that follow.

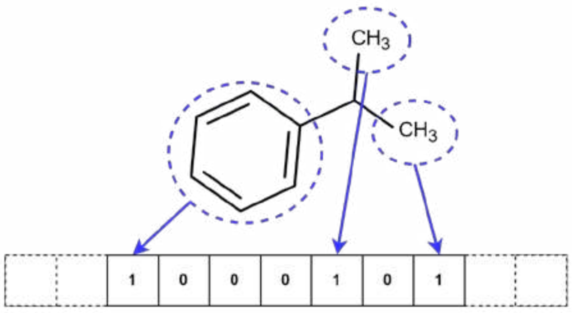


Figure 2. Encoding a chemical compound structure into bit strings of 0 s and 1 s.

2.1.1. Preparing ChEMBL dataset

The first step in Generating Binary Fingerprints called LINGO Profiles Fingerprints (LPFPs) is to prepare the dataset. In this study, we have decided to use the ChEMBL database[26]. The database contains a huge number of drug-like compounds, including over 2.3 million compounds and over 15,000 target proteins and activity classes as well as information on the compounds that were tested and its structures and abstracted bioactivities. The special characteristics possessed by the activity classes in ChEMBL are due to the ligands within it that interacts with certain target proteins.

There are two main categories to which a particular activity class can be classified into and they are known as homogenous and heterogenous. Classes whose compounds share a large number of common fragments are called homogenous. Whereas on the contrary, classes with fewer common fragments that are shared between their compounds are described as heterogeneous. The structure of the compounds within a particular activity class decides which category an activity class belongs to. Since it is a difficult task to observe and analyse each and every compound structure, we have calculated the Mean Pairwise Similarity (MPS) value for each activity class, which can be computed using Equation (1) that is also found and used in the study of Arif et al.[27]:

$$MPS = \frac{Similarity\ Value}{Number\ of\ Actives\ in\ the\ Activity\ Class} \tag{1}$$

The MPS value can be defined as the similarity of chemical structures within each activity class. The MPS reading for an activity class can be obtained by performing pairwise similarity between one random reference structure and every other compound in the activity class using the Tanimoto coefficient. A higher MPS value indicates homogenous activity classes while a lower MPS value indicates heterogeneous activity classes. Six activity classes with different number of compounds were selected and their respective MPS values were computed and they are shown in the table below.

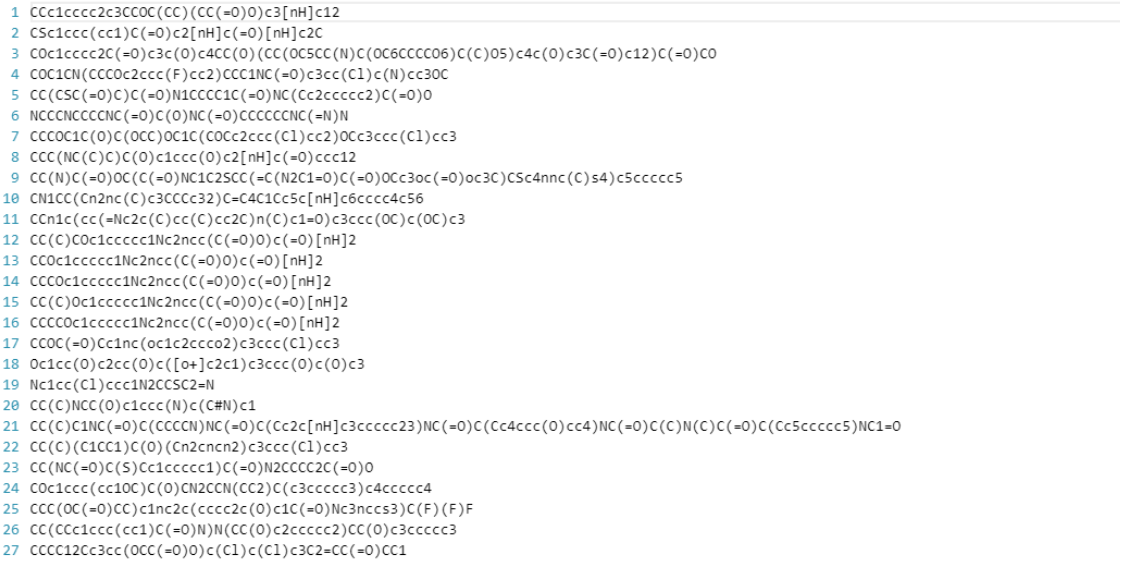
Table 1 shows the homogeneous and heterogeneous activity classes with their respective MPS values that were calculated. It is also important to note that intermediate activity classes refer to activity classes in which their MPS values ranges between the homogenous and heterogenous activity classes. After categorizing the activity classes into homogeneous, intermediate and heterogeneous classes, it is time for us to generate the LINGO Profiles for each compound in all the activity classes.

Table 1. Homogenous and heterogeneous activity classes together with their respective number of compounds and MPS values.

ID	Type	Activity classes	Number of compounds	MPS value
12659	Homo	Prostanoid DP receptor	204	0.21355893
117		Somatostatin receptor 2	62	0.21584493
20174	Inter	G protein-coupled receptor 44	1244	0.17281741
130		Dopamine D3 receptor	1540	0.17229444
52	Hetero	Alpha-2a adrenergic receptor	420	0.15204343
11365		Cytochrome P450 2D6	534	0.15767914

2.1.2. Generating LINGO Profiles

The SMILES representation of each compound for each activity class, as depicted in **Figure 3**, is important for us in order to generate LINGO Profiles.

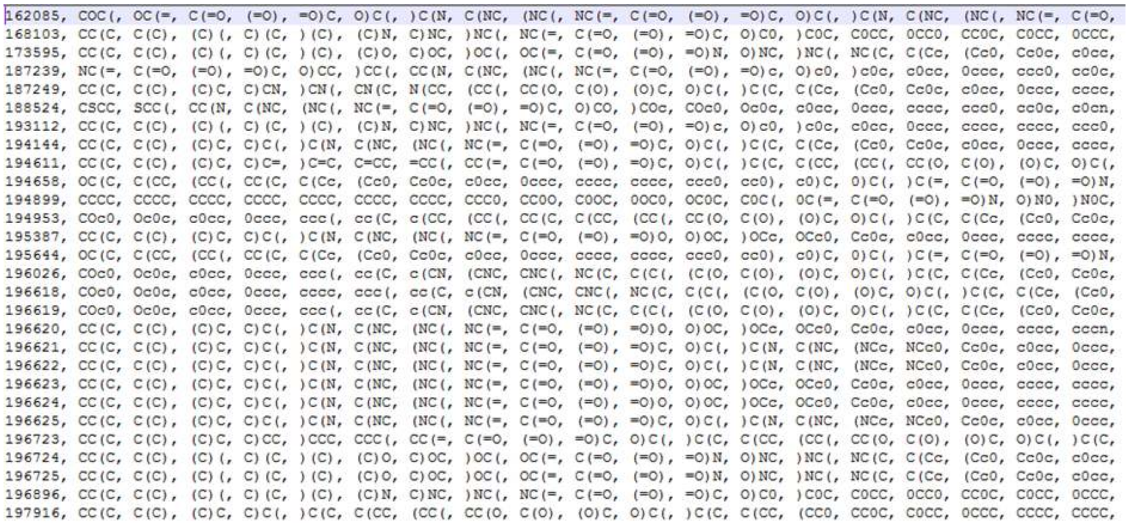


```
CCc1cccc2c3CCOC(CC)(CC(=O)O)c3[nH]c12
CSc1ccc(cc1)C(=O)c2[nH]c(=O)[nH]c2C
COc1cccc2C(=O)c3c(O)c4CC(O)(CC(OC5CC(N)C(OC6CCCO6)C(C)O5)c4c(O)c3C(=O)c12)C(=O)CO
COC1CN(CCCOc2ccc(F)cc2)CCC1NC(=O)c3cc(Cl)c(N)cc3OC
CC(CSC(=O)C)C(=O)N1CCCC1C(=O)NC(Cc2ccccc2)C(=O)O
NCCCNCCCCNC(=O)C(O)NC(=O)CCCCCCNC(=N)N
CCCOC1C(O)C(OCC)OC1C(COCc2ccc(Cl)cc2)OCc3ccc(Cl)cc3
CCC(NC(C)C)C(O)c1ccc(O)c2[nH]c(=O)ccc12
CC(N)C(=O)OC(C(=O)NC1C2SCC(=C(N2C1=O)C(=O)OCc3oc(=O)oc3C)CSc4nnc(C)s4)c5ccccc5
CN1CC(Cn2nc(C)c3CCCc32)C=C4C1Cc5c[nH]c6cccc4c56
CCn1c(cc(=Nc2c(C)cc(C)cc2C)n(C)c1=O)c3ccc(OC)c(OC)c3
CC(C)COc1ccccc1Nc2ncc(C(=O)O)c(=O)[nH]2
CCOc1ccccc1Nc2ncc(C(=O)O)c(=O)[nH]2
CCCOc1ccccc1Nc2ncc(C(=O)O)c(=O)[nH]2
CC(C)Oc1ccccc1Nc2ncc(C(=O)O)c(=O)[nH]2
CCCCOc1ccccc1Nc2ncc(C(=O)O)c(=O)[nH]2
CCOC(=O)Cc1nc(oc1c2ccco2)c3ccc(Cl)cc3
Oc1cc(O)c2cc(O)c([o+]c2c1)c3ccc(O)c(O)c3
Nc1cc(Cl)ccc1N2CCSC2=N
CC(C)NCC(O)c1ccc(N)c(C#N)c1
CC(C)C1NC(=O)C(CCCCN)NC(=O)C(Cc2c[nH]c3ccccc23)NC(=O)C(Cc4ccc(O)cc4)NC(=O)C(C)N(C)C(=O)C(Cc5ccccc5)NC1=O
CC(C)(C1CC1)C(O)(Cn2cncn2)c3ccc(Cl)cc3
CC(NC(=O)C(S)Cc1ccccc1)C(=O)N2CCCC2C(=O)O
COc1ccc(cc1OC)C(O)CN2CCN(CC2)C(c3ccccc3)c4ccccc4
CCC(OC(=O)CC)c1nc2c(cccc2c(O)c1C(=O)Nc3nccs3)C(F)(F)F
CC(CCc1ccc(cc1)C(=O)N)N(CC(O)c2ccccc2)CC(O)c3ccccc3
CCCC12Cc3cc(OCC(=O)O)c(Cl)c(Cl)c3C2=CC(=O)CC1
```

Figure 3. An example of several compounds in our ChEMBL dataset with their respective SMILES notation.

With SMILES, we are able to fragment it into overlapping substrings of a fixed size known as LINGOs. In other words, LINGO Profiles are the hologram of the SMILES representation of a molecule in which they are generated by fragmenting the SMILES notation into a fixed size. Before fragmenting the SMILES notation, all element names composed of two characters in the original SMILES are substituted by the names of one character[17]. For example, the 'Cl' and 'Br' present in the SMILES notations are replaced with 'L' and 'R' respectively and the ring numbers are set to '0'. For instance, 'c1ccccc1' and 'c2ccccc2' will be represented as 'c0ccccc0'. The reason for this normalization of the SMILES string is to reduce the number of possible LINGOs to be generated and it prevents, in general, a unique reconstruction of the original SMILES representation from its holographic representation (LINGOs).

After the normalization process is completed, the simplified SMILES strings will then undergo the fragmentation phase. A total number of $n-(q-1)$ substrings (fragments) of length q is extracted from a SMILES string of length n. The length of q that we will be using is 4, as it is optimal and as every ASCII character can be encoded into 8 bits, a LINGO of size 4 can be encoded into 32 bits of information, which is precisely the size of a word on common hardware. A q-LINGO is a q-character string, including letters, numbers and symbols such as '(', ')', '[', ']' and '#' and it is obtained by stepwise fragmentation of a canonical SMILES representation of all the compounds. This process was done using Java programming and at the end of the fragmentation process, the compounds in each of the activity class will have LINGO Profiles assigned to them in the output file as depicted in **Figure 4**.



```
162085, COC(, OC(=, C(=O, (=O), =O)C, O)C(, )C(N, C(NC, (NC(, NC(=, C(=O, (=O), =O)C, O)C(, )C(N, C(NC, (NC(, NC(=, C(=O,
168103, CC(C, C(C), (C)(, C)(C, )(C), (C)N, C)NC, )NC(, NC(=, C(=O, (=O), =O)C, O)C0, )C0C, C0CC, 0CC0, CC0C, C0CC, 0CCC,
173595, CC(C, C(C), (C)(, C)(C, )(C), (C)O, C)OC, )OC(, OC(=, C(=O, (=O), =O)N, O)NC, )NC(, NC(C, C(Cc, (Cc0, Cc0c, c0cc,
187239, NC(=, C(=O, (=O), =O)C, O)CC, )CC(, CC(N, C(NC, (NC(, NC(=, C(=O, (=O), =O)c, O)c0, )c0c, c0cc, 0ccc, ccc0, cc0c,
187249, CC(C, C(C), (C)C, C)CN, )CN(, CN(C, N(CC, (CC(, CC(O, C(O), (O)C, O)C(, )C(C, C(Cc, (Cc0, Cc0c, c0cc, 0ccc, cccc,
188524, CSCC, SCC(, CC(N, C(NC, (NC(, NC(=, C(=O, (=O), =O)C, O)CO, )COc, COc0, Oc0c, c0cc, 0ccc, cccc, ccc0, cc0c, c0cn,
193112, CC(C, C(C), (C)(, C)(C, )(C), (C)N, C)NC, )NC(, NC(=, C(=O, (=O), =O)c, O)c0, )c0c, c0cc, 0ccc, cccc, cccc, ccc0,
194144, CC(C, C(C), (C)C, C)C(, )C(N, C(NC, (NC(, NC(=, C(=O, (=O), =O)C, O)C(, )C(C, C(Cc, (Cc0, Cc0c, c0cc, 0ccc, cccc,
194611, CC(C, C(C), (C)C, C)C=, )C=C, C=CC, =CC(, CC(=, C(=O, (=O), =O)C, O)C(, )C(C, C(CC, (CC(, CC(O, C(O), (O)C, O)C(,
194658, OC(C, C(CC, (CC(, CC(C, C(Cc, (Cc0, Cc0c, c0cc, 0ccc, cccc, cccc, ccc0, cc0), c0)C, 0)C(, )C(=, C(=O, (=O), =O)N,
194899, CCCC, CCCC, CCCC, CCCC, CCCC, CCCC, CCCC, CCC0, CC00, C00C, 00C0, 0C0C, C0C(, 0C(=, C(=O, (=O), =O)N, O)N0, )N0C,
194953, COc0, Oc0c, c0cc, 0ccc, ccc(, cc(C, c(CC, (CC(, CC(C, C(CC, (CC(, CC(O, C(O), (O)C, O)C(, )C(C, C(Cc, (Cc0, Cc0c,
195387, CC(C, C(C), (C)C, C)C(, )C(N, C(NC, (NC(, NC(=, C(=O, (=O), =O)O, O)OC, )OCc, OCc0, Cc0c, c0cc, 0ccc, cccc, cccc,
195644, OC(C, C(CC, (CC(, CC(C, C(Cc, (Cc0, Cc0c, c0cc, 0ccc, cccc, cccc, ccc0, cc0), c0)C, 0)C(, )C(=, C(=O, (=O), =O)N,
196026, COc0, Oc0c, c0cc, 0ccc, ccc(, cc(C, c(CN, (CNC, CNC(, NC(C, C(C(, (C(O, C(O), (O)C, O)C(, )C(C, C(Cc, (Cc0, Cc0c,
196618, COc0, Oc0c, c0cc, 0ccc, cccc, ccc(, cc(C, c(CN, (CNC, CNC(, NC(C, C(C(, (C(O, C(O), (O)C, O)C(, )C(C, C(Cc, (Cc0,
196619, COc0, Oc0c, c0cc, 0ccc, ccc(, cc(C, c(CN, (CNC, CNC(, NC(C, C(C(, (C(O, C(O), (O)C, O)C(, )C(C, C(Cc, (Cc0, Cc0c,
196620, CC(C, C(C), (C)C, C)C(, )C(N, C(NC, (NC(, NC(=, C(=O, (=O), =O)O, O)OC, )OCc, OCc0, Cc0c, c0cc, 0ccc, cccc, cccn,
196621, CC(C, C(C), (C)C, C)C(, )C(N, C(NC, (NC(, NC(=, C(=O, (=O), =O)C, O)C(, )C(N, C(NC, (NCc, NCc0, Cc0c, c0cc, 0ccc,
196622, CC(C, C(C), (C)C, C)C(, )C(N, C(NC, (NC(, NC(=, C(=O, (=O), =O)C, O)C(, )C(N, C(NC, (NCc, NCc0, Cc0c, c0cc, 0ccc,
196623, CC(C, C(C), (C)C, C)C(, )C(N, C(NC, (NC(, NC(=, C(=O, (=O), =O)O, O)OC, )OCc, OCc0, Cc0c, c0cc, 0ccc, cccc, cccc,
196624, CC(C, C(C), (C)C, C)C(, )C(N, C(NC, (NC(, NC(=, C(=O, (=O), =O)O, O)OC, )OCc, OCc0, Cc0c, c0cc, 0ccc, cccc, cccc,
196625, CC(C, C(C), (C)C, C)C(, )C(N, C(NC, (NC(, NC(=, C(=O, (=O), =O)C, O)C(, )C(N, C(NC, (NCc, NCc0, Cc0c, c0cc, 0ccc,
196723, CC(C, C(C), (C)C, C)CC, )CCC, CCC(, CC(=, C(=O, (=O), =O)C, O)C(, )C(C, C(CC, (CC(, CC(O, C(O), (O)C, O)C(, )C(C,
196724, CC(C, C(C), (C)(, C)(C, )(C), (C)O, C)OC, )OC(, OC(=, C(=O, (=O), =O)N, O)NC, )NC(, NC(C, C(Cc, (Cc0, Cc0c, c0cc,
196725, CC(C, C(C), (C)(, C)(C, )(C), (C)O, C)OC, )OC(, OC(=, C(=O, (=O), =O)N, O)NC, )NC(, NC(C, C(Cc, (Cc0, Cc0c, c0cc,
196896, CC(C, C(C), (C)(, C)(C, )(C), (C)N, C)NC, )NC(, NC(=, C(=O, (=O), =O)C, O)C0, )C0C, C0CC, 0CC0, CC0C, C0CC, 0CCC,
197916, CC(C, C(C), (C)C, C)C(, )C(C, C(CC, (CC(, CC(O, C(O), (O)C, O)C(, )C(C, C(CC, (CC0, CC0C, C0CC, 0CCC, CCCC, CCCC,
```

Figure 4. An example of several compounds in our ChEMBL dataset together with their respective LINGO Profiles and ID.

2.1.3. Acquiring unique LINGO Profiles

After generating LINGO Profiles for all the compounds in all activity classes, unique LINGO Profiles must be acquired. A Java program was developed for the purpose of extracting unique LINGO Profiles per compound. The unique LINGO Profiles in this context means that there will be no duplicates that will be generated at a different output file. These unique LINGO Profiles that will be produced are in distinct format as we have already removed an identical version of a specific LINGO Profile earlier. This procedure is important as the information obtained here will be used as an input when calculating the frequency of occurrences for all the unique LINGO Profiles generated in order to create binary fingerprints of LINGO Profiles called LPFPs.

2.1.4. Computing the frequency of unique LINGO Profiles

The next step is to compute the frequency of the occurrences of each unique LINGO Profile because this will act as a template for us to generate the fingerprints. Besides, the information obtained here will allow us to determine specifically which LINGO Profiles are the most significant in the ChEMBL dataset. In order for us to compute the frequency of unique LINGO Profiles, two output files are compared whereby the list of LINGO Profiles obtained for all compounds in our ChEMBL dataset are considered as one file (refer to Section 2.1.2) and the other file contains only the unique LINGO Profiles for all compounds in our ChEMBL dataset (refer to Section 2.1.3). The final output file produced is a sorted list of unique LINGO Profiles based on their computed frequency. The results of the frequency of occurrences of all the unique LINGO Profiles are discussed in the results section later.

2.1.5. Producing binary fingerprint for each compound in ChEMBL dataset

The final step is to finally produce the binary fingerprints. The binary fingerprints will be of length 10,000 and they consist of 0 s and 1 s, whereby the ‘0’ bit represents the absence of the unique LINGO Profile while the ‘1’ bit represents the unique LINGO that is present in the compound. The sample output of the LPFPs generated are as shown in **Figure 5**.

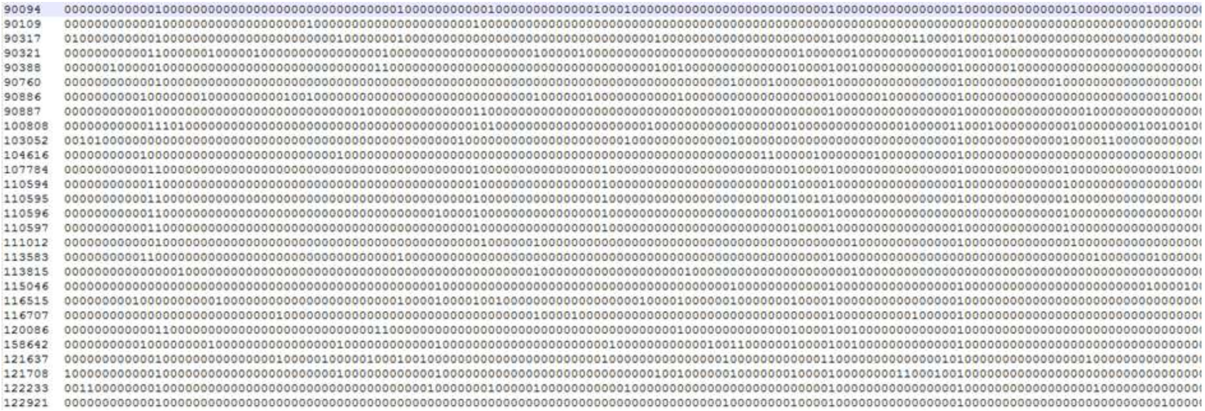


Figure 5. Part of the ChEMBL dataset which display LINGO Profiles Fingerprints (LPFP) and its respective index number for several compounds.

2.2. Phase 2: Evaluating the performance of LINGO Profiles Fingerprints (LPFPs)

2.2.1. Selecting an appropriate machine learning model

In this phase, four machine learning models are developed based on Artificial Neural Networks (ANN), Support Vector Machine (SVM), Naïve Bayes (NB) and Random Forest (RF).

Table 2 shows the size of the machine learning models created when the two molecular descriptors, ECFP4 and LPFP are used. From **Table 2**, it is observed that when ECFP4 is used, we have managed to build all the four predictive models. SVM recorded the lowest model size of about 15 to 30 MB, while ANN registered the highest model size of over 1000 MB. Furthermore, the NB and RF classifiers both have a similar range of model size of about 100 to 200 MB.

Table 2. The comparison of the all the machine learning models' size created using the molecular descriptors ECFP4 and LPFP.

Machine learning models	Size of the model when using ECFP4 (MB)	Size of the model when using LPFP (MB)
ANN	>1000	N/A
SVM	15–30	700–800
NB	100–200	N/A
RF	100–200	N/A

On the other hand, it is interesting to note that only the SVM model was able to be developed using the compounds represented in LPFP. The huge number of features in LPFP, which is ten times more than the features of ECFP4, hindered us to develop the ANN, NB and RF models despite having a large storage space (1 TB) and a competent amount of memory (8 GB) in our machine. Upgrading our machine to a better specification might help us to overcome the problem. However, if the size of the ChEMBL dataset grows bigger, the number of features in the LPFP-represented compounds will also increase.

In other words, the increase in the number of features in LPFP will result in the same problem whereby the machine learning models could not be built due to its huge size in MB. Thus, to perform the DTI prediction using our proposed molecular descriptor, LPFP, we should have a model which is available for this purpose. The unavailability of the ANN, NB and RF models in LPFP and the convincing performance of the SVM classifier in the studies of Nasution et al.[12], Rodríguez-Pérez et al.[13] and that of Heikamp and Bajorath[28] in terms of DTI prediction and drug discovery has convinced us to use the latter for later experiments. Additionally, from **Table 2**, it is also observed that the SVM classifier is the only classifier in which it could be developed for both ECFP4 and LPFP. Hence, SVM is our model of choice

2.2.2. Measuring the performance of ECFP4 and LPFP

To measure the performance of ECFP4 and LPFP in terms of drug-target interaction prediction, the ChEMBL dataset is split into training and testing sets respectively, whereby 70% off the dataset is put aside for training while the remaining 30% is for testing. Furthermore, *k*-fold cross-validation was also applied to the SVM model when performing DTI prediction with the value of *k* is 10 to avoid overfitting. The classification accuracy and the time taken for the compounds represented in LPFP will be recorded per activity class and this process is repeated six times to accommodate all the six ChEMBL activity classes selected as observed in **Table 1**.

Since we have already converted the original SMILES notation of all compounds in our ChEMBL dataset to LINGO Profiles Fingerprints as proposed, it is also vital for us to convert the original SMILES notations of all the compounds in our dataset into ECFP4 fingerprints to be compared with LPFP. Hence, the Chemistry Development Kit (CDK)[29] was used to convert SMILES to ECFP4 and the classification accuracy and the time taken for the compounds represented in ECFP4 will also be recorded separately. Then, at the end of the experiment, the results of the performance and time taken for the compounds represented in LPFP will be benchmarked with the performance and time taken for the compounds represented in ECFP4 respectively.

2.3. Phase 3: Reducing dimensionality of LPFP using Association Rule Mining (ARM)

2.3.1. Implementation of the Apriori algorithm to generate significant patterns

In data mining, association rules are rules that relate one thing to another thing in which two things are closely interrelated[19]. In the context of our work, drug-target interaction prediction is an association-based task in which we are to find drug-target pairs from a list of compounds. Since we have already represented the compounds of all six activity classes in our ChEMBL database as LINGO Profiles in Phase 1, we will now generate significant rules using the Association Rule Mining (ARM) technique. To do so, we have set the *minsupp* and *minconf* parameter to 0.6 and 0.8 respectively before implementing the algorithm on each of the activity class. It is also important to remember that each rule generated is not represented with their ID, since we have already removed the ID from each LINGO Profiles in Phase 1, before performing the experiment. Each LINGO Profile is then inserted into their respective vector space without their ID. Furthermore, rules that pass the threshold level are chosen at the end of the experiment because we believe that it may contain important knowledge in the prediction of DTIs. The rules obtained for each activity class are used to represent all the compounds available in the respective activity class. Additionally, the rules that meet the minimum support and confidence level are then divided into three main categories.

The first category gathers all the rules that have a support level of over 60% ($s = 60\%$). Then, it is followed by the second category whereby it gathers all the rules that have a support level of over 70% ($s = 70\%$) and finally rules that have a support level of 80% ($s = 80\%$) are accumulated together. From each category, fragments are filtered accordingly to eliminate duplicates, and then the unique fragments are used to represent all the compounds in the ChEMBL dataset by encoding them in a binary format. Once the unique fragments are encoded in a binary format, the data is separated into training and testing sets of proportion 70:30, similar to the previous phase and the data is fed into a machine learning algorithm (SVM) in order to predict DTIs. The performance of the model is evaluated in terms of its accuracy and it is then compared with the accuracy of the compounds represented in LPFP which was recorded in Phase 2. In other words, the comparison is done between the compounds represented in significant patterns generated using ARM and the compounds represented in LPFPs.

3. Results and discussion

3.1. Frequency of occurrences of unique LINGO Profiles

The results in this section pertains to the output file obtained from the sorted list of unique LINGO Profiles based on their computed frequency (refer to Section 2.1.4) and they are divided into two categories; the Higher Category and Lower Category. In **Figure 6**, we discovered high frequency (Higher Category) fragments that are frequently present in almost all of the compounds in the ChEMBL dataset.

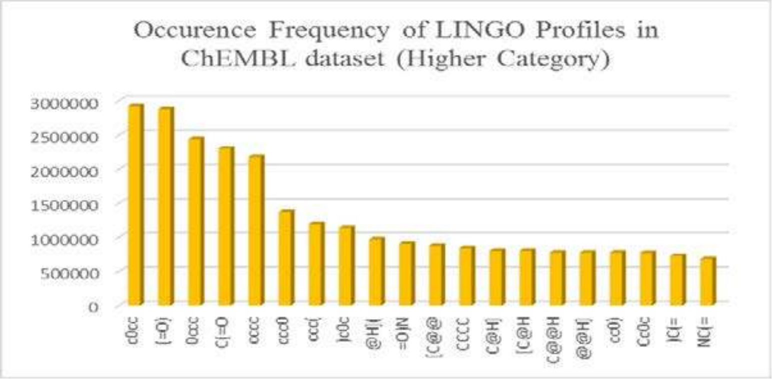


Figure 6. The frequency of occurrences of LINGO Profiles in the ChEMBL dataset (Higher Category).

In addition, repeated occurrence of a particular fragment can be seen in a compound. The fragment with highest frequency of occurrences is ‘c0cc’, which occurs 2,923,854 times. It was then followed by the fragments ‘(=O)’, ‘0ccc’, ‘C(=O’ and ‘cccc’ appearing exactly 2,878,863, 2,436,870, 2,299,177 and 2,176,560 times respectively. A significant change can be observed in the figure wherein the difference between one fragment to another is between the range of 100,000 to 400,000. Nonetheless, there are still other fragments that are in the million range and they are ‘ccc0’, ‘ccc(’ and ‘)c0c’ occurring at 1,367,947, 1,194,443 and 1,137,386 times respectively. Going down the list are the fragments that exist in the hundred-thousand range. The ‘@H](’ fragment is the highest in this sub-category appearing 965,532 times followed by the fragments ‘=O)N’, ‘[C@@’, ‘CCCC’, ‘C@H]’, ‘[C@H’, ‘C@@H’, ‘@@H]’, ‘cc0)’, ‘Cc0c’, ‘)C(=’ and ‘NC(=’, whereby the differences between them are recorded at a range of 100,000 to 300,000. Thus, we can simply conclude that the fragments in this category are gradually changing at a very high rate.

Next, **Figure 7** visualises the frequency of occurrences of the Lower Category fragments or fragments that appear less frequently in all the compounds in our ChEMBL dataset. In other words, the fragments that belongs to this category does not occur as much as those fragments in the Higher Category did. In fact, the probability of the presence of these fragments in any of the compounds in ChEMBL are very low. There are certain compounds which do not even have the fragments depicted in **Figures 6** and **7**. Apart from this, it is also observed that there are no repeated occurrences involving fragments from this category. The fragments from this category are ‘[Hg+’, ‘Hg+]’, ‘g+]c’, ‘(O[H’, ‘-]B0’, ‘]B0O’, ‘O=[F’, ‘=[Fe’, ‘Fe]O’, ‘e]O[‘, ‘]O[F’, ‘O[Fe’, ‘Fe]=’, ‘[Au-‘, ‘Au-]’, ‘u-]S’, ‘-]SS’, ‘]SS(’, ‘.O[H’, ‘Hg]C’ and ‘g]C0’. Interestingly, they all share the same occurrence frequency which is ‘1’. This means that the fragment only appears once in any of the compounds in the dataset and once they appear in any of the compounds in the dataset, they will not appear again in the remaining compounds. The difference in the occurrence frequency between the fragments in this category (Lower Category) and the fragments from the Higher Category is very large.

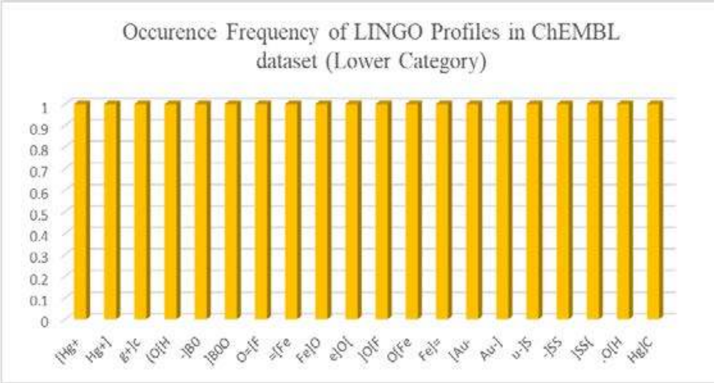


Figure 7. The frequency of occurrences of LINGO Profiles in the ChEMBL dataset (Lower Category).

Moreover, the fragments with a higher frequency of occurrence (Higher Category) tells us that despite the extensive diversity of the ChEMBL dataset, they are present in the majority of the compounds, thus, they play a crucial role in determining the behaviours and characteristics of the compound itself. For example, one of the techniques in virtual screening called similarity searching requires the comparison between a reference structure and the compounds from the database to be done in order to determine the properties of the reference structure. In this case, the fragments that exist in both the reference structure and the database compound plays a huge role as higher coefficient scores between both of them proves that they share certain chemical properties together. Besides, the majority of the compounds in the activity classes that we have selected in this study contains those fragments. This means that the interaction between the compounds with a particular target protein (activity class) are affected by the presence of these fragments.

For the fragments with a lower frequency of occurrence (Lower Category), the extremely low probability of their occurrence in the compounds proof that these fragments did not play a part in determining the behaviour or characteristics of a compound in the ChEMBL dataset. In addition, the lower number of repeated fragments from this category in a particular compound gives us the evidence that those fragments do not solely involve in determining the behaviour or characteristics of a specific compound as it is constantly outnumbered by the fragments from the Higher Category. In terms of drug-target interactions, some activity classes do not even have the fragments from this category in their compounds. The absence indicates that some compounds do not even need those fragments in order to establish an interaction with a particular target protein. Hence, we can conclude that, the fragments from the Higher Category are much more important than the fragments in the Lower Category. From the information that we have obtained above, we have created binary fingerprints or LPFP which consists of unique fragments ranging from the Higher and the Lower categories as to represent the compounds involved in the predicting DTIs and assessing their performance using LPFPs.

3.2. LINGO Profiles Fingerprints (LPFPs) generated

As discussed in Section 2.1.5, binary fingerprints of length 10,723 are generated at the end of Phase 1 (refer to **Figure 5**). The binary fingerprints consist of 0 s and 1 s, in which the '0' bit indicates the absence of a unique LINGO Profile while the '1' bit indicates the presence of a unique LINGO Profile. A sample portion of the LPFP of a compound with the presence and absence of a unique LINGO Profile is visualised in the figure below (refer to **Figure 8**).

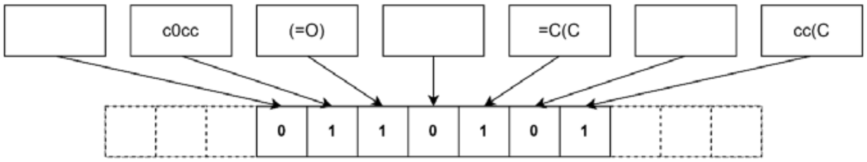


Figure 8. A portion of a LINGO Profile Fingerprint (LPFP) obtained from a compound from our ChEMBL dataset.

3.3. Comparison between the performance of ECFP4 and LPFP

As discussed in Section 2.2, the results of the performance and time taken for the compounds represented in LPFP will be benchmarked with the performance and time taken for the compounds represented in ECFP4. In this section, the said results will be compared, discussed and analysed thoroughly.

From **Table 3**, it is evident that different categories of activity classes (homogenous, heterogeneous and intermediate) produce distinct accuracy values for both descriptors, ECFP4 and LPFP. When LPFP is used, the activity class, Somatostatin receptor 2, recorded the highest accuracy value of 99.01% while Dopamine D3 receptor managed to only achieve 62.81% accuracy (the lowest among all the six activity classes) when SVM was used for its classification purposes. Further investigation led us to the observation of the pattern created among all the activity classes. It turns out that the SVM classifier together with the proposed LPFP representation of the compounds work best on homogenous activity classes, followed by heterogeneous activity classes and finally the intermediate activity classes. There was only a small difference between the average accuracy of heterogeneous and homogenous activity classes. However, a huge gap was recorded between the intermediate activity classes and the homogeneous and heterogeneous activity classes when LPFP was used, as shown in **Table 4**.

Table 3. The classification accuracy and time taken between the compounds represented in LPFP and ECFP4 for each activity class.

Activity class	Accuracy (LPFP)	Time taken for LPFP (mins)	Accuracy (ECFP4)	Time taken for ECFP4 (mins)
Cytochrome P450 2D6	0.8669	33.5629	0.8711	6.1744
Alpha-2a adrenergic receptor	0.8926	33.8055	0.8967	6.5885
G protein-coupled receptor 44	0.6926	33.7916	0.7099	6.4987
Dopamine D3 receptor	0.6281	33.8152	0.6132	6.2299
Somatostatin receptor 2	0.9901	33.0538	0.9901	6.88
Prostanoid DP receptor	0.9554	33.3866	0.9554	6.5764
Average	83.76%	33.5692	83.94%	6.3878

Table 4. The differences in classification accuracy between three categories of activity classes when compounds are represented using LPFPs and ECFP4.

Type of activity classes	Difference in accuracy (ECFP4)	Difference in accuracy (LPFP)
Homogenous	0.000	0.000
Heterogenous	−0.089	−0.093
Intermediate	−0.311	−0.312

Table 4 shows that the SVM classifier used did not perform well when compounds from intermediate classes were used with the LPFP representation. We identified that the main factor which led to this drawback was the nature of the activity classes chosen for this experiment. The majority of the compounds in the ChEMBL dataset belongs to either the heterogeneous or intermediate activity classes. To be specific, the Dopamine D3 receptor contains the highest number of compounds (1540 compounds) followed by the G protein-coupled receptor 44 (1240 compounds), as shown in **Table 1** and both of these activity classes have lower accuracy values (refer to **Table 3**). These activity classes contain compounds that are also present in other activity classes as well. In other words, the group of fragments that are necessary to determine the interactions is not unique to one target protein (activity class), but to various other targets as well. When the machine learning experiment was conducted, the rules inside the SVM model that were generated were not accurate enough to predict the interactions since the rules were made up of fragments that were mixed with other target proteins. For example, in an attempt to predict the accuracy of the Dopamine D3 activity class, the rules to create the model should not be made up of fragments that are solely unique to this activity class. There is no specific target protein or a 'set of fragments' that the model can use to create the rules to determine the interaction between them when using heterogeneous and intermediate activity classes in our experiment. As expected, a much lower accuracy was observed in the intermediate activity classes as they are made up of compounds from homogeneous and heterogeneous classes. On the other hand, homogeneous activity classes, which have a lower number of compounds, managed to obtain a very high accuracy value. This is due to the compounds in this category are unique only to its activity class whereby they do not contain compounds similar to other activity classes. Moreover, the rules created by the model are solely based on fragments derived from those unique compounds itself. When the SVM classifier model was used for DTI prediction, a high accuracy was recorded because the necessary fragments to predict the interactions were present in the rules. Based on this insight, it can be concluded that the nature of the ChEMBL dataset affects the accuracy value of a specific activity class when it is used in predicting DTIs using the SVM classifier.

Furthermore, there were not much of a difference among all the activity classes in terms of the time taken in predicting DTIs using LPFPs. On an average, 33.57 mins was needed to build one classifier model (SVM) using compounds represented in LPFPs and to predict DTIs for each activity class. The heterogeneous activity class, Dopamine D3 receptor took the longest time to build the model and predict DTIs with the total time of 33.8152 mins while the Somatostatin receptor 2 took the fastest to build the model and predict DTIs, totalling to 33.0538 mins. As mentioned earlier, the difference is very minimal, hence, it is apparent that the nature of the ChEMBL dataset does not affect the total time taken. However, it is definitely a possibility that the descriptor used is a factor in the total time taken recorded. The proposed LPFP contains a huge number of features (10,763) which then affects the predictive model to generate a high number of rules for predicting DTIs for each activity class. This is reflected in **Table 2**, where the size of the SVM classifier model built is between the size of 700 to 800 MB. Thus, many rules were generated due to the high number of features present in LPFP. Additionally, some fragments which were not involved in the prediction were necessarily eliminated, leaving only the chosen set of fragments (that were significant) for the prediction to be made for all the compounds in all the six activity classes in our ChEMBL dataset.

In terms of the usage of ECFP4, the Somatostatin receptor 2 activity class recorded the highest accuracy value of 99.01% while the Dopamine D3 receptor managed to only achieved an accuracy value of 61.32%, which is the lowest value among all the other activity classes (refer to **Table 3**). Similar to the case of using

LPFP, the SVM model worked the best with homogeneous activity classes as well, followed by the heterogeneous activity classes and then the intermediate activity classes. A similar pattern was also recorded for the huge difference in the accuracy between the intermediate activity classes and the homogenous and heterogeneous activity classes (refer to **Table 4**). However, the average time taken recorded for the prediction of DTIs for all the activity classes when using ECFP4 is recorded at 6.3878 mins, which is almost five times faster than LPFP. This is probably due to the lower number of features in ECFP4 (1024) which is about ten times less than LPFP's number of features (10,723). Based on **Table 3**, the heterogenous activity class Alpha-2a adrenergic receptor took the longest time to build the SVM model and predict DTIs while the Cytochrome P450 2D6 activity class managed to complete the classification task the fastest with just 6.1744 minutes in predicting DTIs. From the observations, it is obvious that the nature of the ChEMBL dataset do not affect the time taken. The ECFP4 descriptor used, as mentioned above, is of a smaller size in comparison to LPFP, which is the reason why when ECFP4 is used, the average time taken is five times lesser than the average time taken when LPFP is used. This is also reflected in **Table 2**, whereby the size of the SVM model built is only within the range of 15 to 30 MB, meaning that lesser rules were generated when the model was built and DTIs were predicted.

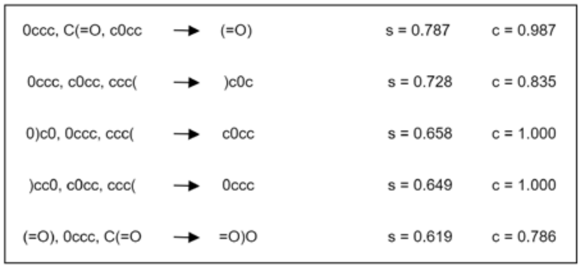
To summarise, the performance of LPFP fell short in comparison to ECFP4 by 0.18% in terms of the accuracy. This small difference proves that LPFPs are able to represent and hold important information pertaining to the compounds that will be used in the next experiment. However, for both homogeneous activity classes, LINGO Profiles Fingerprints managed to achieve the same accuracy values as when ECFP4 was used. Additionally, the accuracy of the Dopamine D3 receptor activity class using LPFP surpassed the accuracy when ECFP4 was used by about 1.49%. It is believed that the differences occur due to the mechanism that was used to encode the fragments that were present in a particular compound. Unlike LPFP, ECFP4 constitutes compound structures by methods of circular atom neighbourhoods. This method allows it to represent both the existence and the non-existence of functionality since both are circumstantial in examining molecular activity. Moreover, the excellence of ECFP4 when performing DTI prediction was also due to their ability to represent an essentially infinite number of different molecular features (including stereo chemical information) and this is crucial in creating an accurate SVM model. On the other hand, LPFP fell slightly in terms of the accuracy scores to ECFP4 due to containing some fragments which were not significant for purpose of DTI prediction. We have taken into account on all the fragments that exist in the ChEMBL dataset and this led to the creation of a SVM model which was less accurate due to the mixture of significant and insignificant features. In terms of the time taken, it was discussed above that there was a huge difference between the LPFP and ECFP4 representation of compounds whereby LPFP took almost five times longer than ECFP4 to predict DTIs for each activity class in the ChEMBL dataset. This is due to the number of features present in the fingerprints with LPFP containing 10,723 features and ECFP4 with 1024 features. In simpler words, the higher the number of features, the longer it takes to predict DTIs. Thus, in the next section, we will be discussing the results on the effects of dimensionality reduction of LPFP using ARM, to improve the performance of our proposed molecular descriptor.

3.4. Dimensionality reduction of LPFP using ARM

As discussed in Phase 3, significant rules for all the compounds in all of the six activity classes are generated using the Association Rule Mining (ARM) technique. The rules generated by ARM possesses information about known DTIs and they are adhered by the minimum support and confidence level whereby the support level values are categorized into three main categories; 60%, 70% and 80% respectively. In other words, the minimum support level set for the experiment is 60% (*s* = 60%). The support variable reflects on how frequent a particular item set appears in a transaction meanwhile the confidence variable is used to further highlight how important a frequent item set is by taking into account the specific fragments into the frequent item set. Furthermore, the fragments within the generated significant rules from LPFP were then used to create a new, compact molecular representation which is lower in size compared to the original LPFP in order to overcome the dimensionality problem in the latter. From the rules generated, the significant ones are agglomerated and the unnecessary rules are removed. This is because the significant rules are the ones containing important fragments which may determine the interaction of a compound with a particular target protein. Moreover, the unnecessary or insignificant rules contain fragments which are also found in the significant rules. Thus, they can be eliminated accordingly as well. By adopting this idea, it is believed that the important fragments chosen can improve the accuracy value and overcome the time consumption problem of LPFP when it is used to represent the compounds in an activity class.

For example, **Figure 9** shows the significant group of fragments generated using ARM that were interacting with the Prostanoid DP receptor activity class with their respective values of support and confidence levels.

There were 162 rules generated in total, but many of them were filtered out since only the significant fragments that are present in the frequent itemset shown in **Figure 9** are considered. Two of the fragments in **Figure 9** managed to achieve a support level higher than 70% while the remaining fragments surpass the minimum support level value of 60%. The fragments such as '0ccc', 'C(=O', 'c0cc', '(=O)', 'ccc(', ')c0c', '0)c0', ')cc0' and '=O)O' appeared frequently among the compounds available in this activity class. Furthermore, it is also important to note that the maximum length for all the significant rules is 4 (refer to Section 2.1.2). Hence, three of the fragments exist as antecedents and the remaining fragment exists as a consequent. Since the significant rules that were identified consist of fragments which are only significant to the Prostanoid DP receptor activity class, we can conclude that the presence of any one of the group of fragments shown in **Figure 9** in an unknown ligand will most probably interact with the Prostanoid DP receptor activity class. The same process of generating significant fragments using ARM was also applied to all the other activity classes (Cytochrome P450 2D6, Alpha-2a adrenergic receptor, G protein-coupled receptor 44, Dopamine D3 receptor and Somatostatin receptor 2) and the analysis on the significant group of fragments generated that were interacting with these activity classes were also comprehensively discussed in the supplementary materials provided (refer to Supplementary Materials).



0ccc, C(=O, c0cc	→	(=O)	s = 0.787	c = 0.987
0ccc, c0cc, ccc(	→	)c0c	s = 0.728	c = 0.835
0)c0, 0ccc, ccc(	→	c0cc	s = 0.658	c = 1.000
)cc0, c0cc, ccc(	→	0ccc	s = 0.649	c = 1.000
(=O), 0ccc, C(=O	→	=O)O	s = 0.619	c = 0.786

Figure 9. The group of fragments (rules) observed in the Prostanoid DP receptor activity class with the support level higher than the minimum value of 6.0.% (s ≥ 60%).

3.4.1 Comparison between the performance of LPFP, ECFP4 and the new and compact representation of compounds using ARM

From the significant fragments obtained earlier for each activity class (refer to **Figure 9** and Supplementary Materials), they were classified into three different groups. The first group contains fragments which have a support level value of above 60% while the second group contains fragments support level value of above 70% and finally, the third group comprises of fragments that have a support level value of at least 80%. These categorisations were made in an attempt to reduce the dimensions of the original LPFP representation. By encoding the compounds using these fragments, an alternative compound representation that met our objective was created and the results of their performance are depicted in **Tables 5** and **6**.

Table 5. The comparison of the accuracy between ECFP4, LPFP and LPFP with a support level value of at least 60%, 70% and 80%.

Activity class	Accuracy of LPFP (*s* = 80%)	Accuracy of LPFP (*s* = 70%)	Accuracy of LPFP (*s* = 60%)	Accuracy of ECFP4	Accuracy of LPFP
Cytochrome P450 2D6	0.8669	0.8835	0.8909	0.8711	0.8669
Alpha-2a adrenergic receptor	0.8943	0.9099	0.9124	0.8967	0.8926
G protein-coupled receptor 44	0.7661	0.9149	0.9281	0.7099	0.6926
Dopamine D3 receptor	0.7008	0.8579	0.8727	0.6132	0.6281
Somatostatin receptor 2	0.9884	0.9884	0.9884	0.9901	0.9901
Prostanoid DP receptor	0.9554	0.9554	0.9587	0.9554	0.9554
Average	86.20%	91.83%	92.52%	83.94%	83.76%

Table 6. The comparison of the time taken between ECFP4, LPFP and LPFP with a support level value of at least 60%, 70% and 80%.

Activity class	Time taken for LPFP (*s* = 80%) (mins)	Time taken for LPFP (*s* = 70%) (mins)	Time taken for LPFP (*s* = 60%) (mins)	Time taken for ECFP4 (mins)	Time taken for LPFP (mins)
Cytochrome P450 2D6	0.1238	0.1519	0.1640	6.1744	33.5629
Alpha-2a adrenergic receptor	0.0965	0.1291	0.1425	6.5885	33.8055
G protein-coupled receptor 44	0.2038	0.1168	0.1305	6.4987	33.7916
Dopamine D3 receptor	0.2563	0.1659	0.1850	6.4987	33.8152
Somatostatin receptor 2	0.0505	0.0536	0.0670	6.4987	33.0538
Prostanoid DP receptor	0.0765	0.0890	0.1000	6.5764	33.3866
Average	0.1346	0.1177	0.1315	6.3878	33.5692

3.4.2. Analysing compacted LPFPs generated using ARM (*s* = 80%)

The compounds that are represented with the refined LPFPs with a support level of over 80% (*s* = 80%) performed better than ECFP4 and LPFP, as shown in **Table 5**. Similar to the situation observed in Section 3.3, the SVM classifier model together with the refined and compact LPFP generated using ARM worked the best on homogeneous activity classes, followed by heterogeneous and intermediate activity classes respectively. Similar to the discussion in Section 3.3 as well, the nature of the activity classes is the factor that affects the outcome of the accuracy regardless of the classification technique used and the type of compound representation implemented. However, the compounds represented using the refined LPFP (*s* = 80%) managed to outperform other descriptors in both the intermediate activity classes. In addition, the refined LPFP also managed to achieve similar accuracies with the rest in the Prostanoid DP receptor activity class. Overall, in terms of the average accuracy, this refined and compact LPFP is better than ECFP4 and LPFP in terms of its classification accuracy whereby it is 2.24% better than ECFP4 and 2.44% better than LPFP. The improvement in terms of accuracy is due to the significant features that were derived from the significant rules that were used to encode the compounds. Despite having a much lower number of features than ECFP4 when performing the classification task, those significant features predicted the outcome accurately. In other words, the features contained in this group of fragments were sufficient to determine the interaction between a compound and a target in any of the six activity classes in our ChEMBL dataset. Furthermore, it is also observed that ECFP4 fell slightly behind the refined LPFP because it contained insignificant features that were not essential for the prediction of DTIs.

On the other hand, this refined LPFP (*s* = 80%) took lesser time to predict DTIs as shown in **Table 6**. For every activity class involved, it took only less than a minute for the model to classify and predict DTIs. When compounds from the Somatostatin receptor 2 activity class were used, the SVM model only took 3.034 s to complete the whole classification process which was the fastest among all the other activity classes involved. Subsequently, the model took the longest time when compounds from the Dopamine D3 receptor were used. Nonetheless, this new and compact LPFP was 47 times much faster than ECFP4 and 249 times much faster than the original LPFP when both of them were put into implementation. It is believed that the important features (fragments) of the compounds that were encoded using this new and compact-sized fingerprints attributed in achieving a higher accuracy value and a shorter time taken in predicting DTIs.

3.4.3. Analysing compacted LPFPs generated using ARM (*s* = 70%)

Unlike the new LPFP (*s* = 80%) as discussed above, the compounds that are represented with the refined LPFPs with a support level of over 70% (*s* = 70%) contained 13 fragments (7 more than LPFP (*s* = 80%)). From **Table 5**, LPFP (*s* = 70%) managed to outperform ECFP4 and LPFP by 7.89% and 8.07% respectively. Additionally, this new LPFP also managed to achieve a higher accuracy reading in each activity class with an exception for the Somatostatin receptor 2, which differs from ECFP4 and LPFP by 0.17%. In addition to that, LPFP (*s* = 70%) performed better than LPFP (*s* = 80%) by 5.63%. The increase in the number of significant features in LPFP (*s* = 70%) has caused the accuracy to be higher than LPFP (*s* = 80%). This means that there were more significant features that was being used in order to encode the compounds and when these compounds were used to predict the interactions, they were able to predict more accurately.

Alternatively, similar to the LPFP (*s* = 80%), LPFP (*s* = 70%) also took a shorter time to predict DTIs, which was approximately 7.062 s on an average to complete the whole classification process as shown in **Table 6**. In fact, it was also much faster than LPFP (*s* = 80%) by 1.014 s. For each of the activity class, LPFP (*s* = 70%) took merely a minute to build the classifier model and accurately predict DTIs. Similar to the situation observed previously, when compounds from the Somatostatin receptor 2 activity class were used, SVM only took 3.213 s to complete the whole classification process, which was the quickest among all the other activity classes involved. Furthermore, SVM when set aside took the longest time to finish the entire procedure when compounds from the Dopamine D3 receptor were utilised. In general, LPFP (*s* = 70%) was

54 times considerably quicker than ECFP4 and 285 times significantly faster than LPFP when two of them were inculcated into usage.

3.4.4. Analysing compacted LPFPs generated using ARM (*s* = 60%)

Moving on, another LPFP was also created based on the rules that had a support level value of above 60%. In this new LPFP (*s* = 60%), a total of 17 fragments were extracted (4 more than LPFP (*s* = 70%) and 11 more times than LPFP (*s* = 80%). The increase in the number of fragments was attributed by the increased number of rules that belonged to this category. As shown in **Table 5**, LPFP (*s* = 60%) overcame the accuracy values of ECFP4 and LPFP by 8.58% and 8.76% respectively. Additionally, LPFP (*s* = 60%) also managed to accomplish a higher accuracy value for each activity class with the exception of the Somatostatin receptor *2* which differs from the other two descriptors (ECFP4 and LPFP) by 0.17%, similar to the case observed in LPFP (*s* = 70%). Besides, LPFP (*s* = 60%) is also deemed superior to LPFP (*s* = 70%) and LPFP (*s* = 80%) by 0.69% and 6.32% respectively. The improvement in terms of accuracy is due to the increase in the number of significant features that were derived from the significant rules that were used to encode the compounds earlier. Despite having a much lower number of features than ECFP4, LPFP, LPFP (*s* = 80%) and LPFP (*s* = 70%), when performing the classification task, those significant features extracted predicted the outcome accurately, resulting in a higher accuracy value.

In terms of the time taken, from **Table 6**, LPFP (*s* = 60%) took 7.89 s on average to complete the whole classification process. It was also much faster than LPFP (*s* = 80%) by 0.186 s, but it is slower than LPFP (*s* = 70%) by 0.828 s. For every activity class involved, it also took merely s to build the classifier and precisely predict DTIs. Similar to the circumstances illustrated by the previous two new LPFPs, when compounds from the Somatostatin receptor 2 activity class were utilised, SVM only took 4.02 s to complete the whole classification process. This was the fastest among all the other activity classes involved. In addition, Dopamine D3 receptor took the longest time to complete its predictions. All in all, LPFP (*s* = 60%) was 48 times impressively faster than ECFP4 and 255 times faster than LPFP when two of them were implemented. Hence, it is prominent that the vital features (fragments) of the compounds that were encoded using this unique smaller size fingerprint accomplished higher accuracy values and reduced the time taken in predicting DTIs. Henceforth, by presenting LPFP (*s* = 60%), this study has successfully figured out on how to beat the dimensionality issues as specified before.

3.4.5. Testing the performance of LPFP on three new activity classes

As thoroughly discussed in the earlier sections of this paper, the convincing performance of the three new LPFPs generated (*s* = 60%), (*s* = 70%) and (*s* = 80%) for all the six activity classes previously has led us to further test and investigate their performance in three new activity classes obtained from ChEMBL, which are the Urotensin II receptor, Focal adhesion kinase 1 and Vanilloid receptor activity classes. These activity classes are selected and categorized into their homogenous, intermediate and heterogeneous nature respectively and the results of the testing are summarized in the table below.

As observed in **Table 7**, all the three new activity classes achieved an accuracy of over 80%. However, in terms of the average, LPFP (*s* = 60%) outperformed the other LPFPs (*s* = 70% and *s* = 80%) by 0.79% and 2.15% respectively. When the homogeneous activity class Urotensin II receptor was used, the accuracy recorded for all the three LPFPs were the highest (over 98%), while the lowest accuracy recorded was when the heterogeneous activity class Vanilloid receptor was utilized. The difference in terms of the results is due to the nature of the activity class itself whereby variations in terms of the number of compounds and the property associated with each activity classes causes the classification accuracy to be different than one another. Conclusively, it is proven that these new LPFPs (*s* = 60%), (*s* = 70%) and (*s* = 80%) were not only able to overcome the high dimensionality problem, but they can also be used to accurately predict the outcome of any activity class with an average classification accuracy of over 90%.

Table 7. The comparison of the classification accuracy of three new activity classes using the three new LPFPs generated earlier.

Type of activity class	Activity class	Accuracy for LPFP (*s* = 60%)	Accuracy for LPFP (*s* = 60%)	Accuracy for LPFP (*s* = 80%)
Homo	Urotensin II receptor	0.9885	0.9872	0.9866
Inter	Focal adhesion kinase 1	0.9643	0.9643	0.9617
Hetero	Vanilloid receptor	0.8787	0.8564	0.8188
Average		94.38%	93.59%	92.23%

4. Conclusion

To conclude, LINGO Profiles can be used to perform DTI prediction by representing the compounds in a binary fingerprint format called LINGO Profiles Fingerprint or LPFP. LPFP is developed by considering all the unique fragments available in the compounds of all the activity classes in the ChEMBL dataset. From the experiments conducted on all six target proteins (activity classes) using the SVM classifier, LPFP was observed to perform quite well alongside the state-of-art ECFP4 fingerprints. However, there is a slight difference in terms of the accuracy and the time taken is much longer when DTIs were predicted using the compounds that are represented using LPFP. This is due to high number of features (high dimensionality) of LPFP that are associated with the compounds in the activity classes. These features are made up of significant and insignificant features respectively. To solve this problem, Association Rule Mining or ARM was used to overcome the dimensionality problem in LPFP. ARM provides knowledge pertaining to the prediction of drug-target interactions based on the rules that are generated. The rules generated using ARM contains significant fragments that are believed to hold important information about the compounds that exist in all the activity classes involved. Hence, three new variants of LPFPs were developed using features identified by ARM with support values of 60%, 70% and 80%. As a proof, the accuracy obtained using these three new LPFPs; LPFP (*s* = 60%), LPFP (*s* = 70%) and LPFP (*s* = 80%) managed to surpass the accuracy of the original LPFP as well as the ECFP4 fingerprints (over 80% consistently). Besides, the time taken to build the classifier model are also significantly reduced (250 times faster than ECFP4 and LPFP) when compounds are represented with a lower number of features. To put it briefly, the ARM technique can act as an alternative to other in silico methods when predicting DTIs. In the near future, there are various other experiments that could be conducted as expansions to the ones depicted in this paper. For example, more activity classes from various categories are to be included to assess the performance in terms of the accuracy and the time taken to predict DTIs. Furthermore, the reduced LPFPs can also be benchmarked with different kinds of molecular descriptors such as ECFP2, ECFP6, ECFP8, MACCS as well as FCFP. Since different molecular descriptors have different algorithms in encoding the compounds, the variation in the accuracy scores ranging across different activity classes can be observed accordingly. It is also proven that the significant fragments contained within the rules are vital because the presence of any group of fragments within a compound will determine the interaction between the said compound with a specified target protein.

Supplementary materials

Figure S1: the group of fragments (rules) observed in the Somatostatin receptor 2 activity class for *s* ≥ 60%; Figure S2: the group of fragments (rules) observed in the Alpha-2a adrenergic receptor activity class for s ≥ 60%; Figure S3: the group of fragments (rules) observed in the Cytochrome P450 2D6 activity class for s ≥ 60%; Figure S4: the group of fragments (rules) observed in the Dopamine D3 receptor activity class for s ≥ 60%; Figure S5: the group of fragments (rules) observed in the G protein-coupled receptor 44 activity class for s ≥ 60%.

Author contributions

Conceptualization, MJMJ and NHAHM; methodology, MJMJ; software, MJMJ; validation, NHAHM; formal analysis, MJMJ; investigation, MJMJ; resources, MJMJ, AKAK and NHAHM; data curation, MJMJ; writing—original draft preparation, MJMJ, AKAK and NHAHM; writing—review and editing, AKAK; visualization, MJMJ and AKAK; supervision, NHAHM; project administration, NHAHM; funding acquisition, NHAHM.

Funding

This work is supported by the Ministry of Higher Education (MOHE), Malaysia via the Fundamental Research Grant Scheme (FRGS) FRGS/1/2019/ICT02/USM/02/4-203.PKOMP.6711800.

Conflict of interest

The authors declare no conflict of interest.

References

1. Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. *British Journal of Pharmacology* 2011; 162(6): 1239–1249. doi: 10.1111/j.1476-5381.2010.01127.x
2. Drews J. Drug discovery: A historical perspective. *Science* 2000; 287(5460): 1960–1964. doi: 10.1126/science.287.5460.1960
3. Hann M, Green R. Chemoinformatics—A new name for an old problem? *Current Opinion in Chemical Biology* 1999; 3(4): 379–383. doi: 10.1016/S1367-5931(99)80057-X
4. Hung CL, Chen CC. Computational approaches for drug discovery. *Drug Development Research* 2014; 75(6): 412–418. doi: 10.1002/ddr.21222
5. Agamah FE, Mazandu GK, Hassan R, et al. Computational/in silico methods in drug target and lead prediction. *Briefings in Bioinformatics* 2020; 21(5): 1663–1675. doi: 10.1093/bib/bbz103
6. Katsila T, Spyroulias GA, Patrinos GP, Matsoukas MT. Computational approaches in target identification and drug discovery. *Computational and Structural Biotechnology Journal* 2016; 14: 177–184. doi: 10.1016/j.csbj.2016.04.004
7. Chen R, Liu X, Jin S, et al. Machine learning for drug-target interaction prediction. *Molecules* 2018; 23(9): 2208. doi: 10.3390/molecules23092208
8. Yamanishi Y, Kotera M, Kanehisa M, Goto S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 2010; 26(12): 10.1093/bioinformatics/btq176
9. Glick NM, Davies JW, Jenkins JL. Prediction of biological targets for compounds using multiple-category bayesian models trained on chemogenomics databases. *Journal of Chemical Information Modeling* 2006; 46(3):

1124–1133. doi: 10.1021/ci060003g

10. Wen M, Zhang Z, Niu S, et al. Deep-learning-based drug-target interaction prediction. *Journal of Proteome Research* 2017; 16(4): 1401–1409. doi: 10.1021/acs.jproteome.6b00618
11. Prado-Prado F, García-Mera X, Abeijón P, et al. Using entropy of drug and protein graphs to predict FDA drug-target network: Theoretic-experimental study of MAO inhibitors and hemoglobin peptides from Fasciola hepatica. *European Journal of Medicinal Chemistry* 2011; 46(4): 1074–1094. doi: 10.1016/j.ejmech.2011.01.023
12. Nasution AK, Wijaya SH, Kusuma WA. Prediction of drug-target interaction on jamu formulas using machine learning approaches. In: Proceedings of 2019 International Conference on Advanced Computer Science and information Systems (ICACSIS); 12–13 October 2019; Bali, Indonesia. pp. 169–174.
13. Rodríguez-Pérez R, Vogt M, Bajorath J. Support vector machine classification and regression prioritize different structural features for binary compound activity and potency value prediction. *ACS Omega* 2017; 2(10): 6371–6379. doi: 10.1021/acsomega.7b01079
14. Riddick G, Song H, Ahn S, et al. Predicting in vitro drug sensitivity using Random Forests. *Bioinformatics* 2011; 27(2): 220–224. doi: 10.1093/bioinformatics/btq628
15. Shi H, Liu S, Chen J, et al. Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics* 2019; 111(6): 1839–1852. doi: 10.1016/j.ygeno.2018.12.007
16. Vidal D, Thormann M, Pons M. LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *Journal of Chemical Information and Modeling* 2005; 45(2): 386–393. doi: 10.1021/ci0496797
17. Abdo A, Pupin M. LINGO-DL: A text-based approach for molecular similarity searching. *Journal of Computer-Aided Molecular Design* 2021; 35(5): 657–665. doi: 10.1007/s10822-021-00383-9
18. bin Javeed MJ, Malim NHAH. Storage consumption reduction using improved inverted indexing for similarity search on LINGO Profiles. *International Journal of Advanced Computer Science and Applications* 2019; 10(5): 2019. doi: 10.14569/IJACSA.2019.0100505
19. Siswanto S, Liong TH, Shaufiah. Dimensionality reduction for association rule mining with IST-EFP algorithm. In: Proceedings of 2015 3rd International Conference on Information and Communication Technology (ICoICT); 27–29 May 2015; pp. 184–187.
20. Li PH, Lee T, Youn HY. Dimensionality reduction with sparse locality for principal component analysis. *Mathematical Problems in Engineering* 2020; 2020: 1–12. doi: 10.1155/2020/9723279
21. Malavika S, Phil M, Selvam K. Reduction of dimensionality for high dimensional data using correlation measures. Available online: http://www.ripublication.com (accessed on 27 July 2023).
22. Fujiwara T, Kwon OH, Ma KL. Supporting analysis of dimensionality reduction results with contrastive learning. *arXiv* 2019; arXiv:1905.03911. doi: 10.1109/TVCG.2019.2934251
23. Mahmud SMH, Chen W, Jahan H, et al. Dimensionality reduction based multi-kernel framework for drug-target interaction prediction. *Chemometrics and Intelligent Laboratory Systems* 2021; 212: 104270. doi: 10.1016/j.chemolab.2021.104270
24. Terol RM, Reina AR, Ziaei S, Gil D. A machine learning approach to reduce dimensional space in large datasets. *IEEE Access* 2020; 8: 148181–148192. doi: 10.1109/ACCESS.2020.3012836
25. Gardiner EJ, Gillet VJ. Perspectives on knowledge discovery algorithms recently introduced in chemoinformatics: Rough set theory, association rule mining, emerging patterns, and formal concept analysis. *Journal of Chemical Information and Modeling* 2015; 55(9): 1781–1803. doi: 10.1021/acs.jcim.5b00198
26. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. *Nucleic Acids Research* 2017; 45(D1): D945–D954. doi: 10.1093/nar/gkw1074
27. Arif S, Khan NZS, Malim N, Zainudin S. Retrieval performance using different type of similarity coefficient for virtual screening. *Research Journal of Applied Sciences, Engineering and Technology* 2015; 9(5): 391–395. doi: 10.19026/rjaset.9.1418
28. Heikamp K, Bajorath J. Support vector machines for drug discovery. *Expert Opinion on Drug Discovery* 2014; 9(1): 93–104. doi: 10.1517/17460441.2014.866943
29. Steinbeck C, Han Y, Kuhn S, et al. The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information Comput Sciences* 2003; 43(2): 493–500. doi: 10.1021/ci025584y