ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

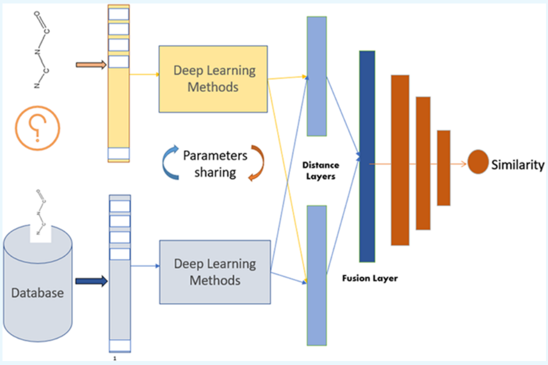
Similarity-Based Virtual Screen Using Enhanced Siamese Deep Learning Methods

Mohammed Khaldoon Altalib* and Naomie Salim

Cite This: *ACS Omega* 2022, 7, 4769−4786

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



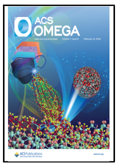
ABSTRACT: Traditional drug production is a long and complex process that leads to new drug production. The virtual screening technique is a computational method that allows chemical compounds to be screened at an acceptable time and cost. Several databases contain information on various aspects of biologically active substances. Simple statistical tools are difficult to use because of the enormous amount of information and complex data samples of molecules that are structurally heterogeneous recorded in these databases. Many techniques for capturing the biological similarity between a test compound and a known target ligand in LBVS have been established. However, despite the good performances of the above methods compared to their prior, especially when dealing with molecules that have homogeneous active structural elements, they are not satisfied when dealing with molecules that are structurally heterogeneous. Deep learning models have recently achieved considerable success in a variety of disciplines due to their powerful generalization and feature extraction capabilities. Also, the Siamese network has been used in similarity models for more complicated data samples, especially with heterogeneous data samples. The main aim of this study is to enhance the performance of similarity searching, especially with molecules that are structurally heterogeneous. The Siamese architecture will be enhanced using two similarity distance layers with one fusion layer to further improve the similarity measurements between molecules and then adding many layers after the fusion layer for some models to improve the retrieval recall. In this architecture, several methods of deep learning have been used, which are long short-term memory (LSTM), gated recurrent unit (GRU), convolutional neural network-one dimension (CNN1D), and convolutional neural network-two dimensions (CNN2D). A series of experiments have been carried out on real-world data sets, and the results have shown that the proposed methods outperformed the existing methods.

1. INTRODUCTION

Drug development is a lengthy and complicated procedure that ends in the creation of new drug production. In the course of conventional drug research and development, a biomolecular target is identified and the experiments for high-performance screening are performed to identify bioactive compounds for specified goals. The development of high-performing research testing is expensive and time-consuming. This process includes specialized laboratories with chemical and biological libraries.[1] In fact, the probability of success is low, and the acceptance and widespread use of approximately 1 out of 5000 identified drug applicants are estimated.[2] The increased computer capabilities, on the other hand, allowed several million chemical compounds to be screened at an acceptable pace and cost. The virtual screening technique is a computational method for searching small molecules in huge libraries and choosing the most likely binding structures with a drug objective.[3−6] Virtual screening (VS) is conducted in the early discovery phases in which broad chemical libraries comprise the most promising lead compounds. In the last few years, the development of drugs has been accelerated by virtual screening (VS). Two main virtual screening techniques exist, namely, structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS).[7] The SBVS techniques seek indirect compounds that are appropriate for the biological objective binding site. The central technology of SBVS methods is molecular docking.[8] On the other hand, the LBVS approach is used constantly for the prediction of molecular properties and for measuring molecular similarity because the method to represent the molecules is easy and accurate. The significance of applications of similarity searching stems from the importance of lead optimization in drug discovery programs, in which close neighbors are looking into an initial lead compound to find decent compounds.[9−12]

Recently, modern deep learning (DL) techniques were introduced in many fields, and they developed in the last years, opening a new door for researchers. The success of DL

Received: August 23, 2021
Accepted: January 17, 2022
Published: February 3, 2022



techniques benefits from the rapid growth of the DL algorithms and the advancement of high-performance computing techniques. Moreover, DL techniques have fewer generalization errors, which allow them to achieve reasonable results on certain benchmarks or competitive tests and make more precise predictions regarding molecular properties.[13−18] Also, features can be automatically discovered from input data using deep learning techniques.[3,19,20] In addition, the Siamese network is commonly used for solving image similarity and text similarity issues. It has been used for more complicated data samples, especially with heterogeneous data samples, with various dimensionalities and type characteristics.[21,22]

Information about different aspects of biologically active compounds is held in a variety of databases. Some databases contain classes of molecules that have structurally homogeneous active elements like the MDL Drug Data Report (MDDR_DR2) data set, and other databases contain classes of molecules that are structurally heterogeneous like MDDR_DR3 and Maximum Unbiased Validation (MUV); however, the vast volume of information stored and complex data samples of molecules that are structurally heterogeneous make it difficult to carry out simple statistical tools. In this study, the Siamese deep learning model will be enhanced by using two distance layers and then a fusion layer that combines the results from two distances layers, it is appropriate to combine them to further improve the similarity measurements between molecules, particularly when dealing with different types of descriptors. In some models, multiple layers have been added after the fusion layer to improve the retrieval recall. In this architecture, several methods of deep learning have been used here which are long short-term memory (LSTM-RNN), gated recurrent unit (GRU-RNN), both are in the recurrent neural network, convolutional neural network-one dimension (CNN1D), and convolutional neural network-two dimensions (CNN2D). The following are the paper's main contributions:

- The Siamese deep learning model will be enhanced using two distance layers and then a fusion layer that combines the results from two distance layers to add further improvements for the similarity measurements between molecules, particularly when dealing with different types of descriptors, and then adding many layers after the fusion layer for some models to improve the retrieval recall.
- In comparison to benchmark approaches, the suggested method demonstrated encouraging results in terms of overall performance, especially when dealing with heterogeneous classes of molecules.

2. RELATED WORK

Similarity-based virtual screening is widely considered to be one of the essential aspects of drug discovery. Several approaches were used to increase the retrieval effectiveness of the similarity searching methods. The 2D similarity methods have become widely used and very common. The fundamental theory behind the calculation of molecular similarity is that structurally similar molecules seem to be more likely to possess similar properties than structurally dissimilar molecules. The purpose of similarity searching, therefore, is to retrieve molecules that are structurally very similar to the reference structures of the consumer. Various coefficient methods allow for the quantification of the similarity/difference between molecule pairs. Many other studies have tested the output of several similarity coefficients, showing that the Tanimoto coefficient surpassed other similarities.[23−26] The Tanimoto coefficient has thus become the most common indicator of the similarity of chemical compounds used in chemoinformatics. Different approaches have been used over the years to enhance the performance of search methods for similarities. Some experiments attempted to incorporate methods from various disciplines. Many parallels between the retrieval of text information and cheminformatics have already indicated that techniques developed for the retrieval of text documents may be employed to improve the similitude of molecular searching.[27] Therefore, some approaches to molecular similarities, such as the Bayesian inference network, used by virtual ligand screening were originally based on text retrieval domains. In virtual screening, for example, Abdo et al.[28,29] have used the Bayesian network, which outpaced Tanimoto. In addition, the reweighting techniques were used in the text field to model retrieval of documents and adapted in the cheminformatic field in the retrieval model.[28,30] The fragment reweighting techniques were also used by Ahmed et al. to strengthen the Bayesian network.[28] Al-Dabagh used the concepts of quantum mechanics theory to enhance molecular similarity searching and molecular ranking of chemical compounds in LBVS.[31] Himmat M. created a new similarity measure by reweighting various bit strings and derived it from existing similarity measures. In addition, the author proposed ranking strategies for developing a substitute ranking technique.[32] Nasser and colleagues employed deep belief networks (DBNs) to reweight molecular features where many descriptors were utilized, each representing distinct relevant aspects, and integrated all new features from all descriptors to provide a new descriptor for similarity searches.[33]

In recent years, new technologies of deep learning (DL) have been adopted and applied in drug discovery and bioinformatics and cheminformatics studies, opening a new door to computational decision making and to assist in the understanding of molecular mechanisms and the development of new therapeutic options for a variety of diseases.[20,34] Gómez-Bombarelli et al. proposed an autoencoder model that produces new molecules by converting discrete molecular representations to multidimensional continuous representations.[35,36] Skalic et al. proposed a new model to produce new molecules by converting the seed compound into a three-dimensional (3D) representation using a variational autoencoder and then sequencing SMILES tokens using convolutional and recurrent neural network systems to explore uncharted areas of the chemical space that still have lead compound-like characteristics.[37] Gao et al. proposed a generative network complex (GNC) model to create new drug like molecules using gradient descent in the latent space of an autoencoder for multiproperty optimization.[38] Hamza et al. used CNN to determine its precision during the prediction of orphan compound activities.[39] Also, Mendolia et al. used convolutional neural networks (CNNs), which are intended to identify a set of candidate compounds for a specific target protein in terms of their biological activity, and both 1D and 2D CNNs were trained separately to test the performance of every single descriptor.[40] Moreover, several researchers have proposed to exploit RNN-based methods for chemoinformatics. The majority of the researchers have utilized the model as a prediction or classifier model. Wan and Zeng proposed a model for compound—protein interaction prediction using DL methods, in which they adopted a commonly used NLP approach called feature incorporation.[41] Their model was built into multidimensional vectors, both ligand details (molecular fingerprints) and protein

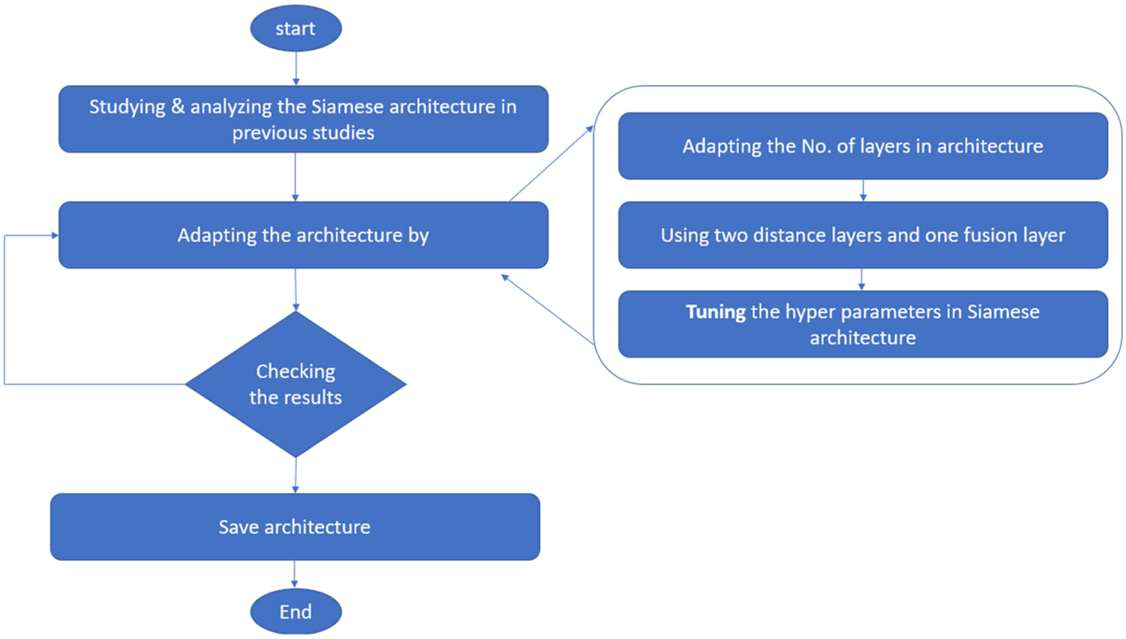

start
Studying & analyzing the Siamese architecture in previous studies
Adapting the architecture by
Adapting the No. of layers in architecture
Using two distance layers and one fusion layer
Tuning the hyper parameters in Siamese architecture
Checking the results
Save architecture
End

Figure 1. Flowchart of steps for enhancing the Siamese architecture.

sequences. Also, SMILES representation of molecules is used in the RNN model. The RNN model was used to learn SMILES' coding grammar, which can be converted into a molecular graph.[42] In addition, Goh et al. used SMILES as an input feature to the RNN model for predicting the molecular properties.[43]

Furthermore, other studies reported that deep learning methods in the Siamese architecture as a similarity model produce the best performance that can be used with more complicated data samples, especially with heterogeneous data samples, with various dimensionalities and type characteristics.[21,22] For example, Yu et al. employed CNN Siamese architectures to assess whether two people are from the same family, allowing missing people to be reunited with their relatives.[44] Jonas et al. used the LSTM Siamese neural network to calculate the similarity between sentences.[45] In this method, an exponential Manhattan distance was used to measure the similarity between two sentences. In the drug discovery domain, Dhami et al. was using images as an input to predict drug interactions in a Siamese convolution network architecture.[46] Jeon et al. proposed a method to use MLP Siamese neural networks (ReSimNet) in structure-based virtual screening (SBVS) to calculate the distance by cosine similarity.[22]

Despite the good performances of the above methods compared to their prior, especially when dealing with molecules that have homogeneous active structural elements like classes of molecules in the MDL Drug Data Report data set MDDR_DR2, however, the performances are not satisfied when dealing with molecules of structurally heterogeneous nature like classes of molecules in the MDL Drug Data Report data set MDDR_DR3 and Maximum Unbiased Validation (MUV) data set. The main goal of this research is to improve the retrieval effectiveness of the similarity model, especially with molecules that are structurally heterogeneous, and because of the power of deep learning for dealing with big data and the power of the Siamese architecture for dealing with complicated data samples, especially with heterogeneous data samples; therefore, they have been used in this study. Many methods of deep learning will be examined as a similarity model through the enhanced Siamese architecture. These methods of deep learning include long short-term memory (LSTM-RNN), gated recurrent unit (GRU-RNN), convolutional neural network-one dimension (CNN1D), and convolutional neural network-two dimensions (CNN2D).

3. METHODS

A Siamese neural network contains two artificial neural networks that are the same, each able to handle the hidden input data representation, which have to be linked to a final layer using a distance layer to predict whether or not two vectors fall under the same category. Since all of the weights and biases are related, the networks that make up the Siamese architecture are called twins, which means that both networks are symmetric. Both error backpropagation and feed-forward perceptron are used by the two neural networks during training. Therefore, it has been used for more complicated data samples, especially with heterogeneous data samples, with various dimensionalities and type characteristics. In this paper, the Siamese deep learning model will be enhanced. Figure 1 shows the flowchart of steps for enhancing the Siamese architecture.

The steps for enhancing the Siamese architecture of deep learning methods include the following:

1 Studying and analyzing many models of Siamese architectures in different fields, like Dhami et al. and Jeon et al. in the field of structure-based virtual screening and Jonas et al. in the text field.

2 All previous studies used one distance layer. In this study, two distance layers are used, and then, one fusion layer combines the results from distance layers. The reason for using more than one distance layer is to further improve the similarity measurements between molecules, particularly when dealing with different types of descriptors.

3 In general, there are two inputs and one output in this architecture; the output value represents the degree of similarity between the inputs. In this study, many layers have been added after the fusion layer for some models to improve the retrieval recall.

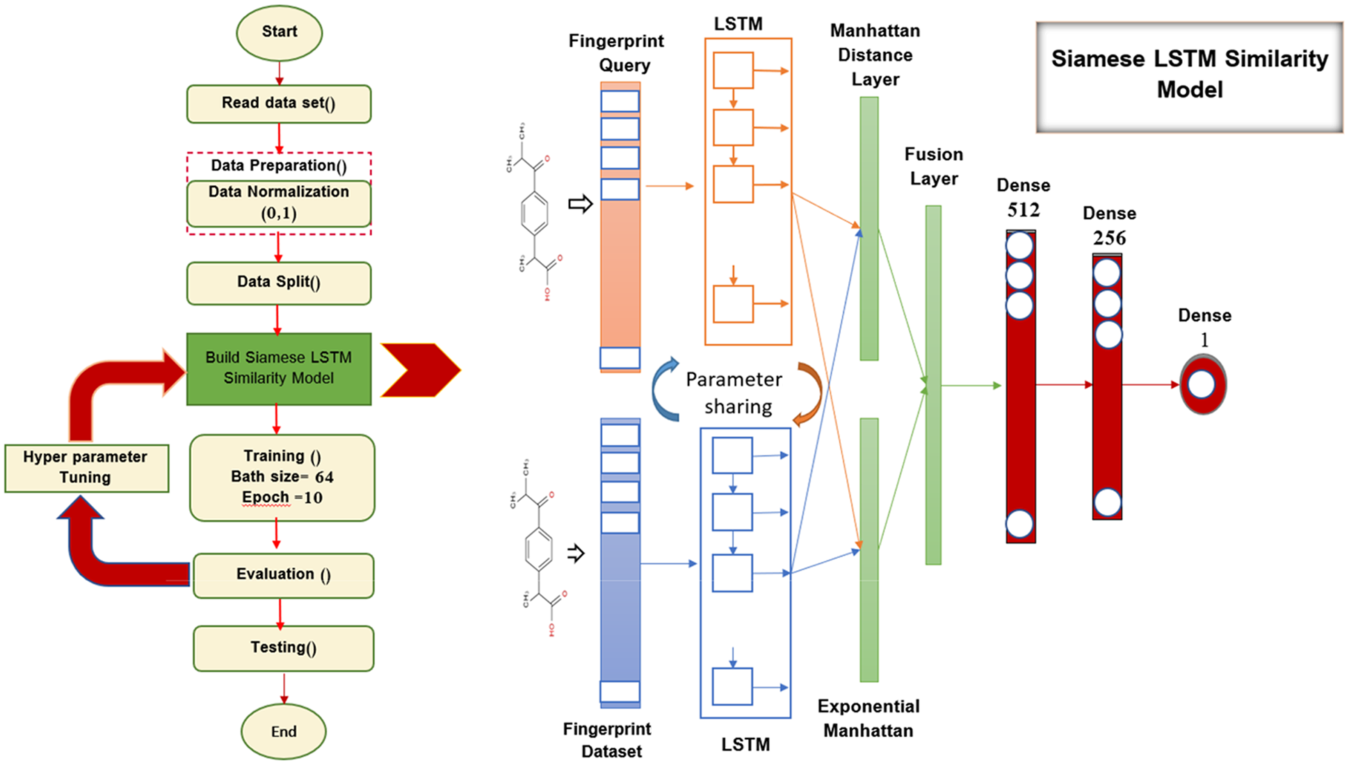


Figure 2. Architecture of the enhanced Siamese RNN-LSTM similarity model.

4 The hyperparameters of the Siamese deep learning similarity model such as the number of epochs and batch size, optimization, and the activation function are tuned to achieve a good retrieval recall result.

Here, four methods of deep learning have been used in this architecture; these methods include these methods include long short-term memory (LSTM-RNN), gated recurrent unit (GRU-RNN), convolutional neural network-one dimension (CNN1D), and convolutional neural network-two dimensions (CNN2D). The following subsections explain each of the methods individually.

3.1. Enhanced Siamese RNN Similarity Model.

Recurrent neural networks are artificial neural networks that form the link between nodes by means of a directed diagram along a time stream. The recurrent neural networks use internal state memory for sequence processing compared to neural feed-forward networks. The recurrent neural networks' dynamic behavior enables them to be very helpful and applicable to audio processing, handwriting recognition, and many such applications. However, recurring neural networks face the problems of vanishing gradients during backpropagation. If the gradient value is extremely small, it cannot lead to effective learning. As a short-term memory solution, LSTM and GRU have been developed. LSTM and GRU have been developed as a solution for short-term memory. They have internal mechanisms, which can monitor the flow of information, called gates.[47]

3.1.1. Enhanced Siamese LSTM Similarity Model.

Long short-term memory (LSTM), is an RNN structure with feedback links that allow everything that a Turing machine can do or compute. A single LSTM unit is made up of a cell, an input gate, an output gate, and a forgotten door, allowing the cell to arbitrarily record the value. The data flow in and out of the LSTM cell is monitored by gates.[48] An enhanced Siamese LSTM structure was used to determine how similar two molecules are, so the architecture has two inputs, one from the query and the other from the fingerprint data set, representing the fingerprint of molecules. The one-output architecture represents the degree of similarity, which means that the output has two classes: if the value is 1, it means high similarity, and if the value is 0, it means high dissimilarity; the weights have also been linked in this architecture so that LSTMa = LSTMb.

In this model, each input layer has two cell dimensions (32,32), each of this matrix is linked to one molecular fingerprint feature, and then each input layer is linked to distance layers; two distances have been used: the first one is the Manhattan distance,[49] which can be represented as

$$d_{AB} = |f_A - f_B| \tag{1}$$

where d_{AB} is the Manhatten distance, f_A is the feature of molecule's query, and f_B is the feature of molecule's data set, and the second distance is the exponential Manhattan distance,[45] which can be given as

$$E_{AB} = \exp(-|f_A - f_B|) \tag{2}$$

where E_{AB} is the exponential Manhatten distance, f_A is the feature of molecule's query, and f_B is the feature of molecule's data set.

Next, a fusion layer is added to fuse between two distance layers (Manhattan, exponential Manhattan). Then, three layers are added after the fusion layer; the cells in these layers are 512, 256, and 1, respectively. The output is one of the two cases: 1, meaning the two input molecules are similar, and 0, meaning the two input molecules are dissimilar. The ReLU activation function has been used for all dense layers except the last one, in which the sigmoid activation function has been used. Moreover, the RMSprop optimizer has been used, the loss function is binary_crossentropy, and the batch size is 64. The

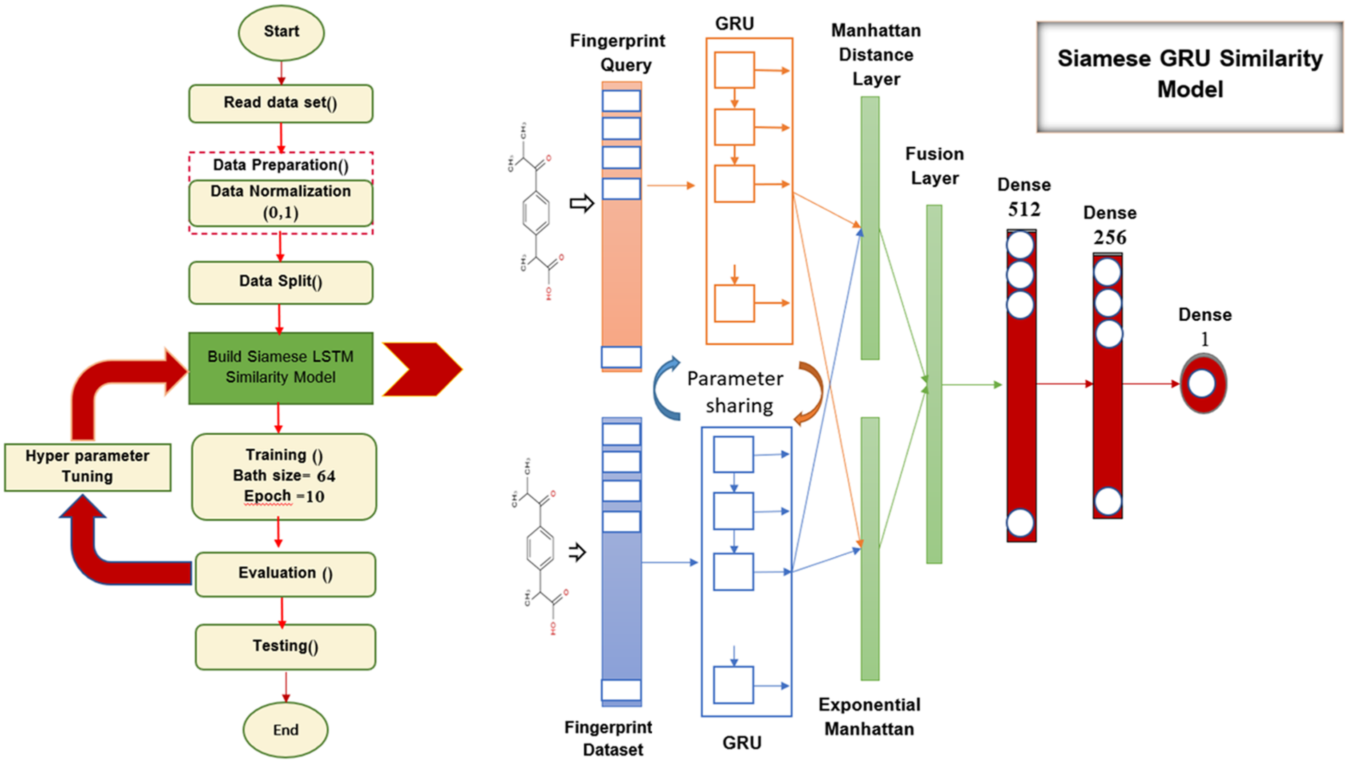


Figure 3. Architecture of the enhanced Siamese RNN-GRU similarity model.

architecture of the Siamese RNN-LSTM similarity model is illustrated in Figure 2.

3.1.2. Enhanced Siamese GRU Similarity Model. The GRU, recognized as the gated recurrent unit, is an RNN architecture that is similar to LSTM units. Instead of the LSTM input, output, and forget gate, the GRU consists of a reset gate and an update gate. The update gate lets the model decide how much of the previous knowledge (from previous time steps) needs to be followed on to the future.[50] An enhanced Siamese GRU framework has been used to determine how similar two molecules are; therefore, the architecture has two inputs, representing the fingerprint of molecules, one from the query and the other from a data set of the fingerprint. Also, the architecture has one output that represents the degree of similarity; 1 means high similarity or 0 means high dissimilarity. Also, the weights have been tied such that GRUa = GRUb in this architecture; each input layer has two dimensions (32,32) of cells, each one connected to one feature of the molecular fingerprint, and then each input layer is connected to distance layers.

Two distances are used (as mentioned in the previous subsection): the first one is the Manhattan distance, and the second distance is the exponential Manhattan distance. Next, a fusion layer is added to fuse between two distance layers (Manhattan, exponential Manhattan). Then, three layers are added after the fusion layer; the cells in these layers are 512, 256, and 1, respectively. The output is one of the two cases: 1, meaning the two input molecules are similar, and 0, meaning the two input molecules are dissimilar. The ReLU activation function has been used for all layers except the last one, in which the sigmoid activation function has been used. Moreover, the RMSprop has been used, binary_crossentropy is the loss function, and 64 is the batch size. Figure 3 demonstrates the architecture of the Siamese RNN-GRU similarity model.

3.2. Enhanced Siamese CNN Similarity Model.

The CNN is a type of high feed-forward network that can be easily trained and generalized compared to other networks with connectivity between the adjacent layers.[51,52] In this work, the Siamese CNN framework has been used to determine how similar two molecules are. CNN1D (one dimension) and CNN2D (two dimensions) have been used in this architecture as follows.

3.2.1. Enhanced Siamese CNN1D Similarity Model. CNNs, whether they have one, two, or three dimensions, function the same way. The difference is the input data structure and how the filtration, often referred to as a convolution kernel or detector of features, travels over the data. In this work, the Siamese CNN1D framework is used to calculate the similarity between a reference structure of molecular and a database structure of molecular based on fingerprints. Thus, the architecture has two inputs, representing the fingerprint of molecules, one from the reference structure (query) and the other from the database structure. Also, the architecture has one output, representing the degree of similarity. If the value is 1, it means high similarity, and if the value is 0, it means high dissimilarity. Also, weights have been tied such that CNN1Da = CNN1Db in this architecture; there are two inputs, each input layer of convolution neural network (1D-CNN) received the molecular fingerprint, followed by another layer of the 1D convolution neural network (1D-CNN), followed by a max pooling size equal 2. The layer is formed by 64 filters with a kernel size equal to 3; the activation function is a rectified linear unit (ReLU), followed by a flatten layer and then a dense layer with a sigmoid activation function. There are two distances used: Manhattan distance and exponential Manhattan distance accordingly. Next, a fusion layer has been added to fuse between two distance layers (Manhattan, exponential Manhattan). Then, one layer has been added after the fusion layer, which represented the output layer. The ReLU activation

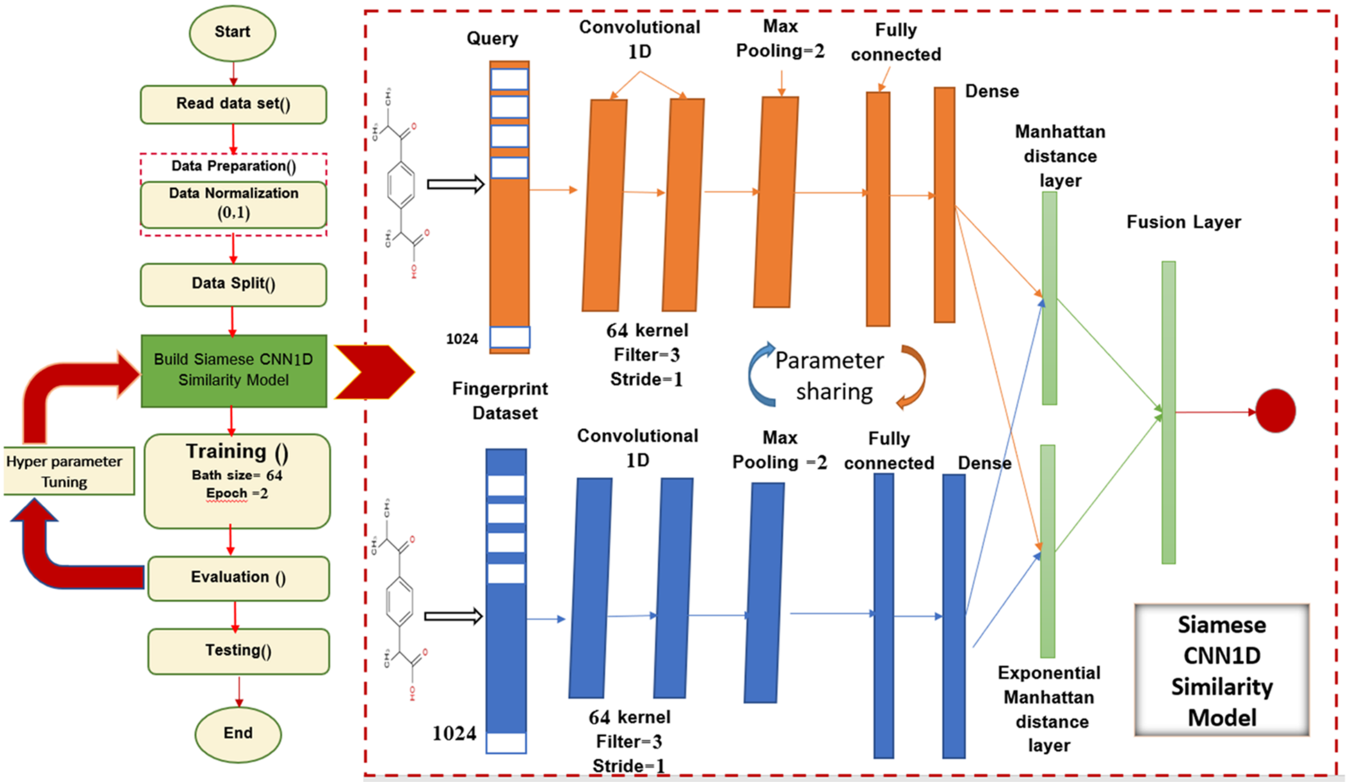


Figure 4. Architecture of the enhanced Siamese CNN1D similarity model.

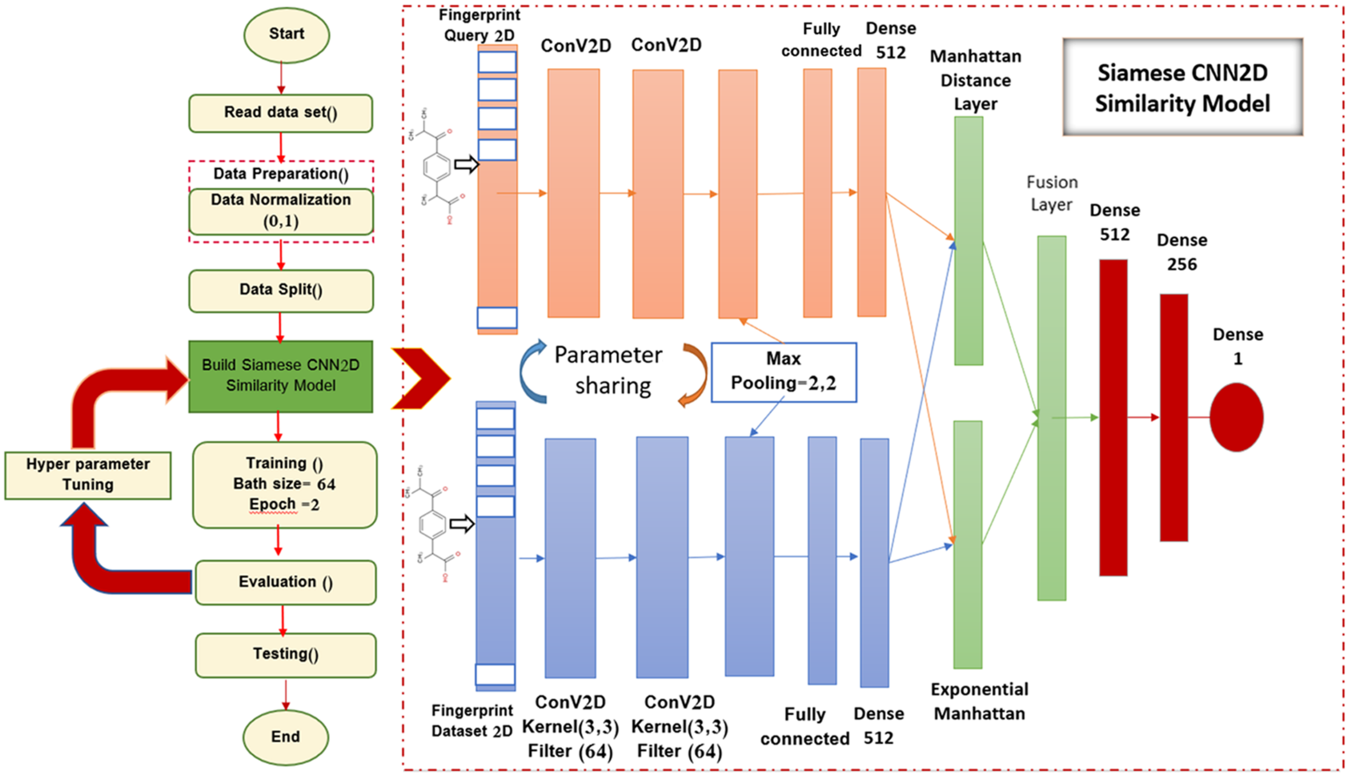


Figure 5. Architecture of the enhanced Siamese CNN2D similarity model.

function is used for all dense layers except the layer before the distance layers and the output layer in which the sigmoid activation function has been used. Moreover, the RMSprop optimizer has been used, binary_crossentropy is the loss function, and 64 is the batch size. Figure 4 demonstrates the architecture of the Siamese CNN1D similarity model.

3.2.2. Enhanced Siamese CNN2D Similarity Model. An enhanced Siamese CNN2D framework has been used to calculate the similarity between a reference structure and a database structure based on 2D fingerprints; therefore, the architecture has two inputs, representing the fingerprint of molecules, one from the reference structure (query) and the other from the database structure. Also, the architecture has one output, which represents the degree of similarity; this means that the output has two classes: if the value is 1, it means high similarity, and if the value is 0, it means high dissimilarity. Also, weights have been tied such that CNN2Da = CNN2Db in this architecture. As mentioned above, there are two inputs: each input layer of the convolution neural network (2D-CNN) received the molecular fingerprint. The layer is formed of 64 filters with a kernel size equal to (3,3); the activation function is a rectified linear unit (ReLU), followed by another layer of a 2D convolution neural network (2D-CNN) formed of 64 filters with a kernel size equal to (3,3), a max pooling size equal to (2,2), a flatten layer, and then a dense layer with a sigmoid activation function.

Two distances are used: the first one is the Manhattan distance, and the second distance is the exponential Manhattan distance. Next, a fusion layer is added to fuse between two distance layers (Manhattan, exponential Manhattan). Then, three layers are added after the fusion layer; the number of cells in these layers are 512, 256, and 1, respectively. The ReLU activation function is used for all dense layers except the layer before the distance layers and the output layer in which the sigmoid activation function has been used. Moreover, the RMSprop optimizer has been used, the loss function is binary_crossentropy, and the batch size is 64. Figure 5 demonstrates the architecture of the Siamese CNN2D similarity model.

4. EXPERIMENTAL DESIGN

4.1. Data Sets. Experiments were conducted using MDL Drug Data Report data sets (MDDR-DS1, MDDR-DS2, and MDDR-DS3)[53] and the Maximum Unbiased Validation (MUV) data set,[54] the most common cheminformatics database. In these databases, all molecules have been translated to the Pipeline Pilot, ECFC-4, and these databases have recently been used by our study community. With ten reference structures chosen randomly from each activity class, the screening experiments were carried out. MDDR-DS1 has 102 516 molecules (active and inactive). The active molecules (about 8300 molecules) comprise 11 activity groups, some with structurally homogeneous active elements and others with structurally heterogeneous (i.e., structurally diverse) active elements. Database MDDR-DS2 also has 102 516 molecules (active and inactive). The active molecules (about 5100 molecules) consist of 10 homogeneous activity classes. Database MDDR-DS3 has 102 516 molecules (active and inactive). The active molecules (about 8600 molecules) consist of 10 heterogeneous activity classes. Tables 1−3 provide descriptions of all three data sets. Each row of the table includes the activity class, the number of molecules belonging to the class, as well as a diversity of groups, which were measured as the average similarity of Tanimoto, computed by ECFC-4 for all pairs of molecules. Rohrer and Baumann recorded the second data collection (MUV), as seen in Table 4. There are 17 interaction groups in this data set, with each class containing up to 30 active and 15 000 inactive molecules. The class composition for this data set indicates that it involves classes with high diversity or more heterogeneous operations. In the previous articles, our research group has used these data collections.

Table 1. MDDR-DS1 Structure Activity Classes

activity index	active molecules	activity class	pairwise similarity
31420	1130	renin inhibitors	0.290
31432	943	angiotensin II AT1 antagonists	0.229
37110	803	thrombin inhibitors	0.180
71 523	750	HIV protease inhibitors	0.198
42731	1246	substance P antagonists	0.149
07701	395	D2 antagonists	0.138
06245	359	5HT reuptake inhibitors	0.122
78374	453	protein kinase C inhibitors	0.120
06235	827	5HT1A agonists	0.133
06233	752	5HT3 antagonist	0.140
78331	636	cyclooxygenase inhibitors	0.108

Table 2. MDDR-DS2 Structure Activity Classes

activity index	active molecules	activity class	pairwise similarity
07707	207	adenosine (A1) agonists	0.229
42710	111	CCK agonists	0.361
31420	1130	renin inhibitors	0.290
64200	113	cephalosporins	0.322
64100	1346	monocyclic-lactams	0.336
64500	126	carbapenems	0.260
64220	1051	carbacephems	0.269
75755	455	vitamin D analogues	0.386
75755	455	vitamin D analogues	0.386
07708	156	adenosine (A2) agonists	0.305

Table 3. MDDR-DS3 Structure Activity Classes

activity index	active molecules	activity class	pairwise similarity
09249	900	muscarinic (M1) agonists	0.111
31281	106	dopamine-hydroxylase inhibitors	0.125
12464	505	nitric oxide synthase inhibitors	0.102
71522	700	reverse transcriptase inhibitors	0.103
43210	957	aldose reductase inhibitors	0.119
12455	1400	NMDA receptor antagonists	0.098
75721	636	aromatase inhibitors	0.110
78351	2111	lipoxygenase inhibitors	0.113
78348	617	phospholipase A2 inhibitors	0.123
78331	636	cyclooxygenase inhibitors	0.108

4.2. Performance Evaluation Measures. The efficiency of the proposed methods is evaluated as follows:

1. The first way to evaluate the performance of the retrieval model is to use the Recall metric, which is the portion of active chemical compounds within the top 1 and 5% of the ranking test set that can be found. This measure has been used in previous studies.[28,31−33,55−66]

The whole data is divided into K sets of equal size: one of them as a test set, and the remaining sets as training sets. Selection of a test set will change in each iteration, and the mean of recall values from all iterations is calculated as the final result. This method is called k-fold cross validation, as shown in Figure 6. In each iteration, ten queries are tested, which are randomly selected from the activity class, and then the mean value of these ten queries is calculated.

Table 4. MUV Structure Activity Classes

activity index	activity class	pairwise similarity
466	S1P1 rec. (agonists)	0.117
644	Rho-Kinase2 (inhibitors)	0.122
600	SF1 (inhibitors)	0.123
689	Eph rec. A4 (inhibitors)	0.113
652	HIV RT-RNase (inhibitors)	0.099
712	HSP 90 (inhibitors) 30	0.106
692	SF1 (agonists)	0.114
733	ER-b-Coact. Bind. (inhibitors)	0.114
713	ER-a-Coact. Bind. (inhibitors)	0.113
810	FAK (inhibitors)	0.107
737	ER-a-Coact. Bind. (potentiators)	0.129
846	FXIa (inhibitors)	0.161
832	cathepsin G (inhibitors)	0.151
858	D1 rec. (allosteric modulators)	0.111
852	FXIIa (inhibitors)	0.150
548	PKA (inhibitors)	0.128
859	M1 rec. (allosteric inhibitors)	0.126

2. Comparison methods: The second way is current approaches that can be used in assessing the results of the proposed model. These approaches include the following.

A. TAN: Over the years, the Tanimoto similarity coefficient has been the search benchmark method in LBVS. The Tanimoto-based model for similarities employs the Tanimoto coefficient in its continuous form, which is suitable to nonbinary fingerprint data.[23]

B. The second method is Bayesian inference in the MDDR data set (DS1, DS2, DS3, and MUV) for the ECFC-4 descriptor. This is an alternative method for calculating the similarity of molecular fingerprints.[29,62]

C. The third method is quantum similarity search SQB-(Complex) in the MDDR data set (DS1, DS2, DS3, and MUV) for the ECFC-4 descriptor. This method utilizes a quantum mechanics approach.[31]

D. SDBN: The latest study is a multidescriptor-based on Stack of deep belief networks method in the MDDR data set (DS1, DS2, and DS3) for ECFC-4, ECFP-4, and EPFP-4 descriptors. The molecular features are re-weighted using deep belief networks.[33]

3. The third significant measure that can be used to evaluate the proposed methods, known as the significance test, is the Kendall W concordance test. This significance test has been used in previous studies.[28,33,55,61,62,64,65] This test can be interpreted as the concordance coefficient, which is a measure of agreement among the raters. Each case is a judge or rater in the Kendall W test, whereas each variable is an object or person being judged. For each variable, thus, the number of ranks is computed. The Kendall W test range is between 0, indicating no agreement, and 1, indicating full agreement. For example, the rank r_{ij} by judge number j, which represents an activity class, where there are n objects and m judges in total, is given to object I as the similarity search tool. It is then possible to calculate the total rank given to object I as[67]

$$\mathfrak{R}_i = \sum_{j=1}^{m} r_{ij} \tag{3}$$

whereas the complete ranks' mean meaning is

$$\bar{\mathfrak{R}} = \frac{1}{2}m(n + 1) \tag{4}$$

The squared deviation sum δ is defined as

$$\delta = \sum_{i=1}^{n} (\mathfrak{R}_i - \bar{\mathfrak{R}})^2 \tag{5}$$

Then, the Kendall W test is defined as

$$W = \frac{12\delta}{m^2(n^3 - n)} \tag{6}$$

The Kendall's W statistical values can be between zero and one since the variance of the number of ranks separated by the maximum possible value has been calculated, which happens when all judges are in absolute agreement. This test shows whether a group of judges can make equivalent decisions about the rating of a set of items or not. The definitions used in this analysis suggest that judges were considered to be the behavior groups of each of the data sets, whereas the recall rates of the different search models were considered to be the items. The outcomes of the Kendall coefficient that are related to significance levels are a significant part of this experiment. This implies verifying whether the value of the coefficient may have happened by chance or not. If the value was important (for which both 0.01 and 0.05 cutoff values were used), it was then possible to assign the item an overall ranking.

4. For a more evident comparison between the recall values of the proposed methods and previous studies, the

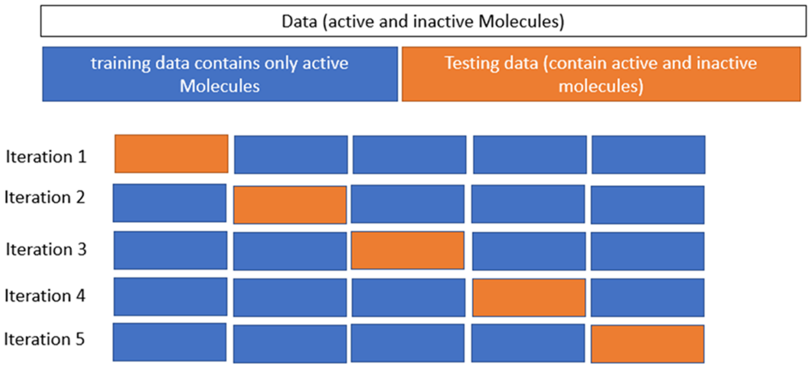

Data (active and inactive Molecules)
training data contains only active Molecules
Testing data (contain active and inactive molecules)
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5

Figure 6. Idea of cross validation for training and testing data.

Table 5. Top 1% Retrieval Results for MDDR-DS1 Data Set for Descriptor ECFC-4

DS1 Retrieval Result 1%	Previous Studies				proposed Methods			
Activity Index					RNN		CNN	
	TAN	BIN	SQB	SDBN	GRU	LSTM	CNN1D	CNN2D
31420	69.69	74.08	73.73	74.21	69.68	70.58	84.58	83.98
71523	25.94	28.26	26.84	27.97	37.29	31.20	59.41	53.59
37110	9.63	26.05	24.73	26.03	28.46	20.31	52.88	42.99
31432	35.82	39.23	36.66	39.79	52.06	48.16	66.41	59.48
42731	17.77	21.68	21.17	23.06	26.67	22.24	38.88	36.70
6233	13.87	14.06	12.49	19.29	16.19	14.72	35.03	27.01
6245	6.51	6.31	6.03	6.27	3.94	5.27	10.68	6.68
7701	8.63	11.45	11.35	14.05	13.11	12.05	16.96	11.92
6235	9.71	10.84	10.15	12.87	8.29	8.00	15.31	12.07
78374	13.69	14.25	13.08	17.47	18.07	11.07	24.60	19.73
78331	7.17	6.03	5.92	9.93	4.61	3.42	8.58	7.06
Mean	**19.86**	**22.93**	**22.01**	**24.63**	**25.31**	**22.46**	**37.57**	**32.84**
No. of Shaded cells	0	0	0	1	0	0	10	0

improvement percentage for each proposed method was calculated using eq 7.[68]

$$\text{improvement}_{\text{method1}} = \frac{\text{recall}_{\text{method1}} - \text{recall}_{\text{method2}}}{\text{recall}_{\text{method1}}} \times 100\% \quad (7)$$

For instance, the improvement percentage of GRU was calculated using the improvement equation with Tan, BIN, SQB, and SDBN. Next, the mean value was calculated; if the result value was positive, there was an improvement in retrieval recall of GRU compared with previous studies, and if the result value was negative, the retrieval recall of GRU was worse. Next, the mean value of improvement overall classes was calculated. Here, this will apply to all proposed methods. However, the improvement percentage for each previous method was also calculated compared with the proposed methods, for example, the improvement percentage of TAN, compared with GRU, LSTM, CNN1D, and CNN2D, then, the mean value was calculated for each class, and then the mean value of all classes in the data set was calculated.

5. RESULTS AND DISCUSSION

The ECFC-4 descriptor's experimental findings on the MDDR-DS1, MDDR-DS2, MDDR-DS3, and MUV data sets are provided in Tables 5−12, respectively, using 1 and 5% cutoffs. The results of the proposed methods of deep learning compared to the benchmark TAN and previous studies BIN, SQB, and SDBN are recorded in these tables. For the top 1% and 5%of the activity class, each row in the tables lists the recall values, and in each row, the best recall rate is shaded. In the tables, the mean rows relate to the average of all activity classes when combined, and the rows of shaded cells are the total number of shaded cells have the top values for each technique over the full range of classes of activity. The distribution of results in tables is provided in boxplots in Figures 7−14.

The MDDR-DS1 recall values for the 1 and 5% cutoffs recorded in Tables 5 and 6, respectively, showed that the proposed Siamese deep learning approaches were obviously superior to the benchmark TAN method and other studies. In addition, among other Siamese deep learning strategies, the CNN1D approach gives the best retrieval recall results in Table 5 in each of mean and the number of shaded cells, when compared, followed by the CNN2D method, GRU, SDNB, BIN, LSTM, SQB, and TAN. The boxplot in Figure 7 shows the comparison among methods for distribution of results in MDDR-DS1 at the top 1%, in view of maximum values, upper quartile values, mean values, median values, and lower quartile values. So, the top four methods in view of maximum values are CNNID, CNN2D, GRU, and LSTM; in upper quartile values are CNNID, CNN2D, GRU, and LSTM; in mean values are CNNID, CNN2D, GRU, and SDBN; in median values are CNNID, CNN2D, SDBN, and GRU; and in lower quartile values are CNNID, CNN2D, SDBN, and GRU. Also, by comparison, the CNN1D approach offered the best retrieval recall results in Table 6, in each of mean and the number of shaded cells, followed by the CNN2D method, GRU, LSTM, SDNB, BIN, SQB, and TAN. The boxplot in Figure 8 shows the comparison among methods for distribution of results in MDDR-DS1 at the top 5%, in view of maximum values, upper

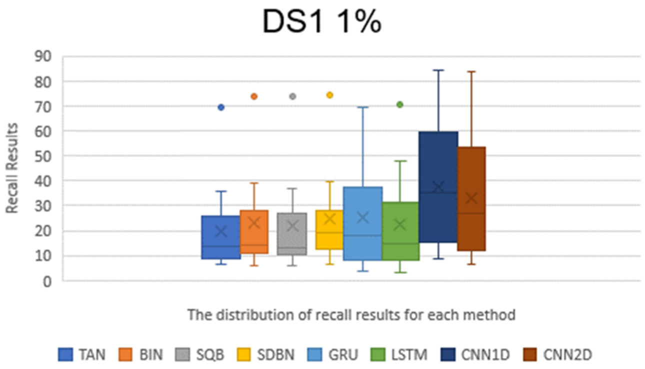


Figure 7. Boxplot for recall result distribution for each method in MDDR-DS1 at the top 1%.

Table 6. Top 5% Retrieval Results for MDDR-DS1 Data Set for Descriptor ECFC-4

DS1 Retrieval Result 5%	Previous Studies				Proposed Methods			
Activity Index					RNN		CNN	
	TAN	BIN	SQB	SDBN	GRU	LSTM	CNN1D	CNN2D
31420	83.49	87.61	87.22	89.03	84.46	83.41	87.35	87.69
71523	48.92	52.72	48.7	65.17	65.40	63.75	79.61	71.96
37110	21.01	48.2	45.62	41.25	50.60	42.80	76.00	67.51
31432	74.29	77.57	70.44	79.87	86.55	82.63	91.83	87.01
42731	29.68	26.63	24.35	31.92	47.76	43.50	57.52	57.43
6233	27.68	23.49	20.04	29.31	35.53	34.51	62.76	58.80
6245	16.54	14.86	13.72	21.06	12.62	15.58	28.90	19.27
7701	24.09	27.79	26.73	28.43	37.39	39.70	42.25	34.71
6235	20.06	23.78	22.81	27.82	25.12	25.93	40.36	31.24
78374	20.51	20.2	19.56	19.09	38.98	27.78	48.27	47.24
78331	16.2	11.8	11.37	16.21	15.84	13.91	25.02	21.81
Mean	**34.77**	**37.70**	**35.51**	**40.83**	**45.48**	**43.04**	**58.17**	**53.15**
No. of Shaded cells	**0**	**0**	**0**	**1**	**0**	**1**	**10**	**0**

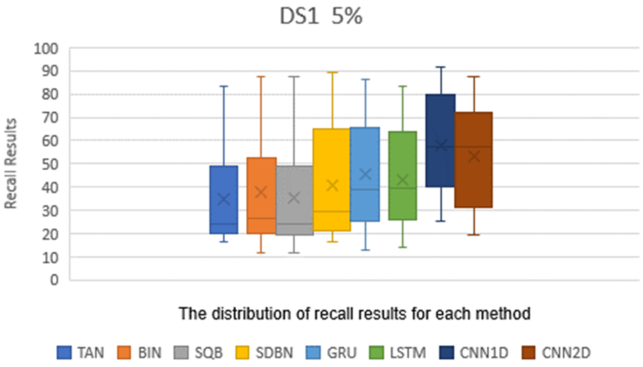


Figure 8. Boxplot for recall result distribution for each method in MDDR-DS1 at the top 5%.

quartile values, mean values, median values, and lower quartile values. So, the top four methods in view of maximum values are CNNID, SDBN, CNN2D, and BIN; in upper quartile values are CNNID, CNN2D, GRU, and SDBN; in mean values are CNNID, CNN2D, GRU, and LSTM; in median values are CNNID, CNN2D, LSTM, and GRU; and in lower quartile values are CNNID, CNN2D, LSTM, and GRU.

Furthermore, the MDDR-DS2 recall values recorded for the top 1% in Table 7 show that the proposed Siamese deep learning method (CNN1D) is clearly superior to the benchmark TAN method and previous studies. The CNN1D method gives the best retrieval recall results in each mean and the number of shaded cells. The second best are SDBN, BIN, and then SQB methods in view of the mean value, followed by Siamese CNN2D, LSTM, GRU, and finally Siamese TAN. The boxplot in Figure 9 shows the comparison among methods for distribution of results in MDDR-DS2 at the top 1%, in view of maximum values, upper quartile values, mean values, median values, and lower quartile values. So, the top four methods in view of maximum values are BIN, SQB, CNN1D, and SDBN; in upper quartile values are SDBN, CNN1D, CNN2D, and BIN; in mean values are CNNID, BIN, SQB, and CNN2D; in median values are CNNID, CNN2D, SDBN, and BIN; and in lower quartile values are CNNID, SDBN, BIN, and SDBN. However, by comparison, the MDDR-DS2 recall values recorded for 5% cutoffs in Table 8 show that the BIN method gave the best retrieval recall results in view of the mean and the number of shaded cells. The second best are SQB, SDBN, CNN1D, CNN2D, LSTM, and finally TAN in view of the mean value. The boxplot in Figure 10 shows the comparison among methods for distribution of results in MDDR-DS2 at the top 5%, in view of maximum values, upper quartile values, mean values, median values, and lower quartile values. So, the top four methods in view of maximum values are BIN, SQB, SDBN, and CNN1D; in upper quartile values are BIN, SQB, SDBN, and CNN1D; in mean values are BIN, SQB, SDBN, and CNN1D; in median values are BIN, SQB, SDBN, and CNN1D; and in lower quartile values are BIN, SQB, SDBN, and CNN1D.

In addition, the MDDR-DS3 recall values recorded for the top 1% and 5% in Tables 9 and 10, respectively, show that the proposed Siamese deep learning methods are clearly superior to the benchmark TAN method and other studies. Likewise, in Table 9, the CNN1D method gives the best retrieval recall results in view of mean and the number of shaded cells, compared to previous studies and other methods of Siamese deep learning. Next, the second one is Siamese CNN2D, followed by SDBN, GRU, BIN, SQB, TAN, and LSTM. The boxplot in Figure 11 shows the comparison among methods for distribution of results in MDDR-DS3 at the top 1%, in view of maximum values, upper quartile values, mean values, median values, and lower quartile values. So, the top four methods in view of maximum values are CNN1D, CNN2D, GRU, and SDBN; in upper quartile values are CNN1D, CNN2D, GRU, and SDBN; in mean values are CNN1D, CNN2D, SDBN, and GRU; in median values are CNN1D, CNN2D, SDBN, and GRU; and in lower quartile values are CNN1D, CNN2D, SDBN, and GRU. However, by comparison, in Table 10, the CNN1D method gives the best retrieval recall results in view of the mean and the number of shaded cells, compared to previous studies and other methods of Siamese deep learning, followed by

Table 7. Top 1% Retrieval Results for MDDR-DS2 Data Set for Descriptor ECFC-4

DS2 Retrieval Result 1%	Previous Studies				Proposed Methods			
Activity Index					RNN		CNN	
	TAN	BIN	SQB	SDBN	GRU	LSTM	CNN1D	CNN2D
7707	61.84	72.18	72.09	83.19	71.07	66.83	93.27	78.05
7708	47.03	96	95.68	94.82	81.42	88.26	94.84	92.13
31420	65.1	79.82	78.56	79.27	65.12	61.79	76.96	67.28
42710	81.27	76.27	76.82	74.81	61.64	58.09	84.55	81.55
64100	80.31	88.43	87.8	93.65	92.22	94.63	97.63	90.02
64200	53.84	70.18	70.18	71.16	63.55	69.10	78.65	67.10
64220	38.64	68.32	67.58	68.71	78.08	76.82	90.81	90.51
64500	30.56	81.2	79.2	75.62	55.92	58.32	71.92	62.16
64350	80.18	81.89	81.68	85.21	81.32	82.05	87.32	84.03
75755	87.56	98.06	98.02	96.52	87.54	86.84	90.95	90.44
Mean	**62.633**	**81.235**	**80.761**	**82.296**	**73.79**	**74.27**	**86.69**	**80.33**
Shaded cells	**0**	**4**	**0**	**0**	**0**	**0**	**6**	**0**

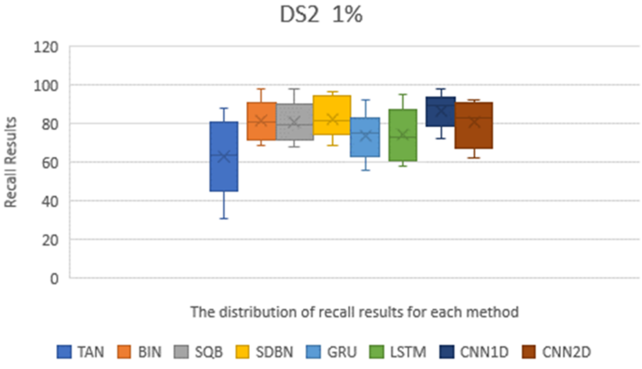


Figure 9. Boxplot for recall result distribution for each method in MDDR-DS2 at the top 1%.

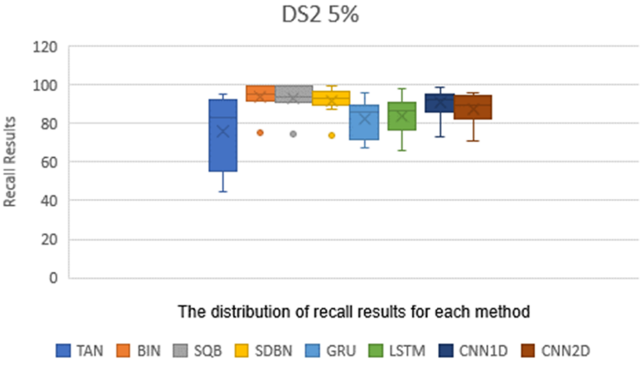


Figure 10. Boxplot for recall result distribution for each method in MDDR-DS2 at the top 5%.

Table 8. Top 5% Retrieval Results for MDDR-DS2 Data Set for Descriptor ECFC-4

DS2 Retrieval Result 5%	Previous Studies				Proposed Methods			
Activity Index					RNN		CNN	
	TAN	BIN	SQB	SDBN	GRU	LSTM	CNN1D	CNN2D
7707	70.39	74.81	74.37	73.9	84.10	85.27	95.85	88.49
7708	56.58	99.61	99.61	98.22	89.10	93.03	94.90	96.06
31420	88.19	95.46	94.88	95.64	81.12	79.27	94.12	87.73
42710	88.09	92.55	91.09	90.12	68.55	65.82	85.64	85.64
64100	93.75	99.22	99.03	99.05	95.92	97.69	98.93	94.86
64200	77.68	99.2	99.38	93.76	72.58	79.03	86.19	72.77
64220	52.19	91.32	90.62	96.01	86.92	87.80	94.07	93.90
64500	44.8	94.96	92.48	91.51	67.52	69.44	73.20	71.04
64350	91.71	91.47	90.78	86.94	87.58	88.86	90.70	89.58
75755	94.82	98.35	98.37	91.6	89.60	89.60	90.99	90.99
Mean	**75.82**	**93.695**	**93.061**	**91.675**	**82.30**	**83.58**	**90.46**	**87.11**
Shaded cells	**1**	**4**	**3**	**2**	**0**	**0**	**1**	**0**

Table 9. Top 1% Retrieval Results for MDDR-DS3 Data Set for Descriptor ECFC-4

Ds3 Retrieval Result 1%	Previous Studies				Proposed Methods			
Activity Index					RNN		CNN	
	TAN	BIN	SQB	SDBN	GRU	LSTM	CNN1D	CNN2D
9249	12.12	15.33	10.99	19.47	22.97	17.20	38.01	38.24
12455	6.57	9.37	7.03	13.29	4.96	3.19	14.21	7.82
12464	8.17	8.45	6.92	12.91	9.90	7.60	25.98	21.74
31281	16.95	18.29	18.67	23.62	43.62	24.57	67.52	67.52
43210	6.27	7.34	6.83	14.23	9.05	3.73	29.34	21.84
71522	3.75	4.08	6.57	11.92	2.84	2.06	12.00	7.26
75721	17.32	20.41	20.38	29.08	27.73	17.06	52.11	48.61
78331	6.31	7.51	6.16	11.93	6.80	5.23	12.41	11.01
78348	10.15	9.79	8.99	9.17	8.75	5.66	13.85	16.20
78351	9.84	13.68	12.5	18.13	4.02	2.66	10.71	8.92
Mean	**9.745**	**11.425**	**10.504**	**16.375**	**14.06**	**8.89**	**27.62**	**24.92**
Shaded cells	**0**	**0**	**0**	**1**	**0**	**0**	**7**	**2**

Table 10. Top 5% Retrieval Results for MDDR-DS3 Data Set for Descriptor ECFC-4

Ds3 Retrieval Result 5%	Previous Studies				Proposed Methods			
Activity Index					RNN		CNN	
	TAN	BIN	SQB	SDBN	GRU	LSTM	CNN1D	CNN2D
9249	24.17	25.72	17.8	31.61	44.11	38.23	61.84	62.00
12455	10.29	14.65	11.42	16.29	13.26	11.11	32.97	20.14
12464	15.22	16.55	16.79	20.9	27.66	21.56	46.12	40.06
31281	29.62	28.29	29.05	36.13	65.62	44.67	78.57	79.90
43210	16.07	14.41	14.12	22.09	23.29	14.24	54.47	43.81
71522	12.37	8.44	13.82	14.68	9.54	7.59	29.19	18.86
75721	25.21	30.02	30.61	41.07	53.46	38.74	77.31	71.50
78331	15.01	12.03	11.97	17.13	19.91	15.97	31.29	28.99
78348	24.67	20.76	21.14	26.93	18.50	13.38	31.89	33.98
78351	11.71	12.94	13.3	17.87	13.28	10.28	30.16	25.07
Mean	**18.434**	**18.381**	**18.002**	**24.47**	**28.86**	**21.58**	**47.38**	**42.43**
Shaded cells	**0**	**0**	**0**	**0**	**0**	**0**	**7**	**3**

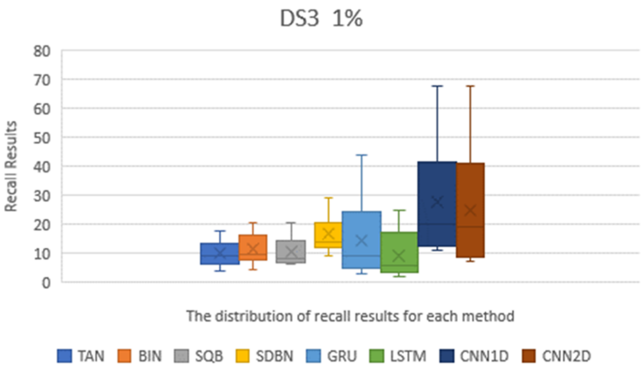


Figure 11. Boxplot for recall result distribution for each method in MDDR-DS3 at the top 1%.

Siamese CNN2D, GRU, SDBN, TAN, BIN, SQB, and finally LSTM. The boxplot in Figure 12 shows the comparison among methods for distribution of results in MDDR-DS3 at the top 5%, in view of maximum values, upper quartile values, mean values, median values, and lower quartile values. So, the top four methods in view of maximum values are CNN1D, CNN2D, GRU, and LSTM; in upper quartile values are CNN1D, CNN2D, GRU, and LSTM; in mean values are CNN1D, CNN2D, GRU, and SDBN; in median values are CNN1D, CNN2D, GRU, and SDBN; and in lower quartile values are CNN1D, CNN2D, SDBN, and GRU.

Moreover, the MUV data set recall values recorded for 1 and 5% cutoffs in Tables 11 and 12, respectively, show that the proposed Siamese deep learning CNN methods are clearly superior to the benchmark TAN method and previous studies. Likewise, in Table 11, the CNN1D Method gives the best

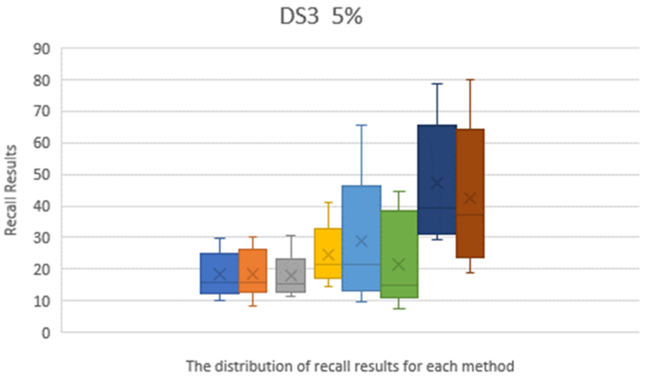


Figure 12. Boxplot for recall result distribution for each method in MDDR-DS3 at the top 5%.

retrieval recall results in view of the mean. Next, the second best are BIN and Siamese CNN2D, followed by GRU, LSTM, SQB, and finally TAN. The boxplot in Figure 13 shows the comparison among methods for distribution of results in MUV at the top 1%, in view of maximum values, upper quartile values, mean values, median values, and lower quartile values. So, the top four methods in view of maximum values are CNN1D, CNN2D, BIN, and SQB; in upper quartile values are CNN1D, CNN2D, BIN, and SQB; in mean values are CNN1D, BIN, CNN2D, and GRU; in median values are BIN, CNN2D, CNN1D, and SQB; and in lower quartile values are BIN, CNN2D, CNN1D, and GRU. By comparison, in Table 12, the CNN1D method gives the best retrieval recall results in view of the mean and the number of shaded cells, followed by CNN2D, BIN, SQB, TAN, GRU, and LSTM. The boxplot in Figure 14 shows the comparison among methods for distribution of results in MUV at the top 5%, in view of maximum values, upper quartile values, mean values, median values, and lower quartile values. So, the top four methods in view of maximum values are CNN1D, CNN2D, BIN, and SQB; in upper quartile values are CNN1D, CNN2D, BIN, and SQB; in mean values are CNN1D, CNN2D, BIN, and SQB; in median values are BIN, CNN2D, SQB, and CNN1D; and in lower quartile values are BIN, CNN2D, GRU, and SQB.

Moreover, the Kendall W concordance test has been used. Table 13 shows the ranking of enhanced Siamese deep learning (RNN-GRU, RNN-LSTM, CNN1D, CNN2D) methods based on previous studies TAN, BIN, SQB, and SDBN using Kendall W test results for MDDR-DS1, MDDR-DS2, MDDR-DS3, and MUV at the top 1% and top 5%. The first method is Tanimoto coefficient TAN, the second method is Bayesian inference (ABDO),[29] the third method is quantum similarity search SQB-Complex (Al-dabagh),[31] and the last method is multidescriptor-based on Stack of deep belief networks (Nasser).[33] For all of the data sets used, the Kendall W test of the top 1% shows that the significance test (P) values are less than 0.05. This means that the enhanced Siamese deep learning methods are significant in all cases with a cutoff of 1%. Therefore, the general ranking of all methods of deep learning indicates that the enhanced Siamese CNN methods are superior to previous studies and benchmark TAN; the overall ranking for methods shows that CNN1D has

Table 11. Top 1% Retrieval Results for MUV Data Set for Descriptor ECFC-4

MUV 1%	Previous Studies			Proposed Methods			
Activity Index				RNN		CNN	
	TAN	BIN	SQB	GRU	LSTM	CNN1D	CNN2D
466	3.1	6.33	1.38	6.67	4.33	6.00	4.00
548	8.62	14.89	11.38	6.00	5.67	13.33	11.33
600	3.79	6.33	5.52	5.67	4.33	5.33	5.67
644	7.59	11	8.97	8.00	5.33	15.33	16.33
652	2.76	7	3.79	4.33	4.33	5.33	6.67
689	3.79	7.33	4.48	4.33	5.67	3.67	5.33
692	0.69	5.33	1.38	5.00	3.67	3.00	4.33
712	4.14	8.22	5.17	4.67	3.67	10.67	6.33
713	3.1	5.89	2.76	5.00	3.67	4.67	5.67
733	3.45	6.67	4.14	3.33	3.67	3.67	4.00
737	2.41	5.11	1.72	3.33	3.67	6.33	6.33
810	2.07	6.78	1.72	3.33	4.33	4.67	3.67
832	6.55	12.55	8.28	8.67	6.67	21.33	12.00
846	9.66	13.11	12.41	13.67	14.33	26.33	20.00
852	12.41	13.78	9.66	8.33	7.00	33.00	19.33
858	1.72	5.11	1.38	4.00	4.00	3.00	3.67
859	1.38	4.89	2.41	3.67	3.67	4.33	4.67
Mean	**4.542**	**8.2541**	**5.091**	**5.76**	**5.18**	**10.00**	**8.20**
Shaded cells	**0**	**10**	**0**	**1**	**0**	**5**	**2**

Table 12. Top 5% Retrieval Results for MUV Data Set for Descriptor ECFC-4

MUV 5%	Previous Studies			Proposed Methods			
Activity Index				RNN		CNN	
	TAN	BIN	SQB	GRU	LSTM	CNN1D	CNN2D
466	5.86	10.44	8.62	10.00	7.33	11.00	8.33
548	22.76	27.22	24.14	14.33	16.33	32.00	31.33
600	11.38	12.89	16.21	10.67	10.33	9.67	12.67
644	17.59	19.67	17.93	18.33	15.00	36.67	31.00
652	7.93	11.67	9.66	8.00	9.00	9.33	12.67
689	9.66	13.22	11.72	14.67	11.67	14.00	9.67
692	4.83	9.22	4.83	9.67	8.33	6.00	8.67
712	10.34	16.45	11.03	9.00	7.33	16.67	13.00
713	7.24	9	5.86	10.33	7.67	7.33	8.67
733	8.97	10.11	8.62	5.33	6.33	6.33	6.67
737	8.28	12	8.28	7.00	7.33	8.33	12.00
810	6.9	13.33	11.03	5.33	7.00	6.67	8.00
832	13.1	20.44	14.83	15.33	12.00	32.00	23.33
846	28.62	26.11	26.9	23.00	28.00	47.00	37.67
852	21.38	23.11	20	16.33	13.33	42.33	33.67
858	5.86	9.11	6.21	9.33	6.33	5.00	9.00
859	8.97	9.44	8.62	9.33	6.33	11.67	10.33
Mean	**11.75**	**14.91**	**12.62**	**11.53**	**10.57**	**17.76**	**16.27**
Shaded cells	**0**	**3**	**1**	**4**	**0**	**8**	**2**

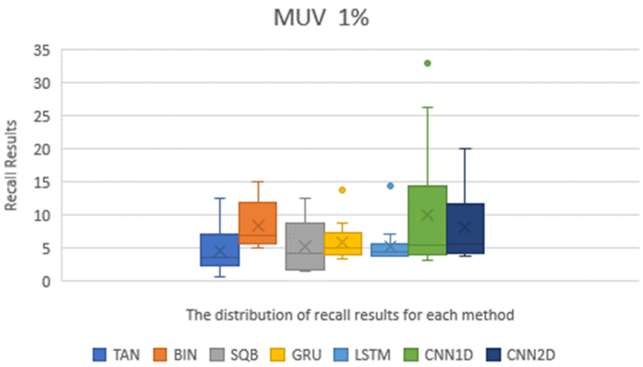


Figure 13. Boxplot for recall result distribution for each method in MUV at the top 1%.

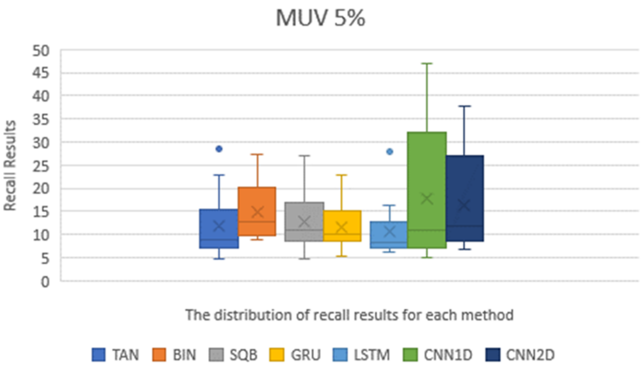


Figure 14. Boxplot for recall result distribution for each method in MUV at the top 5%.

the top rank among other methods in DS1, DS2, DS3 data sets, while BIN method has top rank in the MUV data set.

Same as with the results of the Kendall W test of the top 5%. The results indicate that the probability values (p) related are below 0.05. This denotes that deep learning methods for enhanced Siamese are important in all cases at a cutoff of 5%. As a result, the overall ranking of all methods of deep learning indicates that enhanced Siamese CNN1D is superior to previous studies for DS1 and DS3. In DS2 and MUV, BIN has the top rank at the top 5%. Figures 15 and 16 show the ranking of enhanced Siamese deep learning (RNN-GRU, RNN-LSTM, CNN1D, CNN2D) methods based on TAN, BIN, SQB, and SDBN using Kendall W test results for DS1, DS2, DS3, and MUV at the top 1% and 5%.

For another comparison between the recall values of the proposed methods and prior studies, the improvement percentage is calculated for proposed methods and prior methods for each data set, as shown in Table 14. In the DS1 data set, the proposed CNN methods have positive values at the top 1%, meaning that there is improvement in retrieval recall compared with prior methods; besides that, CNN1D has the top value of improvement percentage, followed by CNN2D, while all previous methods have negative values, meaning that the retrieval recall is worse compared with the proposed methods. For the top 5%, all proposed methods have positive values, meaning that there is improvement in retrieval recall compared with prior methods, and CNN1D has the top value of improvement, followed by CNN2D, GRU, and LSTM, while

Table 13. Ranking of Enhanced Siamese Deep Learning (RNN-GRU, RNN-LSTM, CNN1D, CNN2D) Methods based on TAN, BIN, SQB, and SDBN Using Kendall W Test Results for DS1, DS2, DS3, and MUV at the Top 1% and 5%

data set	retrieval percentage (%)	*W*	*P*	rank methods							
DS1	1	0.64	2.24×10^{-8}	1-CNN1D 7.91	2-CNN2D 6.45	3-SDBN 5.36	4-GRU 4.27	5-BIN 4.00	6-LSTM 3.00	7-SQB 2.55	8-TAN 2.45
	5	0.66	1.1601×10^{-8}	1-CNN1D 7.73	2-CNN2D 6.73	3-GRU 4.91	4-SDBN 4.64	5-LSTM 4.27	6-BIN 4.00	7-TAN 2.64	8-SQB 1.91
DS2	1	0.49	1.471×10^{-5}	1-CNN1D 6.9	2-SDBN 5.8	3-BIN 5.65	4-CNN2D 4.9	5-SQB 4.85	6-LSTM 3.2	7-GRU 2.9	8-TAN 1.8
	5	0.47	2.8157×10^{-5}	1-BIN 6.85	2 SQB 6.25	3-SDBN 5.5	4-CNN1D 5.1	5-CNN2D 4	6-TAN 6	7-LSTM 2.95	8-GRU 2.25
DS3	1	0.64	1.4015×10^{-7}	1-CNN1D 7.45	2-CNN2D 6.45	3 SDBN 6	4-BIN 4.4	5-GRU 3.9	6-SQB 3.1	7-TAN 2.9	8-LSTM 1.8
	5	0.74	7.00×10^{-9}	1-CNN1D 7.7	2-CNN2D 7.3	3-SDBN 5.1	4-GRU 4.9	5-LSTM 3	6-SQB 2.8	7-TAN 2.6	8-BIN 2.6
MUV	1	0.52	9.62×10^{-10}	1-BIN 6.23	2-CNN2D 5.235	3-CNN1D 5	4-GRU 3.76	5-LSTM 3.24	6-SQB 2.71	7-TAN 1.82	
	5	0.33	9.5856×10^{-6}	1-BIN 5.56	2-CNN2D 5.21	3-CNN1D 4.91	4-SQB 3.65	5-GRU 3.47	6-TAN 2.76	7-LSTM 2.44	

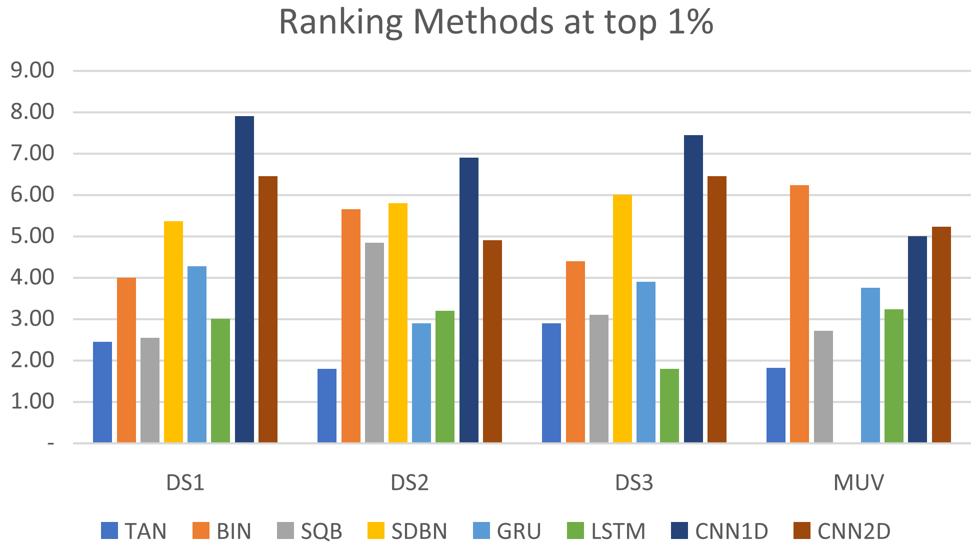


Figure 15. Ranking of enhanced Siamese deep learning (RNN-GRU, RNN-LSTM, CNN1D, CNN2D) methods based on TAN, BIN, SQB, and SDBN using Kendall W test results for DS1, DS2, DS3, and MUV at the top 1%.

all prior methods have negative values at the top 5%, meaning that the retrieval recall is worse compared with the proposed methods.

Also, in the DS3 data set, the proposed methods have positive values at the top 1%, except GRU and LSTM, meaning that there is an improvement in retrieval recall compared with prior methods, and CNN1D has the top value of improvement, followed by CNN2D. The same as with the top 5%, all proposed methods have positive values, except LSTM, meaning that there is improvement in retrieval recall compared with prior methods, and CNN1D has the top value of improvement, followed by CNN2D and GRU, while all prior methods have negative values at the top 1% and 5%, meaning that the retrieval recall is worse compared with the proposed methods.

Moreover, in the MUV data set at the top 1%, the proposed CNN methods have positive values, which means there is improvement in retrieval recall compared with prior methods; also, the previous study on the BIN method has positive values. CNN1D has the top value of improvement, followed by CNN2D and BIN methods. The same as with the top 5%, the proposed methods, except RGU and LSTM, have positive values, meaning that there is improvement in retrieval recall compared with prior methods; also, the previous study on the BIN method has positive values. CNN2D has the top value of

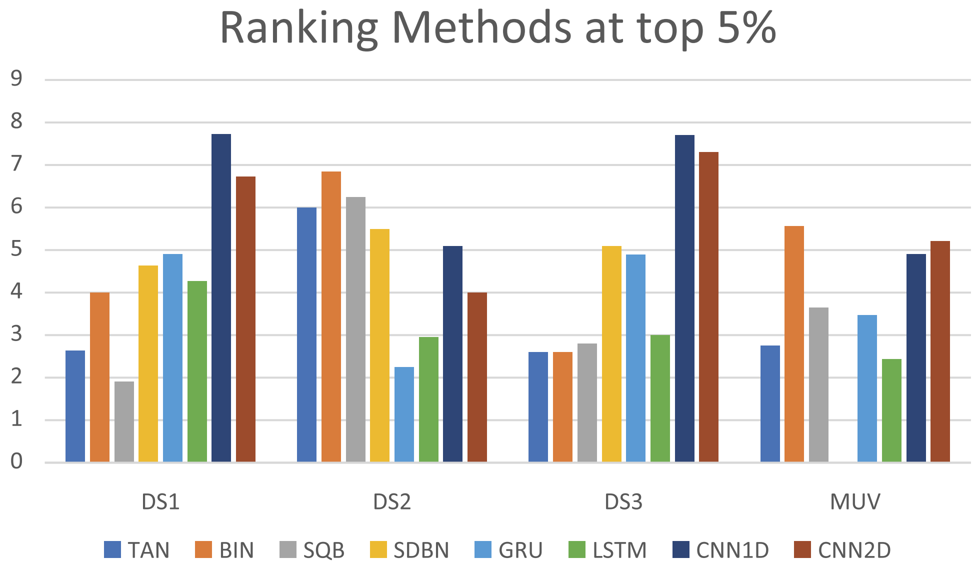


Figure 16. Ranking of enhanced Siamese deep learning (RNN-GRU, RNN-LSTM, CNN1D, CNN2D) methods based on TAN, BIN, SQB, and SDBN using Kendall W test results for DS1, DS2, DS3, and MUV at the top 5%.

Table 14. Improvement Percentage of the Proposed Methods and Prior Methods for Each Data Set

		previous studies				proposed methods			
		TAN	BIN	SQB	SDBN	GRU	LSTM	CNN1D	CNN2D
DS1	top 1%	***−59.037***	***−27.955***	***−34.942***	***−14.029***	**−1.155**	***−15.051***	39.320	25.401
	top 5%	***−57.819***	***−47.508***	***−58.173***	***−31.669***	16.858	14.108	39.509	31.406
DS2	top 1%	***−37.911***	2.400	1.852	3.946	**−5.190**	**−4.989**	11.243	3.606
	top 5%	***−20.062***	7.746	7.137	5.812	**−8.796**	**−7.275**	1.597	**−2.530**
DS3	top 1%	***−79.723***	***−56.487***	***−70.460***	***−6.758***	***−35.122***	***−107.770***	44.872	31.746
	top 5%	***−86.018***	***−87.382***	***−91.183***	***−38.262***	17.277	**−9.266**	56.480	49.208
MUV	top 1%	***−93.350***	16.035	***−77.736***		**−3.198**	***−14.241***	24.255	22.283
	top 5%	***−20.652***	10.123	***−11.820***		***−17.121***	***−25.851***	5.637	10.929

improvement, followed by BIN and CNN1D methods, while GRU and LSTM have negative values, meaning that the retrieval recall is worse compared with previous methods. Also, the TAN, SQB, and SDBN have negative values, meaning that the retrieval recall is worse compared with the proposed methods.

However, in the DS2 data set, the proposed CNN methods have positive values at the top 1%, meaning that there is improvement in retrieval recall compared with prior methods; also, the previous studies have positive values, meaning that there is improvement in retrieval recall compared with the proposed methods, but the proposed CNN1D method has a top value of improvement, followed by SDBN, CNN2D, BIN, and SQB. In the top 5%, only CNN1D has a positive value. On the other side, the previous studies have positive values for BIN, SQB, and SDBN methods and BIN has the top value of improvement, followed by SQB, SDBN, and the proposed CNN1D method.

6. CONCLUSIONS

Many techniques for capturing the biological similarity between a test compound and a known target ligand in LBVS have been established. LBVS is based on the premise that the target-binding behavior of related property compounds will be related. In spite of the good performances of the above methods compared to their prior, especially when dealing with molecules that have homogeneous active structural elements, however, the performances are not satisfied when dealing with molecules that are structurally heterogeneous.

The main goal of this research is to improve the retrieval effectiveness of the similarity model, especially with molecules that have structurally heterogeneous, and because of their powerful generalization, feature extraction capabilities, and the power of deep learning for dealing with big data, also the power of Siamese architecture with dealing with complicated data samples, especially with heterogeneous data samples. Therefore, they have been used in this study. The Siamese deep learning models have been enhanced using two distance layers and then a fusion layer that combines the results from two distance layers and then adding multiple layers after the fusion layer for some models to improve the similarity recall between a test compound and a known target ligand. In this architecture, several deep learning methods have been used, which are LSTM, GRU, CNN1D, and CNN2D. The results showed that the significance of the proposed methods, especially Siamese CNN similarity models, obviously outperformed the standard Tanimoto coefficient (TAN) and previous studies (BIN, SQB, SDNB) at both top 1% and 5%, especially when the model deals with MDDR-DS1, MDDR-DS3, and MUV data sets that include heterogeneous classes.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.1c04587.

- data set description and experimental results and figures (PDF)

  (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Mohammed Khaldoon Altalib − *School of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia; Computer Science Department, College of Education for Pure Sciences, University of Mosul, 41002 Mosul, Iraq;* orcid.org/0000-0001-7748-1840; Email: khaldoon@graduate.utm.my

Author

Naomie Salim − *School of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia;* orcid.org/0000-0001-8509-3055

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c04587

Notes

The authors declare no competing financial interest.
●The MDL Drug Data Report (MDDR) data set is owned by www.accelrys.com. A license is required to access the data. ●Maximum Unbiased Validation (MUV) data sets are freely available at http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html.●Software License: Python 3.7 in environment Anaconda/Spyder was used with the following libraries: TensorFlow, Theano, Keras, NumPy, pandas, and math. The license of statistics application (IBM spss) is licenseapp.utm.my

■ ACKNOWLEDGMENTS

This work is supported by the Ministry of Higher Education, Malaysia (Project No. R.J130000.7828.4F985) and the Ministry of Higher Education and Universiti Teknologi Malaysia (Project No. Q.J130000.2551.21H38). Mohammed Khaldoon Altalib would like to thank the Islamic Development Bank (IsDB) for my Ph.D. scholarship and the Mosul University for allowing me and supporting me to continue my study for Ph.D.

■ REFERENCES

(1) Hertzberg, R. P.; Pope, A. J. High-throughput screening: new technology for the 21st century. *Curr. Opin. Chem. Biol.* **2000**, *4*, 445−451.

(2) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Economics* **2016**, *47*, 20−33.

(3) Carpenter, K. A.; Cohen, D. S.; Jarrell, J. T.; Huang, X. Deep learning and virtual drug screening. *Future Med. Chem.* **2018**, *10*, 2557−2567.

(4) Lavecchia, A.; Di Giovanni, C. Virtual screening strategies in drug discovery: a critical review. *Curr. Med. Chem.* **2013**, *20*, 2839−2860.

(5) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862−865.

(6) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening—an overview. *Drug Discovery Today* **1998**, 3, 160−178.

(7) Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S. H. Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J.* **2012**, *14*, 133−141.

(8) Chaudhary, K. K.; Mishra, N. A review on molecular docking: novel tool for drug discovery. *Databases* **2016**, *3*, No. 1029.

(9) Brown, N. Chemoinformatics—an introduction for computer scientists. *ACM Comput. Surv.* **2009**, *41*, 1−38.

(10) Willett, P. Similarity Searching Using 2D Structural Fingerprints. In *Chemoinformatics and Computational Chemical Biology*; Springer, 2010; pp 133−158.

(11) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58−63.

(12) Muegge, I.; Mukherjee, P. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin. Drug Discovery* **2016**, *11*, 137−148.

(13) Narang, S.; Elsen, E.; Diamos, G.; Sengupta, S., Exploring Sparsity In Recurrent Neural Networks, 2017, arXiv:1704.05119. arXiv.org e-Print archive. https://arxiv.org/abs.1802.00730.

(14) Fukunishi, Y. Structure-based drug screening and ligand-based drug screening with machine learning. *Comb. Chem. High Throughput Screening* **2009**, *12*, 397−408.

(15) Melville, J. L.; Burke, E. K.; Hirst, J. D. Machine learning in virtual screening. *Comb. Chem. High Throughput Screening* **2009**, *12*, 332−343.

(16) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X.; et al. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* **2015**, *113*, 184−215.

(17) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513−530.

(18) Gao, K.; Nguyen, D. D.; Sresht, V.; Mathiowetz, A. M.; Tu, M.; Wei, G.-W. Are 2D fingerprints still valuable for drug discovery? *Phys. Chem. Chem. Phys.* **2020**, *22*, 8373−8390.

(19) Jing, Y.; Bian, Y.; Hu, Z.; Wang, L.; Xie, X.-Q. S. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. *AAPS J.* **2018**, *20*, 58.

(20) Rifaioglu, A. S.; Atas, H.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings Bioinf.* **2019**, *20*, 1878−1912.

(21) Chicco, D. Siamese Neural Networks: An Overview. *Artificial Neural Networks*, 2021; pp 73−94.

(22) Jeon, M.; Park, D.; Lee, J.; Jeon, H.; Ko, M.; Kim, S.; Choi, Y.; Tan, A.-C.; Kang, J. ReSimNet: drug response similarity prediction using Siamese neural networks. *Bioinformatics* **2019**, *35*, 5249−5256.

(23) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **2015**, *7*, 20.

(24) Cai, C.; Gong, J.; Liu, X.; Gao, D.; Li, H. Molecular similarity: methods and performance. *Chin. J. Chem.* **2013**, *31*, 1123−1132.

(25) Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J. Mol. Graphics Modell.* **2010**, *29*, 157−170.

(26) Syuib, M.; Arif, S. M.; Malim, N. *Comparison of Similarity Coefficients for Chemical Database Retrieval*, 2013 1st International Conference on Artificial Intelligence, Modelling and Simulation; IEEE, 2013; pp 129−133.

(27) Willett, P. Textual and chemical information processing: Different domains but similar algorithms *Information Res.* 2000, 52.

(28) Ahmed, A.; Abdo, A.; Salim, N. Ligand-based virtual screening using Bayesian inference network and reweighted fragments *Sci. World J.* 2012, *2012*, DOI: 10.1100/2012/410914.

(29) ABDO, A. *Similarity-Based Virtual Screening Using Bayesian Inference Network*; Universiti Teknologi Malaysia: Malaysia, 2010.

(30) de Castro, P. A.; de França, F. O.; Ferreira, H. M.; Coelho, G. P.; Von Zuben, F. J. Query expansion using an immune-inspired biclustering algorithm. *Nat. Comput.* **2010**, *9*, 579−602.

(31) Al-Dabagh, M. M. *Quantium Inspired Probability Approaches in Ligend-Based Vitual Screen*; Universiti Teknologi Malaysia: Malaysia, 2017.

(32) Himmat, M. H. I. *New Similarity Measures for Ligand-Based Virtual Screening*; Universiti Teknologi Malaysia: Malaysia, 2017.

(33) Nasser, M.; Salim, N.; Hamza, H.; Saeed, F.; Rabiu, I. Improved Deep Learning Based Method for Molecular Similarity Searching Using Stack of Deep Belief Networks. *Molecules* **2021**, *26*, No. 128.

(34) Jing, Y.; Bian, Y.; Hu, Z.; Wang, L.; Xie, X.-Q. S. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. *The AAPS J.* **2018**, *20*, 1−10.

(35) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.

(36) Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G. K deep: Protein−ligand absolute binding affinity prediction via 3d-convolutional neural networks. *J. Chem. Inf. Model.* **2018**, *58*, 287−296.

(37) Skalic, M.; Jiménez, J.; Sabbadin, D.; De Fabritiis, G. Shape-based generative modeling for de novo drug design. *J. Chem. Inf. Model.* **2019**, *59*, 1205−1214.

(38) Gao, K.; Nguyen, D. D.; Tu, M.; Wei, G.-W. Generative network complex for the automated generation of drug-like molecules. *J. Chem. Inf. Model.* **2020**, *60*, 5682−5698.

(39) Hamza, H.; Nasser, M.; Salim, N.; Saeed, F. In *Bioactivity Prediction Using Convolutional Neural Network*, International Conference of Reliable Information and Communication Technology, Springer, 2019; pp 341−351.

(40) Mendolia, I.; Contino, S.; Perricone, U.; Pirrone, R.; Ardizzone, E. In *A Convolutional Neural Network for Virtual Screening of Molecular Fingerprints*, International Conference on Image Analysis and Processing, Springer, 2019; pp 399−409.

(41) Wan, F.; Zeng, J. Deep learning with feature embedding for compound-protein interaction prediction. *bioRxiv*, November 07, 2016, 086033. DOI: 10.1101/086033.

(42) Neil, D.; Segler, M.; Guasch, L.; Ahmed, M.; Plumbley, D.; Sellwood, M.; Brown, N. *Exploring Deep Recurrent Models with Reinforcement Learning for Molecule Design*, 2018.

(43) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep learning for computational chemistry. *J. Comput. Chem.* **2017**, *38*, 1291−1307.

(44) Yu, J.; Xie, G.; Li, M.; Hao, X. In *Retrieval of Family Members Using Siamese Neural Network*, 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, 2020; pp 882−886.

(45) Mueller, J.; Thyagarajan, A. In *Siamese Recurrent Architectures for Learning Sentence Similarity*, Proceedings of the AAAI Conference on Artificial Intelligence, 2016.

(46) Dhami, D. S.; Yan, S.; Kunapuli, G.; Page, D.; Natarajan, S., Beyond Textual Data: Predicting Drug-Drug Interactions from Molecular Structure Images using Siamese Neural Networks, 2019, arXiv:1911.06356. arXiv.org e-Print archive https://arxiv.org/abs/1802.00730.

(47) Neil, D.; Segler, M.; Guasch, L.; Ahmed, M.; Plumbley, D.; Sellwood, M.; Brown, N. In Exploring Deep Recurrent Models with Reinforcement Learning for Molecule Design, https://openreview.net/forum, 2018.

(48) Neil, D. *Deep Neural Networks and Hardware Systems for Event-driven Data*; ETH Zurich, 2017.

(49) Craw, S. Manhattan Distance. In *Encyclopedia of Machine Learning and Data Mining*, Sammut, C.; Webb, G. I., Eds.; Springer US: Boston, MA, 2017; pp 790−791.

(50) Jozefowicz, R.; Zaremba, W.; Sutskever, I. In *An Empirical Exploration of Recurrent Network Architectures*, International Conference on Machine Learning; PMLR, 2015; pp 2342−2350.

(51) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84−90.

(52) Sainath, T. N.; Mohamed, A.-r.; Kingsbury, B.; Ramabhadran, B. In *Deep Convolutional Neural Networks for LVCSR*, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013; pp 8614−8618.

(53) Inc A, *MDL Drug Data Report (MDDR)*; Inc, A.: San Diego, CA, USA.

(54) Rohrer, S. G.; Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.* **2009**, *49*, 169−184.

(55) Ahmed, A.; Salim, N.; Abdo, A. Fragment reweighting in ligand-based virtual screening. *Adv. Sci. Lett.* **2013**, *19*, 2782−2786.

(56) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information. *J. Med. Chem.* **2005**, *48*, 7049−7054.

(57) Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold hopping using clique detection applied to reduced graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503−511.

(58) Wilton, D. J.; Harrison, R. F.; Willett, P.; Delaney, J.; Lawson, K.; Mullier, G. Virtual screening using binary kernel discrimination: analysis of pesticide data. *J. Chem. Inf. Model.* **2006**, *46*, 471−477.

(59) Kogej, T.; Engkvist, O.; Blomberg, N.; Muresan, S. Multi-fingerprint based similarity searches for targeted class compound selection. *J. Chem. Inf. Model.* **2006**, *46*, 1201−1213.

(60) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Model.* **2004**, *44*, 1177−1185.

(61) Abdo, A.; Saeed, F.; Hamza, H.; Ahmed, A.; Salim, N. Ligand expansion in ligand-based virtual screening using relevance feedback. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 279−287.

(62) Abdo, A.; Salim, N.; Ahmed, A. Implementing relevance feedback in ligand-based virtual screening using Bayesian inference network. *J. Biomol. Screening* **2011**, *16*, 1081−1088.

(63) Nasser, M.; Salim, N.; Hamza, H. In *Molecular Similarity Searching Based on Deep Belief Networks with Different Molecular Descriptors*, Proceedings of the 2020 2nd International Conference on Big Data Engineering and Technology, 2020; pp 18−24.

(64) Al-Dabbagh, M. M.; Salim, N.; Himmat, M.; Ahmed, A.; Saeed, F. Quantum probability ranking principle for ligand-based virtual screening. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 365−378.

(65) Al-Dabbagh, M. M.; Salim, N.; Himmat, M.; Ahmed, A.; Saeed, F. A quantum-based similarity method in virtual screening. *Molecules* **2015**, *20*, 18107−18127.

(66) Altalib, M. K.; Salim, N. Similarity-Based Virtual Screen Using Enhanced Siamese Multi-Layer Perceptron. *Molecules* **2021**, *26*, 6669.

(67) Legendre, P. Species associations: the Kendall coefficient of concordance revisited. *J. Agric., Biol., and Environ. Stat.* **2005**, *10*, 226.

(68) Shukur, O. B.; Lee, M. H. Imputation of missing values in daily wind speed data using hybrid AR-ANN method. *Mod. Appl. Sci.* **2015**, *9*, 1.