Detecting potential HIV inhibitors using the Cross Siamese Network

Konrad Witkowski[0009−0004−2916−8672], Agnieszka Duraj[0000−0002−3047−6662], and Piotr S. Szczepaniak[0000−0002−9973−0673]

Institute of Information Technology Lodz University of Technology, al. Politechniki 8, 93-590 Lodz, Poland

Abstract. The issue of in silico analysis plays a crucial role in designing new medicines in modern day pharmaceutical industry. Selecting the best candidate for a new drug among countless molecules is a challenge which can be facilitated by machine learning methods. Following article addresses the problem of computational prediction of Human Immunodeficiency Virus (HIV) inhibition level among molecules. We introduced the cross siamese network (CSN) - a novel architecture based on siamese neural network - generating an embedding aiming to enhance the prediction process. The proposed neural net is a hybrid type model which combines embeddings generated from several subnetwork trained in estimating HIV inhibition level and other chemical properties like: solubility, lipophilicity or toxicological effects. To verify the efficiency of the proposed solution we trained a set of k-nearest neighbors classifiers on starting molecules' fingerprints and embeddings outputted by the experimental models. The results from this test showed that some versions of model enhanced the molecular embeddings, improving their utility for predicting HIV inhibition.

Keywords: In silico drug design · HIV inhibition prediction · siamese neural network

1 Introduction

It is estimated that at the end of 2023 there were 39.9 millions people with HIV, 65% of them living in the WHO African Region [25]. The virus targets the human immune system making it easier for various infections to attack the transmitter. The late HIV infection stage is called acquired immunodeficiency syndrome (AIDS). The disease is not curable, however with proper treatment it is considered to be a manageable health condition. Patients who undergo antiretroviral therapy may in fact lower the level of HIV in their blood [11]. We hope that our work may help in finding better medicines for HIV or any other disease.

In this paper, we propose a novel architecture that aims to embed an Extended-Connectivity Fingerprint (ECFP) [16] molecule representation in such a way that the subsequent classification process may be improved. The main idea consists in

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12

combining outputs of several submodels, each trained to recognize other chemical features (e.g. toxicologial effects), in one common embedding. We hypothesize that this procedure may not only lead to more information-rich vector representations but also spare resources for the laboratory experiments necessary to indicate exact measures of additional chemical features.

The main objective of this experiment is to verify whether our main model is able to make use of auxiliary information gathered from the submodels and, in effect, improve the classification of potential molecules-HIV inhibitors. For this task, we collected data available through MoleculeNet [24] - a publicly available repository of datasets containing molecules and their explored chemical parameters.

This paper is organized as follows. At the beginning of the article we present some of the solutions inspired by siamese neural network used to predict the potential inhibition of HIV and other bioactivities among molecules. In the next part we describe the siamese neural network and explain how it contributed to the architecture of CSN. The end of the article is dedicated to the experiment itself, its results and conclusions.

2 Related works

Similarly to other publications [27, 3] our network uses similarity metric to learn molecule embedding for better prediction. Attempts to improve the siamese network in cheminformatics can be performed in various ways, which means that the novelties may be differentiated for example by such factors as training data selection [27], model architecture [2] or combining multiple predictors' outputs creating this way a hybrid model [3]. In our case CSN fulfills the last 2 criteria.

Zhang et al. [27] introduced similarity based pairing of molecules dedicated for training the siamese network for regression tasks. The selection of training pairs was orchestrated by Tanimoto similarity calculated on ECFP fingerprints of molecules - each molecule from training set was assigned with its most similar counter part. For testing the new method the authors used 3 physiochemical datasets: lipophilicity [7], freesolv [17] and ESOL [8].

In [3] Altalib MK et al. attempted to improve the retrieval recall by connecting models from their earlier publications [2] using different variants of decision fusion layers and features fusion layers. Each of 4 hybrid constructions consists of 2 versions of authorial siamese neural net. To create an enhanced molecule representation a molecular fingerprint is processed parallely by 2 submodels ending up with 2 feature vectors which are merged by a features fusion layer. The role of calculating the similarity measurement lies in a decision fusion layer. The effectiveness of constructed variants was tested on MDL Drug Data Report [1] and Maximum Unbiased Validation MUV [21].

Paykan et al. [10] approached the challenge of predicting the bioactivity level of previously unseen molecules in a slightly different manner. Instead of training a model to specify similarity between 2 molecules they proposed an authorial solution called BioAct-Het which calculates the likelihood of association between the

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12

molecules structure and bioactivity class. BioAct-Het is a heterogenous siamese neural net which merges molecules' embeddings generated by preatrained graph convolutional models (loaded from DGL-LifeSci [14]) with a vector representing predicted bioactivity class. To come out with an embedding for a single bioactivity class the authors introduced Bio-Prof - a novel method mapping Morgan Fingerprints of molecules inside of the training dataset into a single vector whose indices are indicators of how significant a given substructure for occurrence of an explored bioactivity is. The following datasets were used to test the efficacy of the solution: Sider [13], Tox21 [19] and MUV [21].

Li TH et al [15] proposed a siamese neural net inspired model SNRMPACDC to estimate the synergy value of drug combination on different tissue types cell lines. Data for training and testing the model was downloaded from large-scale cancer screeing uploaded by Merck & Co [20]. SNRMPACDC is a 2-piece model from which each unit is responsible for something else. The first part is a siamese neural net which transforms ECFP fingerprint, physicochemical properties and binary toxicophores representation of a molecule into a drug feature vector. Additionally, vectors from both branches of siamese neural net are mixed by Random Matrix Projection (original solution which enabled the model to calculate the interaction potential between 2 molecules) and merged into a single vector representing drug combination features. The second part focuses on utilizing the matrices depicting cell line genomic features and mutation features. These 2 matrices are preprocessed and run through convolutional neural network module, resulting in cell line features. The outputs of 2 part of SNRMPACDC are combined together by Hadamard product and processed by multi-layer network to obtain the synergy value.

3 Model architecture

Cross siamese network is based on the idea that the distance between molecules with similar properties should be smaller than the distance between molecules with radically different characteristics. For creating such representations on hyperplane we have inspired ourselves by the siamese network designed by Bromley et al. [6] and its further applications in chemistry [27, 3, 2].

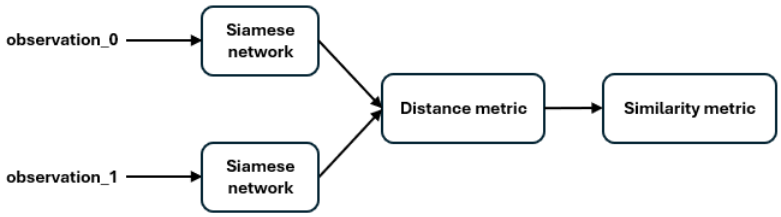

Fig. 1. Simplified architecture of siamese neural network

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12

Siamese neural network consists of 2 parallel neural networks that share the same weights. Calculating the similarity measurement between 2 observations starts with inputting them into separate parts of siamese neural network to generate their embeddings whose remoteness is calculated by distance metric like l2-norm. The task of turning the distance between 2 samples into their similarity level belongs to the similarity metric which is performed at the end of the whole procedure. The simplified architecture of siamese neural network was shown in Figure 1.

3.1 Cross Siamese Network

Cross siamese network is a hybrid model which merges output embeddings of each of n auxiliary siamese neural networks into 1 feature vector $\boldsymbol{f}_{merged}$ at fusion layer phase. The feature vector $\boldsymbol{f}_{merged}$ is calculated as follows:

$$\boldsymbol{f}_{merged} = \sum_{i=1}^{n} \boldsymbol{f}_i \boldsymbol{w}_i, \tag{1}$$

where $\boldsymbol{f}_i$ are feature embeddings generated by submodels and $\boldsymbol{w}_i$ are learnable weights. The main goal of the feature fusion layer consists in generating an enriched vector which in further steps should facilitate the predictions. Nevertheless, such an uncomplex way of combining multiple outputs of submodels may not be sufficient for preserving all of the gained information and result in a too mixed up embeddings.

Before the final output is created the consolidated vector generated by feature fusion layer is processed by convolutional blocks and linear block.

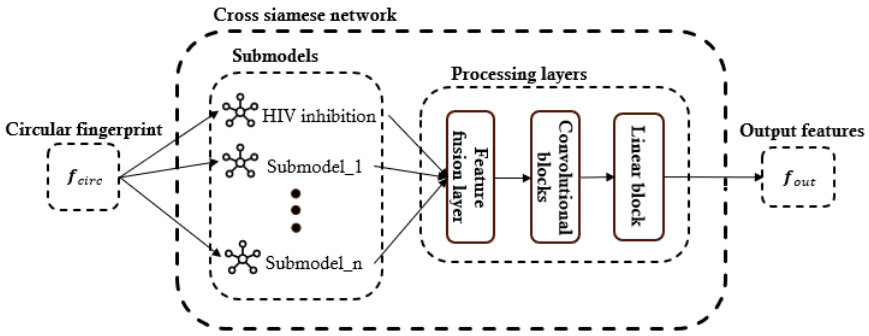

Fig. 2. Architecture of the cross siamese network

The Figure 2 shows the architecture of a network based on cross siamese network approach. This network is designed to analyze chemical and biological data, such as biological activity related to HIV inhibition, which is why the "circular fingerprint" is defined at the top of the diagram. Circular fingerprint (CF),

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12

which was more extensively described in section 4.2, is a way of representing a molecule in a numerical form, based on local molecular structures. Submodels, created using CFs, analyze various aspects of molecular features. The HIV inhibition submodel is a siamese neural net that was trained using molecules tagged with their HIV inhibition level. The remaining submodels analyze other chemical or biological properties. Before the final output is generated the data is processed by: feature fusion layer, convolutional blocks and a linear block. The 2 last structures were presented in Figure 3. To introduce the nonlinearity in the convolutional and linear blocks we used the ReLU activation function which is described by the following equation:

$$f(x) = max(0, x). \tag{2}$$

The convolutional phase is responsible for reorganizing the merged vector and extracting more features from it. It consists of 5 convolutional blocks. Each of these blocks is constructed upon convolutional layer 1d with kernel of size 1, ReLU activation function and batch normalization.

Additionally, 2 residual connections are applied. The input and output number of channels for each convolutional block is 32 except for the last one where the is 32 input channels and 1 output channel.

The linear block, whose role consists in transforming the data into a form suitable for output predictions, is composed of linear layer, ReLU activation function and batch normalization.

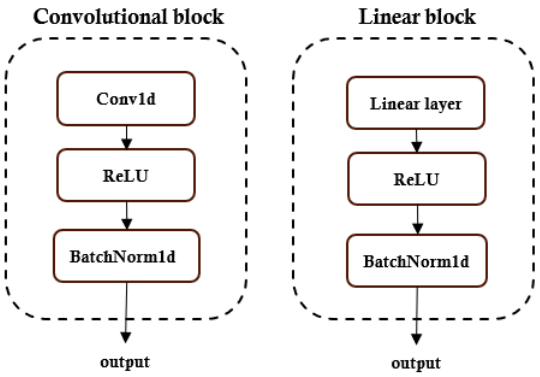

Fig. 3. Structure of convolutional and linear block

Both parts of the CSN serve the purpose of creating a goal-oriented molecular fingerprint whose form in the experiment should be designed to enhance the detection of HIV inhibitors.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12

3.2 Submodels

All of the auxiliary models, which we called siamese mol nets (SMNs), have the same kern substructure responsible for producing the feature vectors. The kern substructure was presented in Figure 4. It is based on 3 pairs of linear and batch normalization layers. The input number of features of the first linear layer is 2048 which corresponds to the length of fingerprints. However, its number of output features is 4096 and the size of the processed vector stays this way until the end, which means that the output feature vector also has 4096 elements.

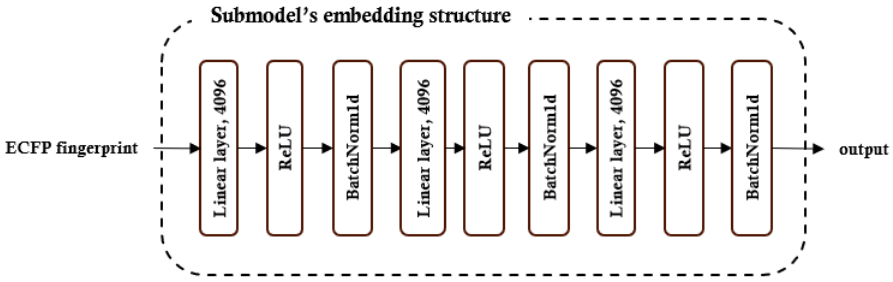

Fig. 4. Schema of kern substructures of the auxiliary models

In the case of the classification task, the structure in Figure 4 suffices for the entire model. On the other hand, the regression task requires additional layers whose goal is to stack feature vectors coming from 2 molecules, calculate their averages for each index, and using several linear layers generate the predicted real value output. The first linear layer reduces the 4096 element vectors to their equivalents of length 2048. This number is further condensed to 64 by the second linear layer. The output receives its final shape by the third and final layer, which generates a single-element vector. The described processing flow was shown in Figure 5.

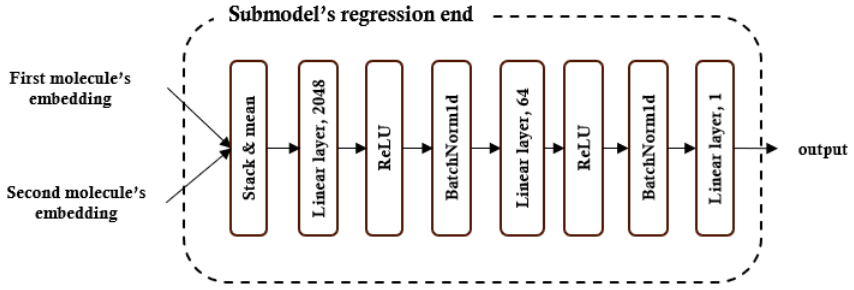

Fig. 5. Schema of submodel's regression end

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12

In both scenarios, the shape of the output has a significant value for the task the model was trained for. In the case of classification, a large number of output vector elements allows to embed it in the molecular representation space more accurately. In contrast, the regression end outputs a single-element vector which is the estimated distance between two molecules' embeddings.

4 Experiment

The goal of the experiment consisted in testing the new architecture in several configurations and verifying whether its embeddings can enhance the classification of potential HIV inhibitor. To perform the experiment, we decided to train the models on publicly available molecular datasets. Finally, we gathered the results from the whole experiment and described them in the next chapter.

4.1 Datasets

To compose the CSNs we chose to use the datasets from MoleculeNet. In order to elastically download the molecules and their ECFP fingerprints we utilized the DeepChem [23] library written in Python.

The starting point for our experiment was the HIV dataset provided by Drug Therapeutics Program which tested 40 000 molecules for inhibition of HIV replication and assigned to each of the explored compound a label describing their level of activeness: confirmed active, confirmed moderately active and confirmed inactive. The authors of DeepChem chose to merge the first 2 labels into a single category i.e. 1 [26], while treating all remaining samples as 0. For training and testing purposes, we used the scaffold splitter - a dedicated cheminformatics tool that identifies the core substructures (scaffolds) of the molecules. Although the scaffold splitter was designed to resemble real-world conditions in cheminformatics, it has certain limitations. This method divides the dataset by identifying ring structures to ensure that the training and testing sets remain distinct. However, this approach may disadvantage molecules that lack such structural features.

In order to build the CSNs we selected 3 auxiliary groups of molecules:

- LIPO is a part of data provided by AstraZeneca and stored by ChEMBL [7]. It contains 4200 molecules with their lipophilicity scores measured by octanol/water distribution. The train and test parts of the dataset were divided using scaffold splitter.
- Delaney contains information about solubility of 1128 molecules. The train and test parts of the dataset were divided using scaffold splitter.
- TOX21 [19] dataset consists of molecules and their toxicity measurement understood as compound activity in all nuclear receptor signaling pathways. For the purpose of experiment we choose following categories: androgen receptor (NR_AR), androgen receptor ligand binding (NR_AR_LBD), androgen receptor aryl hydrocarbon receptor (NR_AR_AHR) and aromatase receptor (NR_AROMAT). The train and test parts of the dataset were divided using scaffold splitter.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12

4.2 Extended-Connectivity Circular Fingerprint

The ECFP is a molecular fingerprint created using a refinement of Morgan Algorithm [18]. The final fingerprint $\boldsymbol{f}$ is a vector of a predefined length S. Its form is a product of R layers that during processing of a single molecule update the information assigned to single atoms. The process of generating an ECFP fingerprint $\boldsymbol{f}$ (of length S) of a molecule operates as follows:

1. Initialize the final fingerprint with zeros: $\boldsymbol{f} \leftarrow \mathbf{0}_S$
2. For each layer L loop over each atom a and perform following operations:
   (a) concatenate atom feature vectors of neighbors of atom a to an auxiliary vector $\boldsymbol{v}$: $\boldsymbol{v} \leftarrow [\boldsymbol{r}_a, \boldsymbol{r}_0, ..., \boldsymbol{r}_n]$
   (b) update atom's a feature vector: $\boldsymbol{r}_a \leftarrow hash(\boldsymbol{v})$
   (c) retrieve an index i of fingerprint $\boldsymbol{f}$ to update its value to 1: $i \leftarrow mod(\boldsymbol{r}_a, S)$
3. Output fingerprint $\boldsymbol{f}$

The final fingerprint is a binary vector whose final form is independent of the task for which it was created.

4.3 Training process

Depending on the predicted value the training process required different loss function which we describe in further details in the following subsection.

For the LIPO and Delaney datasets, where the labels have real values, we decided to measure the quality of the models with the mean square error:

$$L(y_i, \hat{y_i}) = \sum_{i=1}^{n} (y_i - \hat{y_i})^2, \tag{3}$$

where y_i stands for the absolute value of difference between labels of 2 selected molecules, $\hat{y_i}$ is the output of model and n is the number of pairs in mini batch.

For other datasets whose task consisted in binary classification we used triplet margin loss [4] which can be formulated in the following way:

$$L(a_i, p_i, n_i) = max\{\|a_i - p_i\|_2 - \|a_i - n_i\|_2 + margin, 0\}, \tag{4}$$

where a_i is the i-th anchor, p_i is the i-th positive sample, n_i is the negative sample. In case of the experiment the margin was set to 1. To force the models to focus on more challenging examples we decided to use the hard batch-hard mining which for a given observation looks inside of a batch for the nearest negative sample and for farthest positive sample. Additionally, we made sure that the samples from the minority class were equally split across batches. We hoped that this way we could achieve a more stable training process.

To mitigate the risk of overtraining the models in classification task we came up with an idea of boosting the loss of those triplets where the anchor came from the minority class as follows:

$$w_i = \begin{cases} \frac{\text{number of negative samples}}{\text{number of positive samples}}, & a_i \text{ is a positive sample.} \\ 1, & a_i \text{ is a negative sample} \end{cases} \tag{5}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12

We designed the experiment so that one part of the classification models was trained with stable weights and the other part with boosted weights. The training process using these 2 weighting strategies required different number of batch samples to obtain a controllable loss decrease flow. Specifically, for models dedicated for HIV data, we set the batch size to 32 in the boosted weights scenario and 128 in the stable weights scenario. All training processes were performed by Adam optimizer [12], with the learning rate set to 1e-5. The initial weights of the convolutional and linear layers in all models were generated using the Xavier uniform distribution [5], with the gain of 1.0. The starting biases of all layers were manually set to 0.01.

4.4 Evaluation methods

Our primary goal was to determine whether the models could effectively map ECFP fingerprints to equivalent vectors, ensuring that vectors from the same class were closer together, while those from opposite classes were farther apart. To evaluate the quality of the embeddings generated by the models, we used the k-nearest neighbors algorithm for each version of both SMN and CSN. This procedure consisted in training the classifier on embeddings outputted by an explored type of siamese neural network and testing it on the equivalent embeddings derived from the test dataset.

The k-nearest algorithm [22] is a nonparametric classification method. Given a training set $D = \{(x_n, y_n)\}_{n=1}^{N}$ an observation x_i, where $i \in [1, N]$, will be assigned with the most frequent class among k nearest neighbors. For calculating the distance we chose the l2 metric. In the experiment we made use of 2 variants of k-nearest algorithm, both trained on the embeddings from the training set. The first type was set to 4 and the second one to 3 nearest neighbors. This approach ensured that an observation was assigned to the class with the majority representation among its 3 nearest neighbors.

In order to estimate the performance of the classifiers, we used metrics such as accuracy, precision and recall [9].

The accuracy, assessing the overall quality of the classification process, is given as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives.

For measuring how well the algorithm works strictly on the embeddings from the minority class (in this case the positive one) we used:

$$precision = \frac{TP}{TP + FP} \tag{7}$$

and recall formulated in the following way:

$$recall = \frac{TP}{TP + FN}. \tag{8}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12

The calculated metrics belong to the standard procedures for measuring the classification process. Thanks to them, it is possible not only to assess the prediction process but also to notice the potential threat of overemphasizing the majority class.

5 Results

We divided the description of the results obtained from the experiment into 2 subsections addressing the weighting strategies. To measure how well the SMN and CSN performed, we used the results of the k-nearest neighbors classifier (Table 1) trained on the initial ECFP fingerprints from the HIV dataset as a reference point.

Table 1. Performance metrics of k-nearest neighbor trained on starting ECFP fingerprints of HIV dataset.

Phase	Accuracy	Precision	Recall
Train	0,9683	0,6262	0,3807
Test	0,9694	0,5588	0,1462

The abbreviations in the tables that represent model types combine the underlying structure with the task for which the model was trained. The full list of created models looks in the following way:

- SMN_HIV - SMN trained on HIV inhibition data
- CSN_HIV - CSN composed of a SMN trained on HIV inhibition data
- CSN_HIV_LIPO - CSN composed of a SMN trained on HIV inhibition data and a SMN trained on lipophilicity data
- CSN_HIV_TOX_NR_AR - CSN composed of a SMN trained on HIV inhibition data and a SMN trained on compound activity in androgen receptor data
- CSN_HIV_TOX_NR_AROMAT - CSN composed of a SMN trained on HIV inhibition data and a SMN trained on compound activity in aromatase receptor data
- CSN_HIV_TOX_NR_AR_LBD - CSN composed of a SMN trained on HIV inhibition data and a SMN trained on compound activity in androgen receptor ligand binding data
- CSN_HIV_TOX_NR_AR_AHR - CSN composed of a SMN trained on HIV inhibition data and a SMN trained on compound activity in androgen receptor aryl hydrocarbon receptor data
- CSN_HIV_DELANEY - CSN composed of a SMN trained on HIV inhibition data and a SMN trained on solubility data

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12

5.1 Stable weights

The training process based on constant weights, whose loss scores were shown in Table 2, led to CSN_HIV_TOX_NR_AR achieving the smallest loss i.e. 0.86563. However, this version of siamese network did not produce the most effective embeddings for the k-nearest neighbor classifier. As presented in Table 3 the CSN_HIV_TOX_NR_AR was among the weakest models in terms of precision, achieving only 0.6123 on the training dataset and 0.5161 on the testing dataset. These results indicate that the embeddings generated by CSN_HIV_TOX_NR_AR perform even worse for classification purposes than the original molecular fingerprints.

Table 2. Loss results for the normal weights variant of training process

Model type	Loss
SMN_HIV	0.99962
CSN_HIV	0.89432
CSN_HIV_LIPO	1.04699
CSN_HIV_TOX_NR_AR	**0.86563**
CSN_HIV_TOX_NR_AROMAT	1.03654
CSN_HIV_TOX_NR_AR_LBD	0.93201
CSN_HIV_TOX_NR_AHR	0.97224
CSN_HIV_TOX_NR_ER	1.00304
CSN_HIV_TOX_NR_ER_LBD	1.26426
CSN_HIV_DELANEY	0.96148

Table 3. Training and Testing performance metrics (accuracy, precision, recall) for different model variants.

Model type	**Training**			**Testing**		
	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.
SMN_HIV	0.9721	0.8668	0.3011	0.9696	0.8571	0.0462
CSN_HIV	0.9673	0.6137	0.3417	0.9686	0.5110	0.1769
CSN_HIV_LIPO	0.9666	0.6255	0.2670	0.9691	0.6000	0.0922
CSN_HIV_TOX_NR_AR	0.9664	0.6123	0.3289	0.9688	0.4857	0.1632
CSN_HIV_TOX_NR_AROMAT	0.9666	0.6140	0.3433	0.9686	0.4857	0.1323
CSN_HIV_TOX_NR_AR_LBD	0.9678	0.6310	0.3117	0.9686	0.5926	0.1231
CSN_HIV_TOX_NR_AR_AHR	0.9670	0.6755	0.2281	0.9686	0.5568	0.0769
CSN_HIV_DELANEY	0.9683	0.6250	0.3856	0.9696	0.5439	0.2385

The most effective molecular fingerprints were undeniably generated by SMN_HIV, enabling the k-nearest neighbor classifier to achieve a precision of 0.8668

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12

on the training set and 0.8571 on the test set. It should be noted that in this case the recall levels obtained at the training phase (0.3011) and the testing phase (0.0462) were noticeably different. This difference can be likely attributed to the scaffold splitter, which divided the main dataset in a such way that the molecules of the same class from the training and testing phase were as structurally various as possible. The fingerprints generated by SMN_HIV significantly enhanced the starting ECFP versions in terms of precision.

Among the CSNs, the best performance was delivered by the embeddings of CSN_LIPO and CSN_HIV_TOX_NR_AR_LBD, achieving precision values of 0.6 and 0.5926 on the test set, respectively. However, these scores are comparable with those achieved on the initial ECFP fingerprints.

In the normal weighting scenario the loss value did not align with the performance of the k-nearest classifier. The most effective neural net, SMN_HIV, reached a loss of 0.9962, which was not indicative of its ability to generate high-quality embeddings. This may suggest that the CSNs may have focused on optimizing different components of the triplet loss compared to SMN_HIV. The CSNs may owe the biggest decrease in their loss values to the minimized distance between samples from the same class. While this ensures that observations within the same class are closely grouped, it does not necessarily establish a clear distinction between different classes.

5.2 Boosted weights

In the boosted weighting variant of loss calculation (Table 4), CSN_HIV_TOX_NR_AROMAT emerged as the most promising model, achieving a loss of 1.56074. It is worth noticing that the CSN_HIV_TOX_NR_AR, which had the lowest level under the normal weighting strategy, also performed well with a loss of 1.67667.

Table 4. Loss results for the boosted weights of training process

Model type	Loss
SMN_HIV	2.0035
CSN_HIV	1.69165
CSN_HIV_LIPO	2.09365
CSN_HIV_TOX_NR_AR	1.67667
CSN_HIV_TOX_NR_AROMAT	**1.56074**
CSN_HIV_TOX_NR_AR_LBD	1.81694
CSN_HIV_TOX_NR_AHR	2.06221
CSN_HIV_TOX_NR_ER	1.72272
CSN_HIV_TOX_NR_ER_LBD	1.93214
CSN_HIV_DELANEY	1.68214

Nevertheless, the CSN_HIV_TOX_NR_AROMAT did not reach the best results for k-nearest classifier among models trained with boosted weights. As

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12

presented in Table 5 the CSN_HIV_TOX_NR_AROMAT's precision levels (0.6498 on the training dataset and 0.6 on the testing dataset) were surpassed by SMN_HIV's which reached 0.8788 on the training dataset and 0.7143 on the testing dataset.

The best CSN (in terms of precision) turned out to be the CSN_HIV_TOX_NR_AHR which scored 0.6708 on the training dataset and 0.8235 on the testing dataset.

However, this successful score did not align with the equivalent from the scenario with stable weights. The experiment revealed that for CSNs there is no definitive pattern indicating how the neural net will perform under different weighting strategies. This suggests that, in more complex models, the observations themselves may play a more critical role in training than the weighting strategies. Additionally, the inconsistent results could also be explained by the datasets consisting of different molecules for each task.

Table 5. Training and Testing performance metrics (accuracy, precision, recall) for different model variants.

Model type	Training			Testing		
	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.
SMN_HIV	0.9829	0.8788	0.6299	0.9713	0.7143	0.1538
CSN_HIV	0.9672	0.6423	0.2784	0.9677	0.4000	0.0462
CSN_HIV_LIPO	0.9671	0.6118	0.3287	0.9684	0.5000	0.1538
CSN_HIV_TOX_NR_AR	0.9685	0.6471	0.3482	0.9703	0.6333	0.1462
CSN_HIV_TOX_NR_AROMAT	0.9672	0.6498	0.2711	0.9691	0.6000	0.0692
CSN_HIV_TOX_NR_AR_LBD	0.9676	0.6267	0.3312	0.9699	0.5938	0.1462
CSN_HIV_TOX_NR_AR_AHR	0.9693	0.6708	0.3523	0.9711	0.8235	0.1077
CSN_HIV_DELANEY	0.9675	0.6153	0.3531	0.9684	0.5000	0.1231

Similarly as under the normal weighting strategy, the achieved loss did not translate into the most effective model. Moreover, none of the evaluated models from both strategies scored an accuracy below 96%, suggesting that the majority class was not overlooked. The key distinction between the variants of SMN_HIV was the relatively high recall level scored by the boosted weighting version: 0.6299 on the training dataset and 0.1538 on the testing dataset. This improvement indicates that the SMN trained using boosted weights works better at detecting rare classes.

6 Conclusions

The main conclusion is that models based on siamese neural nets are able to enhance the classification of potential HIV inhibitors. The embeddings generated by the SMN_HIVs managed to outperform the starting fingerprints leading to

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12

a significant improvement of the precision of the k-nearest neighbors obtained on the test dataset. Specifically, the classifier achieved a precision of 0.5588 using the initial fingerprints, 0.8571 with embeddings generated by the SMN_HIV trained with the stable weights, and 0.7143 with embeddings produced by the SMN_HIV trained with the boosted weights. Furthermore, the boosted weighting strategy allowed to increase the control over triplet loss function enabling a precision-recall trade-off. However, the embeddings generated by the CSNs were comparable in their classification quality to that of the initial fingerprints, raising questions about the underlying reasons for this outcome. In summary, the proposed SMN_HIV architecture, along with the weighting strategy, may find application in the search for potential drugs beyond HIV. Nevertheless, the perspective of CSN remains an important topic for further discussion.

References

1. Accelrys: Mdl drug data report (mddr). http://www.accelrys.com, accelrys Inc.: San Diego, CA, USA; Accessed online on 31 October 2021
2. Altalib, M.K., Salim, N.: Similarity-based virtual screen using enhanced siamese multi-layer perceptron. Molecules **26**, 6669 (2021). https://doi.org/10.3390/molecules26216669
3. Altalib, M.K., Salim, N.: Hybrid-enhanced siamese similarity models in ligand-based virtual screen. Biomolecules **12**(11), 1719 (2022). https://doi.org/10.3390/biom12111719
4. Balntas, V., Riba, E., Ponsa, D., Mikolajczyk, K.: Learning local feature descriptors with triplets and shallow convolutional neural networks. BMVC Proceedings pp. 119.1–119.11 (2016). https://doi.org/10.5244/C.30.119
5. Bengio, Y., Glorot, X.: Understanding the difficulty of training deep feed forward neural networks. International Conference on Artificial Intelligence and Statistics pp. 249–256 (01 2010)
6. Bromley, J., Guyon, I., LeCun, Y., Sickinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. In: Advances in Neural Information Processing Systems. vol. 6, pp. 737–744 (1993)
7. Chemdbl: Chembl3301361. https://www.ebi.ac.uk/chembl/document_report_card/CHEMBL3301361/, accessed 30 September 2024
8. Delaney, J.S.: Esol: Estimating aqueous solubility directly from molecular structure. Journal of Chemical Information and Computer Sciences **44**(3), 1000–1005 (2004). https://doi.org/10.1021/ci034243x
9. Google: Classification: Accuracy, recall, precision, and related metrics. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall, accessed 5 December 2024
10. Heyrati, M.P., Ghorbanali, Z., Akbari, M., Pishgahi, G., Zare-Mirakabad, F.: Bioact-het: A heterogeneous siamese neural network for bioactivity prediction using novel bioactivity representation. ACS Omega **8**(47), 44757–44772 (2023). https://doi.org/10.1021/acsomega.3c05778
11. HIV.gov: What are hiv and aids. https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids, accessed: 2024-09-26
12. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (12 2014)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12

13. Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The sider database of drugs and side effects. Nucleic Acids Research **44**, D1075–D1079 (2016). https://doi.org/10.1093/NAR/GKV1075
14. Li, M., Zhou, J., Hu, J., Fan, W., Zhang, Y., Gu, Y., Karypis, G.: Dgl-lifesci: An open-source toolkit for deep learning on graphs in life science. ACS Omega **6**(41), 27233–27238 (2021). https://doi.org/10.1021/acsomega.1c04017
15. Li, T.H., Wang, C.C., Zhang, L., Chen, X.: Snrmpacdc: computational model focused on siamese network and random matrix projection for anticancer synergistic drug combination prediction. Briefings in Bioinformatics **24**(1), bbac503 (2023). https://doi.org/10.1093/bib/bbac503
16. Micheli, A.: Neural network for graphs: A contextual constructive approach. IEEE Transactions on Neural Networks **20**(3), 498–511 (2009). https://doi.org/10.1109/TNN.2008.2010350
17. Mobley, D.L., Guthrie, J.P.: Freesolv: a database of experimental and calculated hydration free energies, with input files. Journal of Computer-Aided Molecular Design **28**, 711–720 (2014). https://doi.org/10.1007/s10822-014-9747-x
18. Morgan, H.L.: The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. Journal of Chemical Documentation **5**(2), 107–113 (1965). https://doi.org/10.1021/c160017a018, https://doi.org/10.1021/c160017a018
19. National Center for Advancing Translational Studies: Tox21 data challenge 2014. https://tripod.nih.gov/tox21/challenge/data.jsp, accessed 1 October 2024
20. O'Neil, J., Benita, Y., Feldman, I., Chenard, M., Roberts, B., Liu, Y., Li, J., Kral, A., Lejnine, S., Loboda, A., Arthur, W., Cristescu, R., Haines, B.B., Winter, C., Zhang, T., Bloecher, A., Shumway, S.D.: An unbiased oncology compound screen to identify novel combination strategies. Molecular Cancer Therapeutics **15**, 1155–1162 (2016)
21. Rohrer, S.G., Baumann, K.: Maximum unbiased validation (muv) data sets for virtual screening based on pubchem bioactivity data. Journal of Chemical Information and Modeling **49**(2), 169–184 (2009). https://doi.org/10.1021/ci8002649
22. Shi, Y., Yang, K., Yang, Z., Zhou, Y.: Chapter two - primer on artificial intelligence. In: Shi, Y., Yang, K., Yang, Z., Zhou, Y. (eds.) Mobile Edge Artificial Intelligence, pp. 7–36. Academic Press (2022). https://doi.org/https://doi.org/10.1016/B978-0-12-823817-2.00011-5
23. The DeepChem Project: Model classes. https://deepchem.readthedocs.io/en/latest/api_reference/models.html, accessed 21 March 2024
24. The DeepChem Project: Moleculenet. https://deepchem.readthedocs.io/en/latest/api_reference/moleculenet.html, accessed 29 September 2024
25. WHO: Hiv and aids. https://www.who.int/news-room/fact-sheets/detail/hiv-aids, accessed 26 September 2024
26. 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. Chemical Science **9**(2), 513–530 (2017). https://doi.org/10.1039/c7sc02664a
27. Zhang, Y., Menke, J., He, J., Nittinger, E., Tyrchan, C., Koch, O., Zhao, H.: Similarity-based pairing improves efficiency of siamese neural networks for regression tasks and uncertainty quantification. Journal of Cheminformatics **15**, 75 (2023). https://doi.org/10.1186/s13321-023-00744-6

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_12