Detecting potential HIV inhibitors using the Cross Siamese Network

Konrad Witkowski[a], Agnieszka Duraj[a], Piotr S. Szczepaniak[a]

[a] Institute of Information Technology, Lodz University of Technology, Politechniki 8, Lodz 93-590, Poland

25th International Conference on Computational Science
Singapore, 7-9 July 2025

Presentation plan

1. Introduction
2. Models
3. Experiment
4. Summary

Introduction

Human Immunodeficiency Virus (HIV)

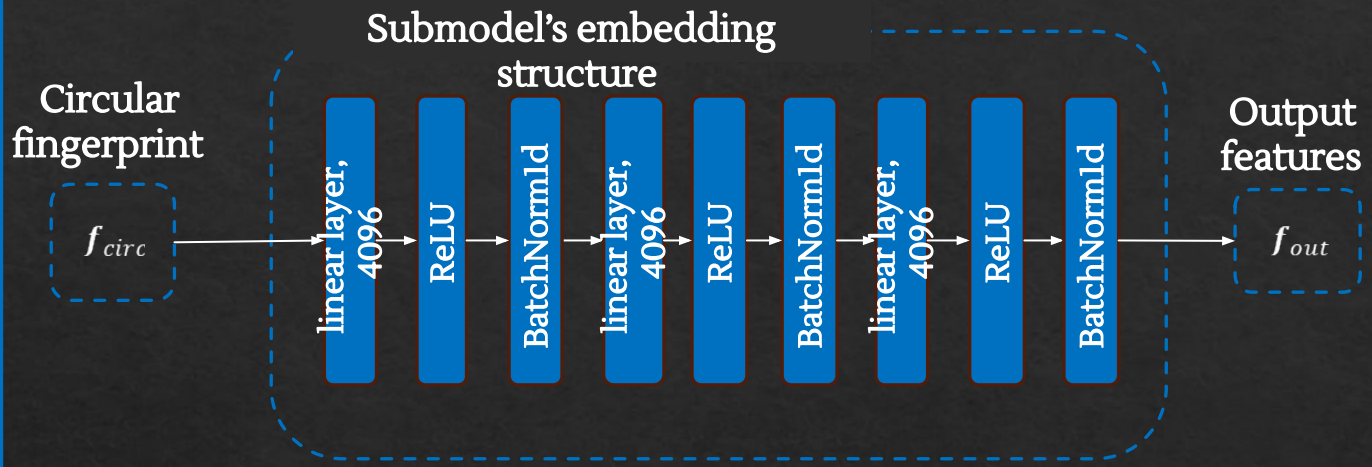
- HIV is a Lentivirus (subgroup of retrovirus) targeting the human immune system. HIV may lead to acquired immunodeficiency syndom (AIDS).
- At the end of 2023 there were approximately 39.9 millions people with HIV, 65% of them living in the WHO African Region.
- AIDS is not curable. However, undertaking an antiviral therapy may slow down the disease and prolong the life expectancy of a patient.

Publication

- The publication introduces a novel machine learning model - Cross Siamese Network (CSN) based on Siamese Network architecture.
- CSN was tested on indicating the HIV inhibitors

Models - Siamese Mol Net (SMN)

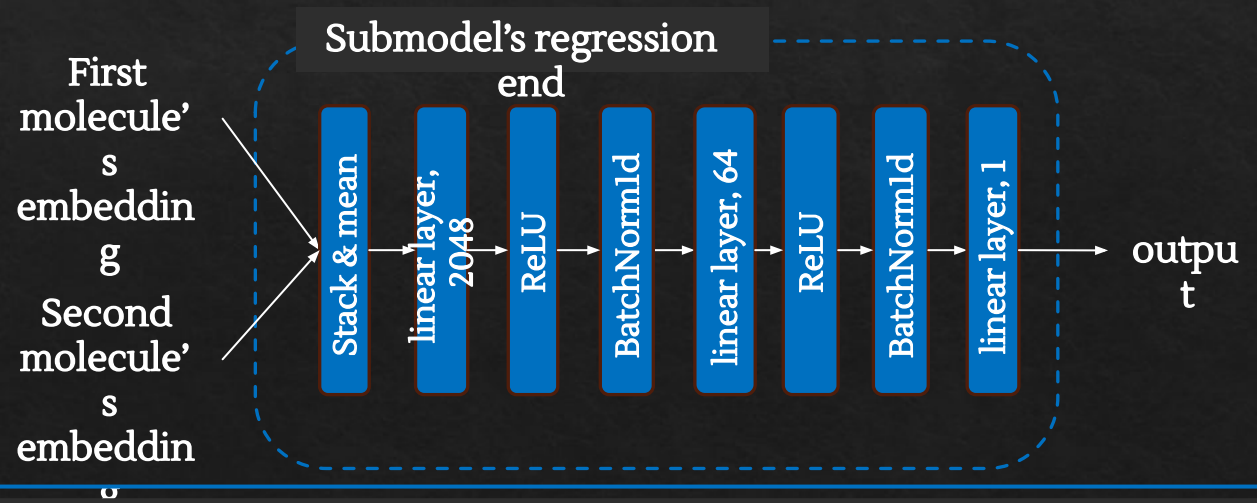
SMN as classificator



- The SMN receives as input Circular fingerprints of length 2048
- The embedding structure generates a vector with dimensionality of 4096

SMN as regressor

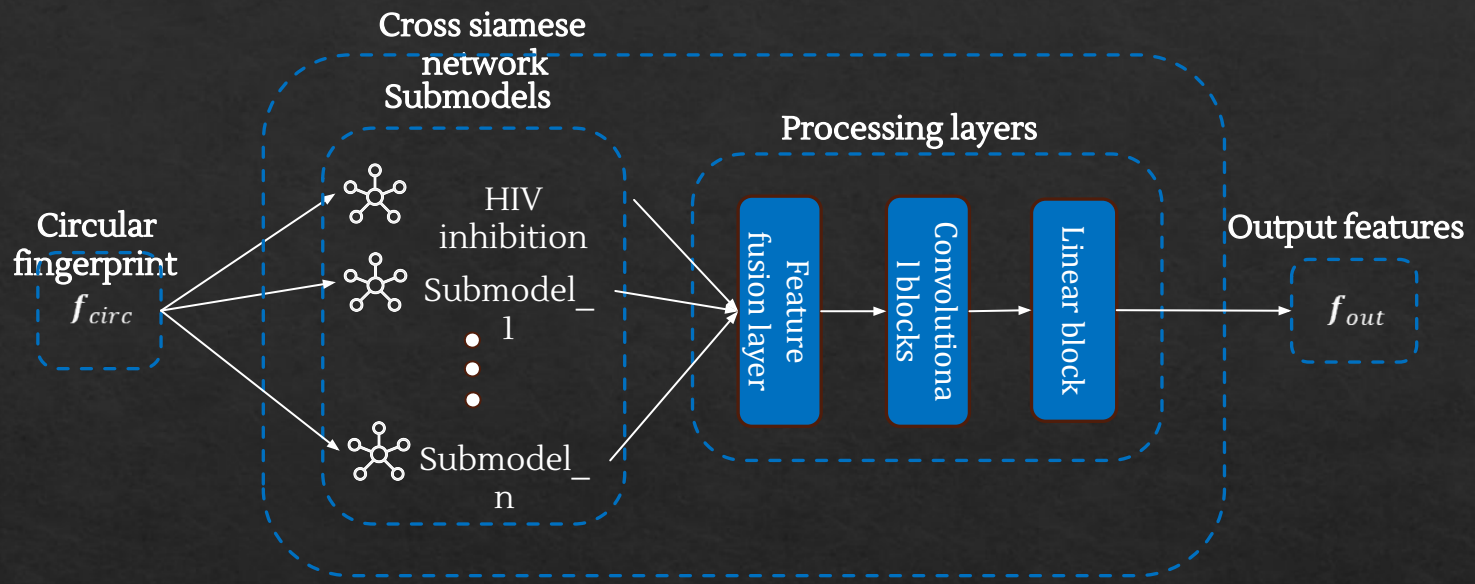
The regression type of SMN stacks the vector representations created by the embedding structure and outputs a single element vector – estimated value



Models - Cross Siamese Network (CSN)

Architecture

- Model consists of several submodels (SMNs) whose outputs is merged by feature fusion layer
- The input f_{circ} for each of SMN is a Circular Fingerprint of length 2048

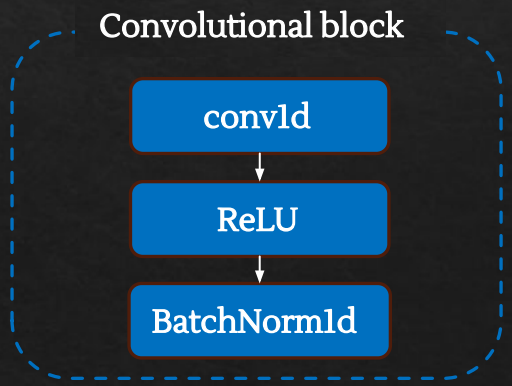


Processing layers

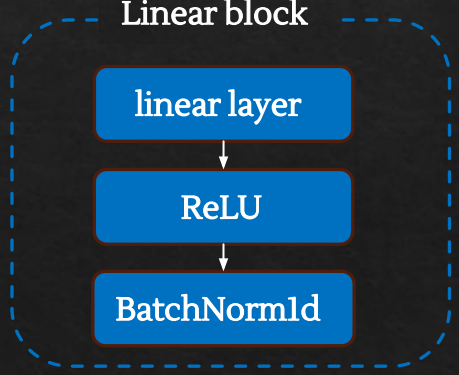
Feature Fusion layer

$$f_{merged} = \sum_{i=1}^{n} f_i w_i$$

5 X



1 X



Experiment - overview

Description

The experiment aimed to verify the efficacy of the new architecture in indicating the potential HIV inhibitors. To conduct the evaluation, we trained a set of auxiliary SMNs which were used as components for the CSNs.

Metrics

Accuracy

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Precision

$$\frac{TP}{TP + FP}$$

Recall

$$\frac{TP}{TP + FN}$$

Reference point

K-nearest neighbors (initial Circular fingerprints)

Phase	Accuracy	Precision	Recall
Train	0.9683	0.6262	0.3807
Test	0.9694	0.5588	0.1462

Experiment - datasets

HIV inhibitors

The datasets consists of **41 719** molecules, each described with one of three labels CA (confirmed active), CM (confirmed moderately active), CI (confirmed inactive).

Class	Count	Share
CA	456	1%
CM	1068	3%
CI	40 195	96%

Auxiliary datasets

Lipo

4 200 molecules with their lipophilicity score measured by octanol/water distribution.

Delaney

1 128 molecules with their solubility scores.

TOX21

12 707 molecules and their toxicity measurement (compound activity in all nuclear receptor signaling pathways). Selected 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).

Experiment - training

Loss functions

RMSE

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

Triplet margin loss

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

Batch construction

For batch composition we utilized hard batch mining. We also made sure that each batch had the same proportion of positive and negative samples.

Weighting strategies

Standard weights

$w_i = 1$

Boosted weights

$$w_i = \begin{cases} \dfrac{number\ of\ neg.\ samples}{number\ of\ pos.\ samples}, a_i\ is\ a\ pos.sample \\ 1, a_i\ is\ a\ neg.\ sample \end{cases}$$

Experiment - results

Standard weights

- The best result in terms of precision was achieved by SMN_HIV – the model reached approximately 0.86 on the test set. Its recall was around 0.05.
- Adding the auxiliary models to the CSN did not lead to better performance.

Model	Training			Testing		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
SMN_HIV	**0.9721**	**0.8668**	**0.3011**	**0.9696**	**0.8571**	**0.0462**
CSN_HIV	0.9673	0.6137	0.3417	0.9686	0.511	0.1769
CSN_HIV_LIPO	0.9666	0.6255	0.267	0.9691	0.6	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	0.614	0.3433	0.9686	0.4857	0.1323
CSN_HIV_TOX_NR_AR_LBD	0.9678	0.631	0.3117	0.9686	0.5926	0.1231
CSN_HIV_TOX_NR_AR_AHR	0.967	0.6755	0.2281	0.9686	0.5568	0.0769
CSN_HIV_DELANEY	0.9683	0.625	0.3856	0.9696	0.5439	0.2385

Boosted weights

- The boosted weighting strategy increased the recall of SMN_HIV to 0.15, but this came at the cost of reduced precision (0.71).
- A similar effect was observed in the case of the CSN network — increased recall simultaneously reduced precision.

Model	Training			Testing		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
SMN_HIV	**0.9829**	**0.8788**	**0.6299**	**0.9713**	**0.7143**	**0.1538**
CSN_HIV	0.9672	0.6423	0.2784	0.9677	0.4	0.0462
CSN_HIV_LIPO	0.9671	0.6118	0.3287	0.9684	0.5	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	0.6	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	0.5	0.1231

Summary

Conclusion

- The introduced architecture was able to enhance the quality of molecular embeddings for indicating potential HIV inhibitors.
- The boosted weighting strategy allowed for control of the precision-recall trade-off during the training process.
- The molecular representations generated by the CSNs were less effective than those produced by the SMNs.

Next steps

- Refine the architecture, training approach and propose a method for visualizing key molecular substructures.
- Develop an algorithm (data splitter) for dividing a set of chemical molecules into training and test subsets.