**ISBN** 978-981-11-0008-6
**Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)**
*Tokyo, Japan, 17-19 June, 2016, pp. 655-659*
doi: 10.18178/wcse.2016.06.115

# Deep Belief Networks for Ligand-Based Virtual Screening of Drug Design

Aries Fitriawan [1], Ito Wasito [1+], Arida Ferti Syafiandini [1], Azminah Azminah [2], Mukhlis Amien [1] and Arry Yanuar [2]

[1] Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
[2] Faculty of Pharmacy, Universitas Indonesia, Depok, Indonesia

**Abstract.** Virtual screening (VS) is a computational technique used in drug discovery. Virtual Screening process usually works by identifying structures that are most likely to bind the target of drug. Virtual screening is usually based on compound similarity or database docking. Thus, the identification for drug compounds based on structure classification still remain as a challenging task. The purpose of this research is to find a new approach for ligand-based virtual screening using machine learning technique. In this paper, the classification has been done by using Deep Belief Networks (DBN) method. The data from *Nicotinamide Adenine Dinucleotide* (NAD) protein target family were used for training and testing the model. This research used four protein target classes from literature and two protein target classes from DUD-E docking website. Feature were obtained from molecular fingerprint descriptor. The experiments result show that DBN method outperform the existing pharmacophore approach.

**Keywords:** deep belief networks, deep learning, drug discovery, virtual screening.

## 1. Introduction

In pharmacology, protein targets are subset of proteins that are affected during drug interaction [1]. Protein targets are usually used to measure the ability of drugs in disease treatment. Compared to the amount of compounds in the whole world, only a small amount of compounds that are proved to be useful for drug composition [2]. Therefore, a computer-aided drug design is done to help reducing the time and cost of laboratory experiment, one of which is virtual screening using pharmacophore analysis [3].

Cheminformatic studies found that computer science approaches, such as similarity measure, bipartite graph and some machine learning techniques are quite effective in finding interaction between drug and its protein targets [4][5][6][7]. For instance, Support Vector Machine (SVM), as one of machine learning methods, it can be employed to classify whether a compound is drug or not [5]. However, its SVM model could not identify the protein targets of each drug. Analysis of chemical compounds similarity have been found by Johnson and Maggiora [8]. That research concluded that compounds with similar structure tend to have similar properties. Utilizing this concept, identification for drug compounds based on structure classification still remain as a challenging task

In this paper, a framework of ligand based virtual screening using Deep Belief Networks (DBN) is proposed. Ligand based virtual screening uses compound similarity as its base. It is usually done using pharmacophore analysis [3]. DBN was first developed by Hinton [9] and utilize maximum likelihood in initiating learning parameters. In order to reduce required amount of time, DBN uses contrastive divergence (CD) with gibbs sampling as optimizer [10]. In this paper, DBN is used for drug compound classification. There are two datasets used for experiments. The first one contains of *Nicotinamide Adenine Dinucleotide*

[+] Corresponding author. Tel.: +62 858-4276-1204
*E-mail address*: ito.wasito@cs.ui.ac.id

(NAD) family. These drug compounds were collected form literature [11]. NAD protein family is commonly used for Alzheimer's and Parkinson disease treatment [12]. The second one is another NAD protein family from DUD-E docking [13]. The result from this paper is expected to produce an alternative approach for ligand-based virtual screening with better performance.

## 2. Methodology

This research was developed under the methodology (see Fig. 1). This methodology aimed to develop and evaluate the DBN architecture for virtual screening of drug design.

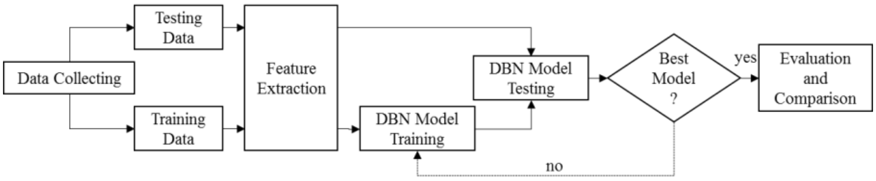


Fig. 1: The scheme for the research method used in developing DBN method for virtual screening of drug design.

## 3. Data and Feature Extraction

### 3.1. Data

This research use two dataset. The first one collected from literature and database, such as ChEMBL database, MUBD-HDACs database, PubMed publication, and ACS publication [11]. The second one is dataset from DUD-E website (dude.docking.org) [13]. Protein targets are taken from NAD protein family. Every set of data consist of drug compounds ligand and their decoy. The ligands will be represented as class 1, meanwhile class 0 for the decoys. The detail of data can be seen at Table 1. This data will be separated as 75% training data and 25% testing data.

Table 1. The number compounds for each data collection

| Protein Target | Ligand | Decoy |
|---|---|---|
| Literature Dataset | | |
| Sirtuin-1 activator (SIRT1 activator) | 48 | 240 |
| Sirtuin-1 inhibitor (SIRT1 inhibitor) | 35 | 175 |
| Histone deacetylases 4 (HDAC4) | 39 | 195 |
| Histone deacetylases 7 (HDAC7) | 24 | 120 |
| DUD-E Docking Dataset | | |
| Histone deacetylases 2 (HDAC2) | 185 | 925 |
| Histone deacetylases 8 (HDAC8) | 170 | 850 |

### 3.2. Feature Extraction

This research used substructure keys-based fingerprint descriptor for feature. Substructure key-based fingerprint provides a bit value based on the existence of a molecular substructure (see Figure 2). The substructure is already registered in the structural keys that have been collected before. There are several substructures keys. PubChem [2] issued as many as 881 structural keys. While research from Klekota and Roth [14] collected as many as 4860 structural keys. This research compared the both of fingerprints as feature.

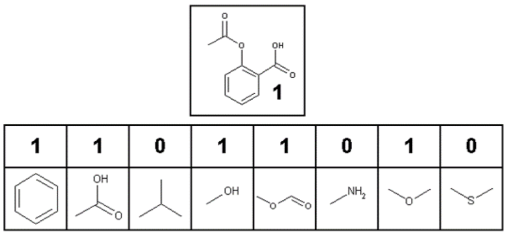


Fig. 2: Illustration for substructure keys-based fingerprint as feature.

# 4. Deep Belief Networks

The idea behind Deep Belief Networks (DBN) [9] is allowing any inter-layer model of Restricted Boltzmann Machine (RBM) to receive a different representation of the data from its output. RBM is a simplification of the Boltzmann Machine models that have the energy formula of joint configuration {v, h} as follows:

$$E(v,h) = -\sum_{i=1}^{gv}\sum_{j=1}^{gh} W_{ij}\, v_i h_j - \sum_{i=1}^{gv} a_i v_i - \sum_{j=1}^{gh} b_j h_j$$

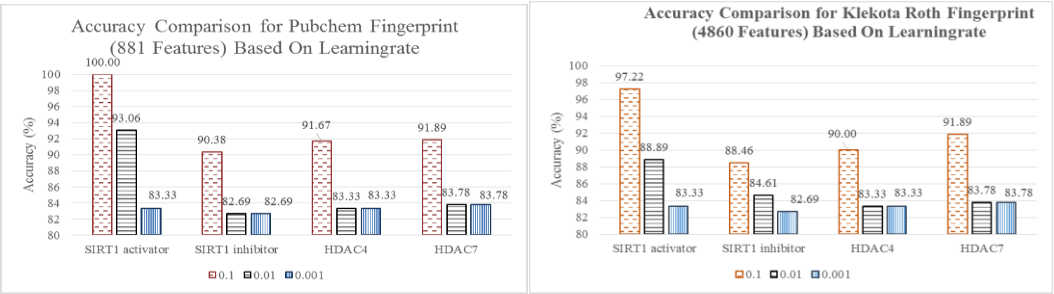
$W_{ij}$ represents the weight of the interaction between the visible units $i$ and hidden units $j$, while $b_i$ and $a_j$ is biased to the hidden units and visible units.

After doing the pre-training data using RBM, the values of the hidden units derived from data can be used as input data for pre-training RBM at the next layer. At the end of layer, the classification function is inserted. This research is using sigmoid function for the classification layer. Training for DBN were done by using Deepnet library for R pogramming.

# 5. Experiments Result

## 5.1. Experiments Result for Literature Data

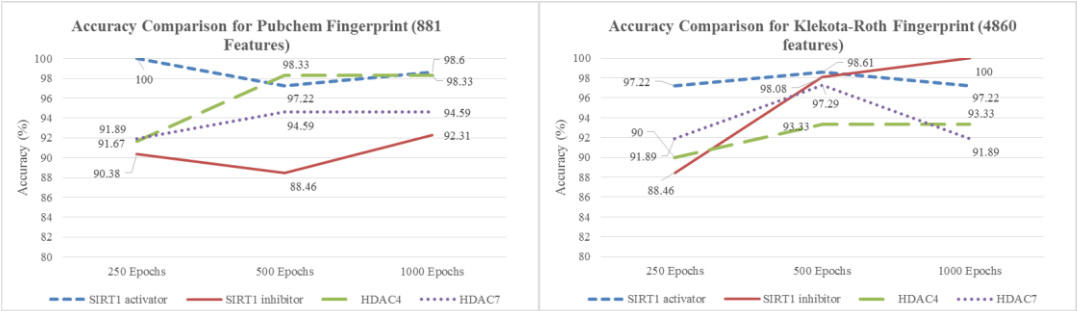
In this section, we use 2 RBM layer for pre-training and 1 output layer for classification. First, we compare the DBN architecture based on learning rate. The comparison of testing accuracy from 25% data between learning rate can be seen in Figure 3. We use 250 training epochs, 440 hidden units for PubChem fingerprint and 2430 hidden units for Klekota-Roth fingerprint to retrieve the best learning rate for the architecture.



(A) (B)

Fig. 3: (A) Accuraccy comparison based on learningrate for PubChem fingerprint feature (B) Accuraccy comparison based on learningrate for Klekota-Roth fingerprint feature.

Can be seen from Figure 3, the best accuracy is obtained from learning rate 0.1. Next step we improve the number of training epochs by 250, 500, and 1000 to find the better model for classification. The comparison for number of training epochs can be seen at Figure 4.



(A) (B)

Fig. 4: (A) Accuraccy comparison based on epochs for PubChem fingerprint feature (B) Accuraccy comparison based on epochs for Klekota-Roth fingerprint feature.

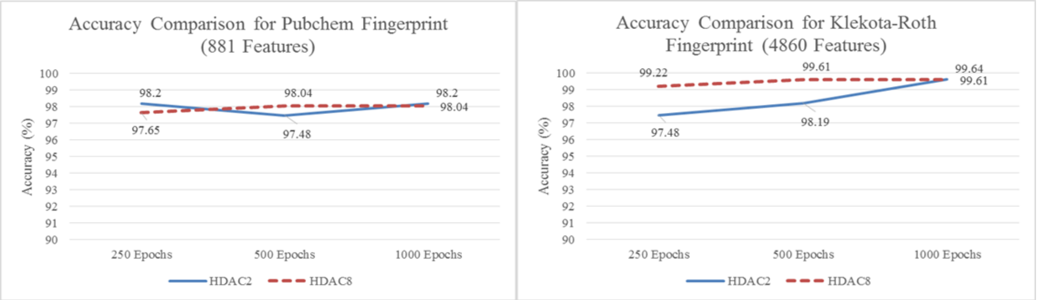
From Figure 4, there is a peculiar mixture between increasing and decreasing of accuracy between epochs. From the average of accuracy between the data, we got for PubChem fingerprints as follows: 93.48% (250 epochs), 94.65% (500 epochs), and 95.97% (1000 epochs). And for Klekota-Roth fingerprints as follows: 91.89% (250 epochs), 96.83% (500 epochs), and 95.61% (1000 epochs). We also tried 100 training epochs, but it produces a poor sensitivity rate, so using epochs under 100 is not recommended. This leads to the next step, the result will be compared with the existing research which is using LigandScout 4.0 for pharmacophore modeling [11]. The comparison results can be seen at Table 2. The comparison result showed that our method outperform the pharmacophore method for literature data.

Table 2. Comparison between our method and existing method

| Protein Target | PubChem Fingerprint Feature<br>(1000 epochs, 0.1 learning rate) | | | Klekota-Roth Fingerprint Feature<br>(500 epochs, 0.1 learning rate) | | | Existing Research [11]<br>(Pharmacophore Modeling) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Accuracy (%) | Sensitivity | Specificity | Accuracy (%) | Sensitivity | Specificity | Accuracy (%) | Sensitivity | Specificity |
| SIRT1 activator | **98.61** | 0.917 | 1.000 | 98.61 | 0.917 | 1.000 | 90.81 | 0.800 | 0.910 |
| SIRT1 inhibitor | 92.31 | 1.000 | 0.907 | **98.08** | 1.000 | 0.977 | 73.21 | 0.591 | 0.735 |
| HDAC4 | **98.33** | 0.900 | 1.000 | 93.33 | 0.800 | 0.960 | 97.93 | 0.539 | 0.990 |
| HDAC7 | 94.39 | 0.667 | 1.000 | 97.29 | 0.833 | 1.000 | **98.21** | 0.308 | 0.992 |
| **Average** | 95.97±3.04 | 0.853±0.014 | 0.977±0.046 | **96.83±2.39** | 0.887±0.089 | 0.984±0.019 | 90.04±11.73 | 0.559±0.202 | 0.907±0.121 |

## 5.2. Experiment Result for DUD-E Docking Data

In this experiment, we use the same configuration as Section 5.1. First, the comparison of DBN architecture based on feature and number of epochs with learning rate 0.1. The result can be seen at Figure 5. For comparison, the ROC was computed to find the AUC for both protein. Then, the best result which achieved from Klekota-Roth feature with 1000 epochs was compared with those AUC from DUD-E docking. Details of the results can be seen in Table 3 and Table 4.



(A) (B)

Fig. 5: (A) Accuraccy comparison based on epochs for PubChem fingerprint feature (B) Accuraccy comparison based on epochs for Klekota-Roth fingerprint feature.

Table 3. Result comparison between features (using 1000 epochs)

| Protein target | PubChem Fingerprint Feature | | | Klekota-Roth Fingerprint Feature | | |
|---|---|---|---|---|---|---|
| | Accuracy (%) | Sensitivity | Specificity | Accuracy (%) | Sensitivity | Specificity |
| HDAC 2 | 98.20 | 0.935 | 0.996 | 99.64 | 0.978 | 1.000 |
| HDAC 8 | 98.04 | 0.884 | 1.000 | 99.61 | 0.977 | 1.000 |
| **Average** | 98.12±0.11 | 0.909±0.036 | 0.998±0.003 | 99.62±0.02 | 0.977±0.001 | 1.000±0.000 |

Table 4. AUC comparison between DBN method and DUD-E

| Protein target | DBN Virtual Screening<br>AUC (%) | DUD-E Docking<br>AUC [12] (%) |
|---|---|---|
| HDAC 2 | 100 | 76.59 |
| HDAC 8 | 98.71 | 79.91 |

## 6. Conclusion

This research has successfully compared the performance of virtual screening between DBN method and the pharmacophore method. The accuracy of DBN methods were significantly increased after increasing the training epochs. DBN method obtained the best accuracy among 97% until 99%. This accuracy is slightly higher than the previous research. The DBN method also compared with DUD-E docking data. The result shows that the AUC from DBN method is higher than the DUD-E method. Finally, DBN method can be used as an alternative way for screening the drug protein target.

## 7. References

[1] J. P. Overington, B. Al-Lazikani, A. L. Hopkins. 2006. How many drug targets are there?.*Nature Review Drug Discovery*. 2006, 5, pp. 993-996.

[2] E. E. Bolton, Y. Wang, P. A. Thiessen, *et al.* PubChem: integrated platform of small molecules and biological activities, *Annual Reports in Computational Chemistry*. 2008, vol. 4, pp. 217–241.

[3] V. K. Vyas, A. Goel, M. Ghate, P. Patel. Ligand and structure-based approaches for the identification of SIRT1 activators. *Chemico-Biological Interaction*. 2015, 228(2015), pp. 9-17.

[4] F. Cheng, C. Liu, J. Jiang, *et al*., Prediction of drug-target interactions and drug repositioning via network-based inference, *PLoS Comput. Biol*. 2012, 8(5):p.e1002503.

[5] S. K. Dhanda, D. Singla, A. K. Mondal, G. P. S. Raghava. DrugMint: a webserver for predicting of drug-like molecules. *Biology Direct*. 2013, 8:28.

[6] J. Y. Shi, S. M. Yiu, Y. Li, *et al.* Predicting drug-target interaction for new drugs using enhanced similarity measures and super-target clustering, *Bioinformatics and Biomedicine (BIBM)*. 2014, pp 45-50.

[7] Y. Liu. Machine learning for drug design. *International Journal of Computer and Information Technology*. 2015, 4(1).

[8] A. M. Johnson, G. M. Maggiora. *Concepts and Applications of Molecular Similarity*. New York: John Willey&Sons. 1990. ISBN 0-471-62175-7.

[9] G. E. Hinton. Learning multiple layers of representation. *Trends in Cognitive Sciences*. 2007, 11(10): pp.428–434.

[10] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*. 2002, 14(8): pp.1771-1800.

[11] A. Yanuar, Azminah, L. Erlina, Andhika. [Drug compounds for NAD protein family]. *Unpublished Raw Data.* 2016.

[12] P. Belenky, K. L. Bogan, C. Brenner. NAD+ metabolism in health and disease. *Trends Biochem. Sci.* 2006, 32 (1): pp. 12–9. doi:10.1016/j.tibs.2006.11.006. PMID 17161604.

[13] M. M. Mysinger, M. Carchia, J. J. Irwin, B. K. Shoichet. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*. 2012, 55(14): pp. 6582-6594. DOI: 10.1021/jm300687e

[14] J. Klekota, and F. P. Roth. Chemical Substructures that Enrich for Biological Activity. *Oxford Journal, Bioinformatics*. 2008, 24(21):2518-25. doi: 10.1093/bioinformatics/btn479.