Malaysian Journal of Chemistry, 2023, Vol. 25(5), 168-174
(Special Issue: *The 4th International Conference on Chemistry and Material Sciences 2022*)

Development of Machine Learning-Based QSAR Model for Virtual Screening of Dipeptidyl Peptidase-4 Inhibitors

Salsabila Safa, A. B. S., Sudarko*, Zulfikar, Anak Agung Istri Ratnadewi and Wuryanti Handayani
Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Jember, Jl. Kalimantan 37, Jember, 68121, Indonesia
*Corresponding author (e-mail: darko@unej.ac.id)

Treatment of type 2 diabetes mellitus is mostly done by inhibiting the DPP-4 protein using an inhibitor compound, however, it may cause headaches and indigestion as its side effect. This study has been focused on the development of the DPP-4 inhibitor as a new drug candidate for type 2 diabetes mellitus using the Machine Learning-based Quantitative Structure-Activity Relationship (QSAR) for the virtual screening process. Training dataset has been obtained from the ChEMBL database with DPP-4 as the target protein (code ChEMBL284), and it is used to find a model which then applied for the virtual screening process of 884 million molecules obtained from the ZINC database. The screening processes are based on the predicted activity (pIC50) values above the experimental activity values of the drugs that were already available and it is then screened again according to Lipinski Rule of 5 to find out the compounds that can be absorbed by the body. The compounds that can be absorbed by the body were then docked using AutoDockVina software to determine the free energy value and interaction pattern between the compound and protein target to get recommendations for a new DPP-4 inhibitor candidate. Result obtained from the best model with an R^2 test value of 0.69 is then used for virtual screening. The results of the virtual screening were 5 compounds that had the highest pIC50 values and not violating Lipinski's Rule of 5. These compounds had codes ZINC341837061, ZINC001359979988, ZINC001707862778, ZINC001722886251 and ZINC001726358542.

Keywords: Diabetes Mellitus Type 2; dipeptidyl peptidase-4; Machine Learning; Quantitative Structure-Activity Relationship (QSAR)

Received: September 2022; Accepted: December 2022

Diabetes mellitus is a type of disease caused by high levels of glucose in the blood that exceed normal conditions (DeFronzo, 2009). Diabetes mellitus is divided into two types, i.e. diabetes mellitus type 1 and diabetes mellitus type 2. The cause of diabetes mellitus type 1 is the inability of the pancreas to produce insulin or autoimmune disease. Diabetes mellitus type 2 is caused by the body's inability to use insulin effectively or insulin resistance (Stenhouwer and Schaper, 2009).

The use of metformin as an antidiabetic drug has a small hypoglycemic effect but has a gastro-intestinal effect of quite high >10% (Gumantara and Oktarina, 2017). The side effects of metformin such as nausea, vomiting, bloating, and diarrhea. These symptoms can be removed by lowering the dose of the drug. Side effects that occur related to the total dose given and increasing the dose too fast. This can be overcome by increasing the dose by 500 mg per day every 2 to 3 weeks, then 3 to 7 weeks until effective dose is achieved (Andayani et al., 2009).

The protein involved with type 2 diabetes mellitus is dipeptidyl peptidase-4 (DPP-4). Treatment of type 2 diabetes mellitus can be done by inhibiting the DPP-4 protein using an inhibitor. DPP-4 or CD26 is a type of membrane peptidase that has 766 amino acids and is widely distributed in many tissues. It is constitutively expressed on epithelial and endothelial cells of a variety of different tissues, for example, the intestine, liver, lung, kidney, and placenta (Havale and Pal, 2009).

DPP-4 is a type of serine peptide present in the human body which is responsible for the decreased activity of the 2 main incretin hormones. There are two types of incretin hormones in the body, namely glucagon-like peptide-1 (GLP-1) and glucose-dependent insulinotropic (GIP). Incretin hormones are defined as gut-related hormones that are released in response to the ingestion of nutrients (food). The inhibition of the incretin hormone has an impact on its activity which is inhibited in the process of insulin excretion. GLP-1 is a type of incretin hormone secreted by L-cells in the distal small intestine in response to food intake, namely oral nutrition (Holst *et al*, 1987). The target of therapy for type 2 diabetes mellitus is to increase the activity of GLP-1 and GIP (Lovshin and Ducker, 2009).

†Paper presented at the 4th International Conference on Chemistry and Material Sciences 2022 (IC2MS2022)

Machine learning is a method of studying data that produces a model using computer assistance (Putra, 2019). This study develops a model using the Machine Learning-based Quantitative Structure-Activity Relationship (QSAR) method which is used for the virtual screening process for new drug candidates for diabetes mellitus type 2. Virtual screening is an in-silico method that can be screened various compounds to find potential ligands for appropriate drug target. The virtual screening method is divided into ligand-based virtual screening and structure-based virtual screening (Carpenter and Huang, 2018).

QSAR is a method that processes data by representing the 2D structure of compounds with complex features (Carpenter and Huang, 2018). QSAR (Quantitative Structure and Activity Relationship) is modeling that requires an in-silico approach with a quantitative relationship between molecular structure and activity (Ekins *et al*, 2007).

The molecular fingerprint is a representation of chemical structure for design in chemical database substructure search but then used for analytical tasks, such as: similarity search, grouping, and classification (Todeschini and Consonni, 2000). Molecular finger-printing enables the prospective prediction of biological activity and properties of compounds relevant to drug development through the use of quantitative activity structures in a model (Muegge and Mukherje, 2016).

Extended-connectivity fingerprints (ECFP6) is a method by which molecular fingerprints are explicitly used to capture molecular features relevant to molecular activity. The topological representation of the ECFP is derived from the refinement process of Morgan's Algorithm and usually folds into a fixed size such as 1024, 2048 or 4096 for further use for predictive modeling tasks (Xu et al, 2017).

Lipinski's rule of 5 is used to determine which compounds could be absorbed by the body (Lipinski et al, 2001). The criteria or parameters of Lipinski's rule of 5 include weight molecule (BM) less than 500 g/mol, octanol/water partition coefficient value (log P) which counts less than 5, donor hydrogen bonds (-NH-, -OH-) are less than 5, the number of hydrogen bond acceptors (-N=, -O-) is less than 10 (Lipinski et al, 2001; Syahputra *et al.*, 2014; Widiandani *et al.*, 2013).

Docking molecular is a modeling technique that is used to predict how a protein (enzyme) interacts with small molecules (ligands) (Pinzi and Rastelli., 2019). The accuracy of the results obtained from the docking process must be carried out to determine the position regarding the conformation of the enzyme and ligand or known as the Root Mean Deviation Square (RMSD) by comparing the positions of the ligand atoms experimentally (Noviardi *et a.l*, 2015).

Research Methods

The process of searching for new drug candidates begins with the retrieval of a dataset in the ChEMBL database of the DPP-4 target protein with the code ChEMBL284. The dataset contains compounds that have been tested for their activity against DPP-4 in the form of SMILES and pIC50 values. The dataset is used for the process of finding the best model.

The best model is then used to predict the activity values of 884 million molecules in the ZINC database. Molecules with the best activity values were screened again using Lipinski's rule of 5 to determine which compounds could be absorbed by the body. Compounds that can be absorbed by the body are then docked using AutoDockVina software to determine the value of the free bond energy between the compound and the target protein. The final step in this research is to observe the interaction between the ligand and the protein (enzyme) with the help of PyMOL and Biovia Discovery Studio (BDS).

Tools and Materials

This research uses a computer with a Linux operating system and the Python version 3.9 equipped with the PyTorch module (PyTorch, n.d.), RDKIT, Ray-Tune (Ray Tune - Fast and Easy Distributed Hyperparameter Tuning, n.d.) (Bento *et al*, 2020) Pandas, NumPy, sklearn. The AutoDockTools 1.5.6 application is used for the preparation of ligands and macromolecule, the PyRx application is used for the docking process, the PyMOL application, and Biovia Discovery Studio for the molecular visualization process. The material used in this study using the material in the form of a protein target for the enzyme DDP-4 downloaded from RCSB Protein Data Bank with code (PDB ID: 2ONC). The ligand structure of the virtual screening results is down-loaded from the ZINC20 database.

METHODS

Model Search

The activity/pIC50 values of the target protein DPP-4 with the code ChEMBL284 was downloaded from the ChEMBL database. A total of 4138 datasets are divided into training data and testing data. The optimization process on the QSAR model is carried out by running 375 jobs from the combination of neural network hyperparameters used in the machine learning training process. Hyperparameter of neural network consists of size of fully connected layer 1 (FC1), size of fully connected layer 2 (FC2), size of fully connected layer 3 (FC3), drop out rate (DO), learning rate (LR) and batch size (BS). FC1, FC2 and FC3 are varied from 64 to 2048, BS is varied from 64 to 256.

The result of the optimization process is in the form of the best model which will be used for the activity value prediction process (pIC50).

Virtual Screening with ZINC Database

Virtual screening was carried out by ranking predicted pIC50 values of 884 million compounds from the ZINC20 database (Irwin *et al*, 2020). Molecules, with predicted activity values higher than averaged drug activity values, were screened again using Lipinski's rule of 5. Virtual screening with Lipinski's rule of 5 aims to determine the body's ability to absorb these molecules.

Molecular Docking

The next treatment is the molecular docking process. The target protein was downloaded from the RCSB Protein Data Bank with PDB ID 2ONC (Dipeptidyl peptidase-4) in pdb format. The ligands contained in the molecule were removed using Biovia Discovery Studio and then the molecular format was changed to pdbqt format using AutodockTools.

Structure of molecules with high predicted activity values are downloaded on the ZINC20 database page with sdf format, then converted into pdbqt format using AutodockTools. The target protein with pdbqt format is then docked with the molecules resulting from the screening process. The analysis was carried out by taking 5 molecules with the smallest free energy values in the docking process and then proceeding with the analysis process using the PyMOL.

The purpose of the analysis process is to determine the location of the active site and the interaction of proteins that bind to prospective ligands or prospective inhibitors of new drug candidates for type 2 diabetes mellitus. The activity score of the docking process shows the activity of the DPP-4 complex with the ligand then sorted based on Gibbs free energy to get the recommendation for DPP-4 inhibitors as a candidate for drug of diabetes mellitus type 2.

RESULTS AND DISCUSSION

Model Search

The combination of hyperparameter neural network for QSAR optimization consists of 375 different models. These models are tested and optimized with Ray-Tune to get result with minimum value R2test. The best model is found to have hyperparameter as follows: FC1(512), FC2(128), FC3(128), DO(0.4), BS(64) and LR(0.001). $R^2_{training}$, $R^2_{validation}$ and R^2_{test} values for the best model are 0.79, 0.60 and 0.69 respectively.

Virtual Screening with ZINC Database

The best model was used for virtual screening of 884 million compounds from the ZINC database and 1550 molecules were found to have predicted activity above the experimental activity values of the drug. The available drug for diabetes mellitus type 2 is Saxagliptin, Vildagliptin, Linagliptin, Alogliptin, and Sitagliptin (Mantzoros and Christos, 2018).

Screening with Lipinski's Rule of 5

The result of virtual screening of molecules of the ZINC database is then screened with Lipinksi's rule of 5. Lipinski's rule of 5 screening results which are zero aberrations are compounds that do not violate Lipinski's rule of 5 and have the possibility of being absorbed by the body. 1501 molecules are found not violating Lipinski's rule of 5.

Molecular Docking

DPP-4 protein with PDB code 2ONC was downloaded from the Protein Data Bank. These proteins or macro-molecules have 4 types of subunits different, namely the A chain, B chain, C chain, and D chain. Preparation of protein using the Biovia Discovery Studio, namely by separating the A chain from the B chain, C chain, and D chain, removing water molecules (H_2O), as well as other unused molecules. The ligands are separated because they are bound to the active site which can prevent other ligands from binding. A water molecule (H_2O) is removed so as not to disturb the bond during the docking process in the form of binding of the ligand to water molecules through hydrogen bonds. The prepared structure is then saved in pdbqt format (Nurfitriyana, 2010). Protein after preparation can be seen in Figure 1.

Molecular docking results in the form of Gibbs free energy and RMSD value. Gibbs free energy shows the energy of the interaction between the ligand and the protein. Interaction between ligand and protein calculated with molecular docking can be seen on Figure 2.

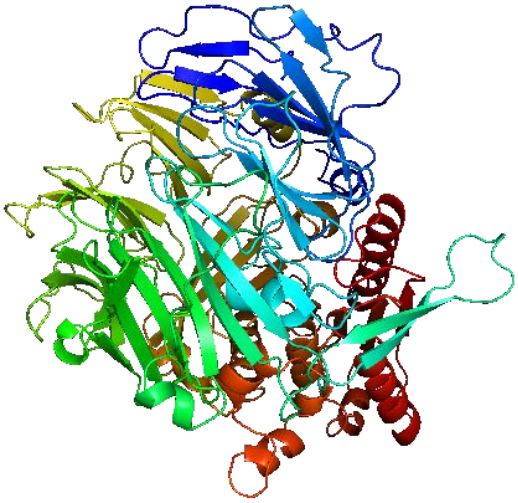

Figure 1. Protein DPP-4 after preparation.

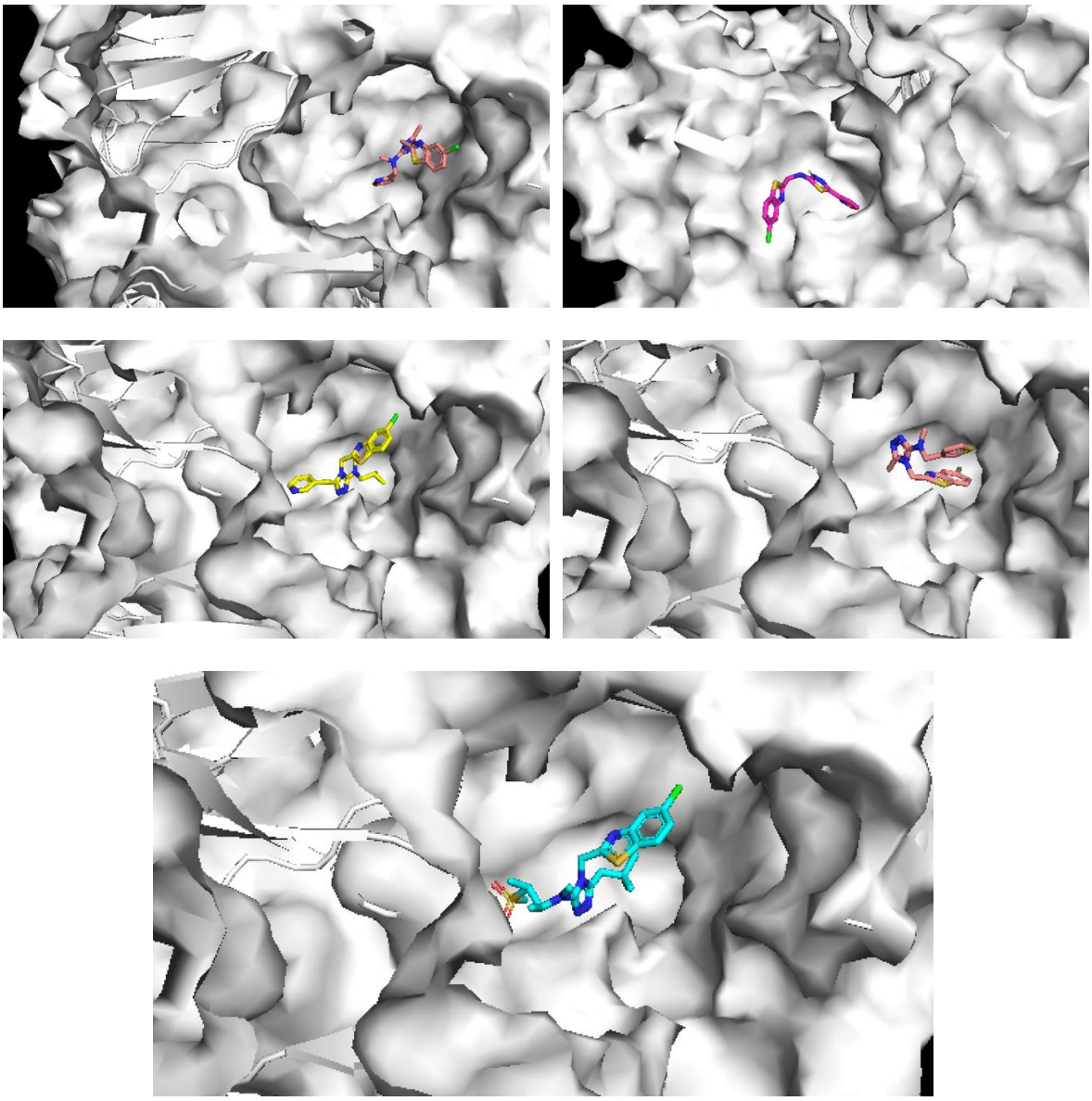

Figure 2. Interaction of ligand and protein.

Tabel 1. Result of Molecular Docking.

Ligand	Binding Affinity (kcal/mole)	RMSD (Å)
ZINC001359979988	-9.4	0
ZINC001722886251	-8.3	2.8
ZINC000341837061	-8.2	4.6
ZINC001707862778	-7.8	1.8
ZINC001726358542	-7.4	2.4

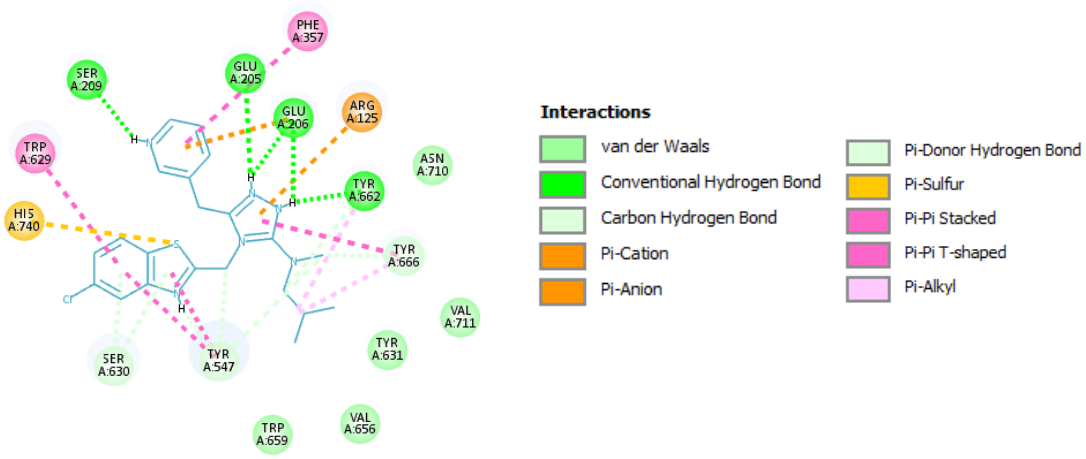

Figure 3. Visualization of Docking DPP-4.

Table 1 shows docking results of 5 ligands interact with the DPP-4 protein determined by the value of the Gibbs free energy. The entire Gibbs free energy value of the test was negative (<0) for the DPP-4 protein. This indicates that all the test ligands have the same binding affinity to protein (Muttaqin et al., 2019). The lowest Gibbs free energy value is ZINC001359979988 at -9.4 kcal/mole. Meanwhile, the ZINC001726358542 ligand has the highest free energy value high is -7.4 kcal/mol. The smaller (negative) free energy indicates the greater the energy of the interaction that occurs between the protein and the ligand (Frimayanti *et al.*, 2018).

Based on figure 3, the interaction of the ligand with proteins are hydrogen bonds and carbon-hydrogen bonds. Hydrogen bonds interact with the residues Ser 209, Glu 205, Glu 206, and Tyr 662. Carbon-hydrogen bonds interact on the residues Ser 630, Tyr 547, and Tyr 666. Amino acid is non-polar namely Val 656 and Val 711, where non-polar amino acid residues tend to form interactions hydrophobic in protein. The presence of hydrophobic interactions can minimize the interaction of non-polar residues with water. The molecular docking result that has the best potential is ZINC001359979988 with the lowest Gibbs free energy of -9.4 kcal/mole.

CONCLUSION

The optimization model with 375 combinations of hyperparameters obtained the best model with an R^2 test value is 0.69. The best model was then used for virtual screening of 884 millions of molecules from ZINC database, where it was found that 1501 molecules with predicted activity values above the experimental activity values of the drugs and not violating Lipinski's rule of 5. 5 Molecules with lowest Gibbs free energy obtained from molecular docking analysis are ZINC000341837061, ZINC001359979988, ZINC001707862778, ZINC001722886251, and ZINC-001726358542. These 5 molecules can be a new potential candidate for DPP-4 inhibitor.

REFERENCES

1. Andayani, T. M., Ibrahim, M. I. M. and Asdie, A. H. (2009) Pengaruh kombinasi terapi sulfonilurea, metformin, dan acarbose pada pasien diabetes mellitus tipe 2. *Majalah Farmasi Indonesia*, **20(4)**, 224–230.

2. Bento, A. P., Hersey, A., Félix, E., Landrum, G., Gaulton, A., Atkinson, F., Bellis, L. J., De Veij, M. and Leach, A. R. (2020) An open source

chemical structure curation pipeline using RDKit. *Journal of Cheminformatics*, **12(1)**, 1–16. https://doi.org/10.1186/s13321-020-00456-1.

3. Carpenter, K. A. and Huang, X. (2018) Machine Learning-based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: A Review. *Current Pharmaceutical Design*, **24(28)**, 3347–3358. https://doi.org/10.2174/1381612824666180607124038.

4. Defronzo, R. A. (2009) From the triumvirate to the ominous octet: A new paradigm for the treatment of type 2 diabetes mellitus. *Diabetes*, **58(4)**, 773–795. https://doi.org/10.2337/db09-9028.

5. Ekins, S. and Mestres, J. (2007) In silico pharmacology for drug discovery: applications to targets and beyond. *Br J Pharmacol*., **152(1)**, 21–37. https://doi.org/10.1038/sj.bjp.0707306.

6. Frimayanti, N., Mora, E. and Anugrah, R. (2018) Study of Molecular Docking of Chalcone Analoque Compound as Inhibitors for Liver Cancer Cells HepG2. *Computer Engineering and Applications Journal*, **7(2)**, 147–158. https://doi.org/10.18495/comengapp.v7i2.260.

7. Gumantara, M. P. B. and Oktarina, R. Z. (2017) Perbandingan Monoterapi dan Kombinasi Terapi Sulfonilurea-Metformin terhadap Pasien Diabetes Melitus Tipe 2. *Majority*, **6(1)**, 55–59.

8. Havale, S. and Pal, M. (2009) Medicinal chemistry approaches to the inhibition of dipeptidyl peptidase-4 for the treatment of type 2 diabetes. *Bioorg Med Chem*., **17(5)**, 1783–1802. https://doi.org/10. 1016/j.bmc.2009.01.061.

9. Holst, J., Orksov, C., Nielsen, O. and Schwartz, T. (1987) Truncated glucagon-like peptide I, an insulin-releasing hormone from the distal gut. *FEBS Let*., **211(2)**, 169–174. https://doi.org/10. 1016/0014-5793(87)81430-8.

10. Irwin, J. J., Tang, K. G., Young, J., Dandarchuluun, C., Wong, B. R., Khurelbaatar, M., Moroz, Y. S., Mayfield, J. and Sayle, R. A. (2020) *ZINC20□A Free Ultralarge-Scale Chemical Database for Ligand Discovery*. https://doi.org/10.1021/acs.jcim.0c00675.

11. Lipinski, C. A., Lombardo, F., Dominy, B. W. and Feeney, P. J. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, **46(1–3)**, 3–26. https://doi.org/10.1016/S0169-409X(00)00129-0.

12. Lovshin, A. and Ducker, J. (2009) Incretin-based therapies for type 2 diabetes mellitus. *Nat Rev. Endocrinol*., **5(5)**, 262–269. https://doi.org/10.1038 /nrendo.2009.48.

13. Mantzoros and Christos (2018) Insulin resistance: Definition and clinical spectrum - UpToDate. *UpToDate*, 1–17.

14. Muegge, I. & Mukherjee, P. (2016) An overview of molecular fingerprint similarity search in virtual screening. *Expert Opinion on Drug Discovery*, **11(2)**, 137–148. https://doi.org/10.1517/17460441.2016.1117070.

15. Muttaqin, F. Z., Ismail, H. and Muhammad, H. N. (2019) Studi Molecular Docking, Molecular Dynamic, dan Prediksi Toksisitas Senyawa Turunan Alkaloid Naftiridin sebagai Inhibitor Protein Kasein Kinase 2-α pada Kanker Leukemia. *Pharmacoscripts*, **2(1)**, 49–64.

16. Nurfitriyana, F. (2010) Penambatan Molekuler beberapa Senyawa Xanton dari Tanaman *Garcinia magostana Linn*. Pada Protease HIV-1. *Skripsi, Depok: Program Studi Farmasi: Universitas Indonesia*.

17. Noviardi, H., Studi Farmasi, P. & Tinggi Teknologi Industri dan Farmasi Bogor, S. (2015) Potensi senyawa bullatalisin sebagai inhibitor protein leukotrien A4 hidrolase pada kanker kolon secara in silico. *Fitofarmaka: Jurnal Ilmiah Farmasi*, **5(2)**, 65–73. https://doi.org/10.33751/JF.V5I2.410.

18. Pinzi, L. & Rastelli, G. (2019) Molecular Docking: Shifting Paradigms in Drug Discovery. *International Journal of Molecular Sciences*, **20(18)**. https://doi.org/10.3390/IJMS20184331'\.

19. Putra, J. W. G. (2019) Pengenalan Konsep Pembelajaran Mesin dan Deep Learning Edisi 1,4. *Tokyo: Departemen of Computer Science, Tokyo Institute of Technology.*

20. Protein Target DPP-4 (2ONC) (2022) Diakses pada 1 September 2022 dari PDB: https://www.rcsb.org/

21. Stenhouwer, C and Schaper, N. (2009) Therapeutic strategies in diabetes in *Cardiology Clinics*, **23**, 2. *Clinical Publishing*.

22. Syahputra, G., Ambarsari, L. and Sumaryada, T. (2014) Simulasi Docking Kurkumin Enol, Bisdemetoksikurkumin dan Analognya sebagai Inhibitor Enzim12-Lipoksigenase. *Jurnal Fisika*, **10(1)**, 55–67.

23. Todeschini, R. and Consonni, V. (2000) Handbook of Molecular Descriptors. *Wiley-VCH: Weinheim, Germany*.

24. Widiandani, T., Siswandono, S. Hardjono, R. Sondakh, Istifada and Zahra, R. (2013) Docking dan Modifikasi Struktur Senyawa Baru Turunan Parasetamol. *Berkala Ilmiah Kimia Farmasi*, **2(1)**, 41–45.

25. Xu, Z., Wang, S., Zhu, F. and Huang, J. (2017) Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery. *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology*, 285–294. https://doi.org/10.1145/3107411.3107424.