Development of Molecular Autoencoders as Generators of Protein Inhibitors: Application for Prediction of Potential Drugs Against Coronavirus SARS-CoV-2

Mikita Shuldau
United Institute of Informatics
Problems of NAS of Belarus
Minsk, Belarus
nickshuldov29@gmail.com

Artsemi Yushkevich
United Institute of Informatics
Problems of NAS of Belarus
Minsk, Belarus
artsemi.yushkevich@gmail.com

Ivan Bosko
United Institute of Informatics
Problems of NAS of Belarus
Minsk, Belarus
vanya_384@mail.ru

Alexander Tuzikov
United Institute of Informatics Problems
of NAS of Belarus
Minsk, Belarus
tuzikov@newman.bas-net.by

Alexander Andrianov
Institute of Bioorganic Chemistry
of NAS of Belarus
Minsk, Belarus
alexande.andriano@yandex.ru

Abstract. **A generative autoencoder for the rational design of potential inhibitors of the SARS-CoV-2 main protease able to block the catalytic site of this functionally important viral enzyme was developed.**

Keywords: **SARS-CoV-2, X77, main protease, deep learning, generative autoencoder, semi-supervised learning, virtual screening, molecular docking, binding free energy calculations, anti-SARS-CoV-2 drugs**

I. Introduction

To date, computer-aided drug design has become an important tool allowing one to significantly reduce the time and costs required for developing novel therapeutic agents. In recent years, computer-assisted mathematical and statistical models, such as machine learning, are increasingly being used for drug design and discovery. Despite these methods becoming more common in chemoinformatics, their potential in this field is yet to be revealed.

Generative models have proven to be promising in tasks of text [1] and image [2] generation, including generation of medical images like X-ray ones. Despite the traditional similarity-based virtual screening of chemical databases, such as PubChem (https://pubchem.ncbi.nlm.nih.gov/) [3], provide wide possibilities for identification of novel potential drugs, it has certain disadvantages compared to generative statistical models. One of the major incentives to use generative models is a better exploration of a molecular feature space. Similarity-based search provides exploration of focused chemical space, limited by search space diversity of compounds at disposal, while generative statistical models allow one to cover molecular feature space of much wider chemical diversity. The second reason for generative model superiority is conditional sampling. Generation of new molecules from a chemical space is not the only option: predicted binding free energy could be used as an additional dimension, which allows one to generate molecules from a subset of investigated chemical space with a preset binding affinity.

This study is devoted to the development of the generative autoencoder based on a linear molecule representation in the Simplified Molecular Input Line Entry System (SMILES) format [4]. One of the core ideas behind using SMILES, or more precisely, vectorized SMILES for model training was the recovery capabilities of such data. As was shown in a study [5], generative models provide a decent ground for screening results enrichment, however use of descriptors like fingerprints may complicate the recovery of the chemical structures themselves. In contrast, SMILES-based embeddings are supposed to be a good alternative to using fingerprints in deep learning chemoinformatics approaches when the ability to restore the structure of chemical compounds is important. That is why the SMILES embeddings were chosen as the architectural basement for the constructed generative autoencoder to generate potential inhibitors of the selected protein target.

II. Materials and Methods

A. Training Set Preparation

The developed generative autoencoder is built to be specific for a target protein, and, therefore, the training

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

dataset should include compounds potentially active against the selected protein. As noted before, the autoencoder developed here was adopted and applied for generation of potential anti-SARS-CoV-2 inhibitors, and, accordingly, a virtual compound library of potential anti-SARS-CoV-2 agents was formed for preparation of a training dataset. The preparation procedure of this molecular library was as follows:

a) Pharmacophore-based Virtual Screening: To identify small-molecule compounds potentially active against SARS-CoV-2 main protease (M^{pro}), the pharmacophore-based virtual screening was performed using the Pharmit server software (http://pharmit.csb.pitt.edu/) [6]. Seventeen pharmacophore models were built based on 6 peptidomimetics and 10 small-molecule inhibitors of SARS-CoV reported in a study [7], using web-server PharmaGist [8]. Virtual screening was performed in the nine Pharmit molecular libraries containing over 213.5 million chemical structures, resulting in a set of 711102 compounds that satisfied one of the seventeen constructed pharmacophore models. The Pubchem API wrapped in Python 3 (https://www.python.org/) module PubChemPy (https://pubchempy.readthedocs.io/) was used to additionally enrich the screened dataset with potential inhibitors based on the selected PubChem compounds by the similarity search with a Tanimoto similarity coefficient of 0.8.

b) Molecular Docking: Compounds identified by the pharmacophore-based virtual screening and PubChem similarity search were subject to the preliminary molecular docking with the unliganded SARS-CoV-2 M^{pro} structure. The compounds were then filtered based on the values of the docking scoring function with the threshold of –7 kcal/mol, which corresponds to the standard activity threshold of 10 μM commonly used in vitro screening. The dataset of 353467 potentially active compounds was subject to the refining molecular docking with the unliganded SARS-CoV-2 M^{pro} structure. Analysis of the distribution of scoring function values after the refining docking resulted in the filtration of successfully docked compounds above the selected threshold of –6 kcal/mol.

c) SMILES space revision and vectorization: Based on a linear SMILES notation, the dataset of selected compounds was cleared from those containing non-recognizable atoms, non-abundant isotopes, other than druglike (H, C, N, O, P, S, F, Cl, Br, I) atoms or those which molecular weight was above the selected threshold of 1000 Da. Structure representations of the prepared compounds in the linear notation SMILES were obtained by Python 3 using the RDKit (http://www.rdkit.org/) module which was also used previously for the described dataset cleaning. Based on the frequency distribution of SMILES elements in the prepared dataset, compounds possessing at least one SMILES element with frequency less than 0.001 were filtered out. Finally, distribution of SMILES lengths was investigated and compounds with SMILES representation longer than 120 characters were eliminated. After all the filters applied, the dataset consisted of 342102 distinct ligands and their corresponding SMILES. The SMILES were vectorized into a matrix according to the maximum length and symbols vocabulary size, with the added start and end symbols represented by "!" and "E".

The obtained 342102 compounds combined with the corresponding values of molecular docking scoring function formed the dataset which was split into the training, validation, and test sets comprising 70%, 15% and 15% of the original dataset, respectively. When forming the subsets, a stratified split was used to preserve equal energy distributions within all 3 sets. The validation set was used to evaluate the model's ability to reconstruct the input SMILES during training, while test SMILES were used to sample new compounds from by adding distortion to their latent representation. Thus, the corresponding datasets for model training, validation and generation of new molecules were prepared.

B. 3D Structures Generation for Generated Molecules

To evaluate the ability of deep learning model to generate novel compounds active towards the target protein, molecular docking of these molecules should be performed. In doing so, 3D structures of the generated molecules are required. To obtain these 3D structures from a linear notation SMILES, a script was developed in Python 3 using the RDKit module. The generation pipeline included the following steps: SMILES input, SMILES validity check, 2D coordinates generation, 3D coordinates generation, optimization of the structure in the MMFF94 force field, addition of hydrogen atoms, and re-optimization in the MMFF94 force field. Generation of 3D coordinates was performed using the ETKDGv3 [9] algorithm.

C. Molecular Docking

a) Preparation of Protein Structure: The crystal structure of the unliganded SARS-CoV-2 M^{pro} was taken from the Protein Data Bank (PDB ID: 6Y84; https://www.rcsb.org/pdb/). This SARS-CoV-2 M^{pro} structure was prepared by adding hydrogen atoms and annotating atoms with partial charges by Gasteiger scheme [10] followed by the structure optimization in the UFF force field [11] using the OpenBabel

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

software [12]. The structure of SARS-CoV-2 M^{pro} prepared in this way was used for the preliminary and refining molecular docking both during the dataset preparation as well as for molecular docking of the generated compounds.

b) Preparation of Ligand Structures: Prior to the preliminary docking, preparation of the ligand structures was the same as described for the SARS-CoV-2 M^{pro} structure. This procedure was performed using OpenBabel but included an additional step of rotatable bonds identification which is auto-made by this software. However, prior to the refining molecular docking, the ligand structures were prepared via the following two steps: i) optimization in the MMFF94 force field [13, 14] to remove steric clashes and addition of hydrogen atoms that are absent in the initial structure, both using the RDKit module in Python 3, ii) addition of partial charges by Gasteiger scheme and rotatable bonds identification using MGLTools (http://mgltools.scripps.edu/). It should be noted that before subjecting the generated compounds to molecular docking, 3D structures of novel ligands were obtained from the generated linear SMILES notations, as described above. Further preparation steps of the generated compounds for molecular docking included the addition of partial charges by Gasteiger scheme and rotatable bonds identification performed by MGLTools.

c) Molecular Docking Settings: The preliminary molecular docking was performed using the QuickVina 2 [15] program, and refining molecular docking was carried out by AutoDock Vina [16] software, both in the approximation of rigid receptor and flexible ligands. In both cases, the grid box included the catalytic site of SARS-CoV-2 M^{pro} with the following parameters: $\Delta X = 19$ Å, $\Delta Y = 21$ Å, $\Delta Z = 23$ Å centered at $X = -20$ Å, $Y = 19$ Å, $Z = -26$ Å. Thus, the grid box volume was $19\times2 \times23 = 9177$ Å^3. The value of the exhaustiveness parameter defining the number of individual sampling "runs" was set to 10 and 50 for preliminary and refining docking, respectively.

D. Deep learning

a) Models Architectures: Two deep learning models have been developed, namely an unsupervised SMILES-based Long Short-Term Memory (LSTM) [17] autoencoder (model I) and a semi-supervised SMILES-based LSTM autoencoder (model II) in which the value of binding free energy was used as an additional dimension in latent space to learn from the docked compounds and a value to sample around in the generation mode.

Model I (Fig. 1, I) takes vectorized SMILES as input, which follow through the LSTM layer. The peculiarity of this model is defined by the fact that LSTM output itself is not used, instead the hidden and cell states vectors are derived, which are concatenated together and then put through a dense layer. The output of this dense layer serves as a latent vector or SMILES embeddings in the context of the autoencoder. The embeddings are fed to two dense layers in parallel, creating initial hidden and cell state inputs for the LSTM layer in the decoder part. There is also the decoder input layer used as input for the decoder LSTM, which in the training mode receives

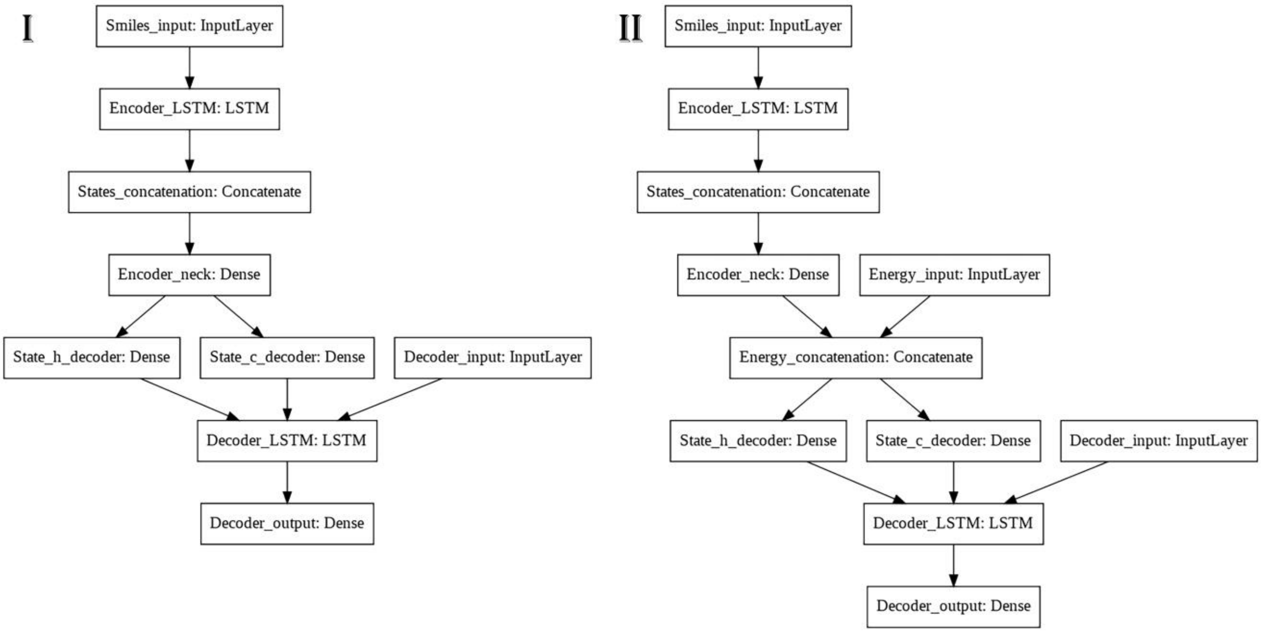

Fig. 1. Architectures of the autoencoder models: (I) Unsupervised (embeddings) model;
(II) Semi-supervised (embeddings and energy) model

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

the same vectorized SMILES as the encoder input, and, as a conventional LSTM generative model, it predicts the next symbol. However, in the generation mode, decoder input starts the generation process with a start symbol only, embeddings are used to predict the initial states of decoder LSTM and they basically define which kind of SMILES will be generated.

The so-defined energy model (Fig. 1, II) differs from model I in the neuron responsible for the binding affinity value situated on the latent layer of the model. While model I allows one to generate molecules from random SMILES embeddings as well as adding noise to SMILES embeddings of ligands with predicted affinity, the energy model enables one to generate new ligands with a preset property of binding affinity, in addition to attempts to manipulate SMILES embeddings of the given ligands to try to improve their structures after decoding and thus increase the value of binding affinity.

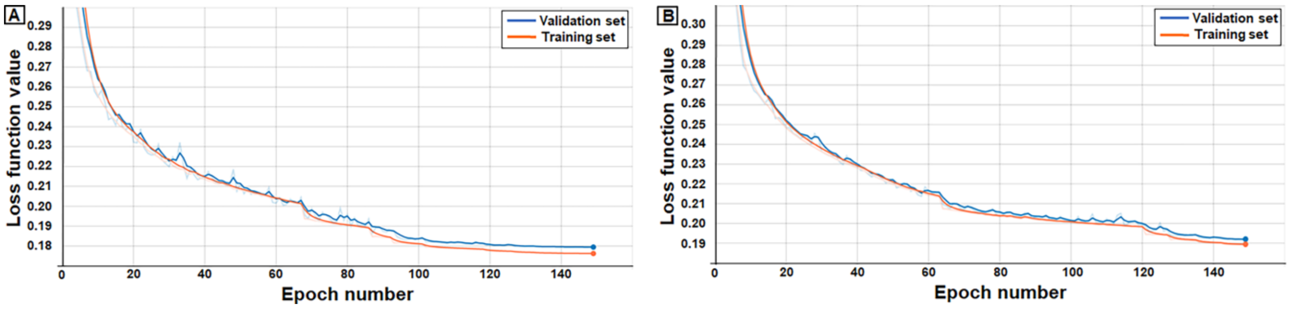

Fig. 2. Train and validation losses for autoencoder model: (A) Semi-supervised (embeddings & energy) model; (B) Unsupervised (embeddings) model

b) Models Training: Both models were composed layer by layer using TensorFlow 2.1 (https://www.tensorflow.org/) high-level API. The models were subject to 150 epochs of training, additionally "Reduce learning rate on a plateau" and "Early stopping" callbacks were used to help the model converge to a better local minima and also avoid overfitting. Stochastic gradient descent optimization method Adam [18] was used as an optimizer with 0.005 learning rate initial value and the categorical cross-entropy loss function was chosen. The loss score progress for both models is shown in Fig. 2.

E. Deep Learning-based Compounds Generation

Two methods of generation were subjects of our consideration: generation from random numbers drawn from normal distributions, where distributions parameters were derived using test data distribution on the latent layer for each vector component. For this method, the process of generation for model II implied setting an a priori value of binding free energy to approximate generated ligands with. Experiments for different thresholds were carried out. The major idea of the second method of generation was to sample best ligands from the test set, to try to add noise to their embeddings. This approach is supposed to change the reconstructed ligand, and, in the case of model II, also increase the predicted binding affinity, forcing the generation of more promising ligands. The combinations of two autoencoder models and two generation methods are summarized in Table I.

III. Results and Discussion

As noted above, both models were tested using each of two generation methods. The results obtained were evaluated based on the values of binding affinity predicted by molecular docking, as well as by comparing these values with those calculated for the reference compounds used in the virtual experiments as a positive control.

The results of compound generation common for all of the conducted simulations are summarized in Table II.

A. Embeddings Model, Random Vectors Generator

Despite the fact, that this model does not use both reference compounds for generation and values of binding energy, it is capable of generating new potential inhibitors of the selected target only by generating compounds from the embeddings distributions inferred from the training data. In the set of generated compounds, the share of molecules with the predicted values of binding free energy less than –9 kcal/mol was 3.2%, which exceeds the same share in the training dataset (1.8%) by almost 2 times.

B. Embeddings Model, Test Set Compounds Used To Generate New Compounds From

This generation method utilizes compounds available and tries to generate new potential inhibitors from them. The share of generated compounds with high binding affinity is considerably larger, with 38%

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

of compounds exhibiting the predicted values of binding free energy less than –8 kcal/mol and 4% of compounds showing the predicted values less than –9 kcal/mol.

TABLE I. Description of Investigated Combinations of Generative LSTM Autoencoder Models and Generation Methods

Model	Generation starting point description	Generation process description
Unsupervised (Embeddings model)	Random number vectors drawn from fitted normal distributions	Random numbers are used as embeddings and fed to the decoder
Unsupervised (Embeddings model)	Compounds with free binding energy less than –9 kcal/mol, sampled from test set	Embeddings for these compounds are calculated, then distortion is added and updated embeddings are fed to the decoder
Semi-supervised (Energy model)	Random number vectors drawn from fitted normal distributions and a preset free binding energy value	Random vectors are used as embeddings and are passed as latent layer inputs along with a preset free binding energy value
Semi-supervised (Energy model)	Compounds with free binding energy less than –8 kcal/mol, sampled from test set and improved free binding energy values	Embeddings for these compounds are calculated, then distortion is added and updated embeddings along with improved free binding energy values are passed to the decoder

C. Energy Model, Generation from Random Numbers with a Set of Preset Thresholds of Binding Affinities

The semi-supervised model utilizes a version of “style and content” disentanglement for molecular data. According to the results obtained, 31% and 64% of generated compounds showed the values of binding energy within the deviations from the pre-defined energy value equal to 1 kcal/mol and 2 kcal/mol, respectively.

D. Energy Model, Test Set Compounds Used as Starting Points to Generate More Compounds, Energy Threshold is Shifted Towards Lower Energy by 0.5 and 1.0 kcal/mol Steps

This combination of the model and method proved to generate the top compounds throughout all four modes of generation. 52% of generated compounds are located within 1 kcal/mol deviation from the reference compounds, while 16% of generated compounds have the values of binding free energy lower than –9 kcal/mol.

IV. Conclusion

Two generative autoencoder models for prediction of novel drugs against SARS-CoV-2 were developed and applied to identify potential inhibitors able to block the catalytic site of the coronavirus main protease. The designed generative models combined with molecular docking proved their great potential to enrich screening pipelines with new compounds with desired properties. The generative power of the designed models is confirmed by the fact that out of 4805 successfully generated compounds only one compound was found in the original dataset. This indicates the richness of unexplored chemical space and proves an importance of development and application of generative models in drug design and discovery.

TABLE II. Compounds Generation Results Using Two Generative LSTM Autoencoder Models and Two Generation Methods

Model name, generation method	Number of generated compounds	Number of successfully docked compounds	Number of compounds with the predicted binding free energy less than –8 kcal/mol	Lowest predicted binding free energy, kcal/mol	Fraction of generated compounds with a lower binding affinity compared to reference compounds or energy threshold
Embeddings, random vectors	1000	986	277	–10.6	–
Embeddings, reference compounds	2543	2518	967	–10.3	> 10.0 %
Energy, random vectors and energy value	600	594	161	–9.3	> 17.4 %
Energy, reference compounds and improved energy values	662	658	266	–10.4	> 12.2 %

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

Acknowledgments

This study was financed by a grant of the Belarusian Republican Foundation for Fundamental Research (project F21COVID-002). The authors are grateful to EPAM Systems for support of the project and interest in the study.

References

[1] T. Iqbal and S. Qureshi, "The survey: Text generation models in deep learning," Journal of King Saud University - Computer and Information Sciences, April 2020.

[2] V. Sorin, Y. Barash, E. Konen, and E. Klang, "Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) – A Systematic Review," Academic Radiology, vol. 27, no. 8, pp. 1175–1185, August 2020.

[3] S. Kim et al., "PubChem in 2021: new data content and improved web interfaces," Nucleic Acids Research, vol. 49, no. D1, pp. D1388–D1395, January 2021.

[4] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," J. Chem. Inf. Comput. Sci., vol. 28, no. 1, pp. 31–36, February 1988.

[5] A. M. Andrianov, G. I. Nikolaev, N. A. Shuldov, I. P. Bosko, A. I. Anischenko, and A. V. Tuzikov, "Application of deep learning and molecular modeling to identify small drug-like compounds as potential HIV-1 entry inhibitors," Journal of Biomolecular Structure and Dynamics, pp. 1–19, 15 April 2021.

[6] J. Sunseri and D. R. Koes, "Pharmit: interactive exploration of chemical space," Nucleic Acids Research, vol. 44, no. W1, pp. W442–W448, July 2016.

[7] T. Pillaiyar, M. Manickam, V. Namasivayam, Y. Hayashi, and S.-H. Jung, "An Overview of Severe Acute Respiratory Syndrome–Coronavirus (SARS-CoV) 3CL Protease Inhibitors: Peptidomimetics and Small Molecule Chemotherapy," J. Med. Chem., vol. 59, no. 14, pp. 6595–6628, July 2016.

[8] D. Schneidman-Duhovny, O. Dror, Y. Inbar, R. Nussinov, and H. J. Wolfson, "Deterministic Pharmacophore Detection via Multiple Flexible Alignment of Drug-Like Molecules," Journal of Computational Biology, vol. 15, no. 7, pp. 737–754, September 2008.

[9] S. Wang, J. Witek, G. A. Landrum, and S. Riniker, "Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences," J. Chem. Inf. Model., vol. 60, no. 4, pp. 2044–2058, April 2020.

[10] J. Gasteiger and M. Marsili, "A new model for calculating atomic charges in molecules," Tetrahedron Letters, vol. 19, no. 34, pp. 3181–3184, Januray 1978.

[11] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations," J. Am. Chem. Soc., vol. 114, no. 25, pp. 10024–10035, December 1992.

[12] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," Journal of Cheminformatics, vol. 3, no. 1, p. 33, October 2011.

[13] T. A. Halgren, "Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94," Journal of Computational Chemistry, vol. 17, no. 5–6, pp. 490–519, 1996.

[14] P. Tosco, N. Stiefl, and G. Landrum, "Bringing the MMFF force field to the RDKit: implementation and validation," Journal of Cheminformatics, vol. 6, no. 1, p. 37, July 2014.

[15] A. Alhossary, S. D. Handoko, Y. Mu, and C.-K. Kwoh, "Fast, accurate, and reliable molecular docking with QuickVina 2," Bioinformatics, vol. 31, no. 13, pp. 2214–2216, July 2015.

[16] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," Journal of Computational Chemistry, vol. 31, no. 2, pp. 455–461, 2010.

[17] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, November 1997.

[18] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv:1412.6980 [cs], January 2017, Accessed: July 14, 2021. [Online]. Available: http://arxiv.org/abs/1412.6980.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.