



Article

Neural Networks in the Design of Molecules with Affinity to Selected Protein Domains

Damian Nowak ¹, Rafał Adam Bachorz ² and Marcin Hoffmann ^{1,*}

¹ Quantum Chemistry Department, Faculty of Chemistry, Adam Mickiewicz University in Poznan, Uniwersytetu Poznanskiego 8, 61-614 Poznan, Poland

² Institute of Medical Biology of Polish Academy of Sciences, Lodowa 106, 93-232 Lodz, Poland

* Correspondence: marcin.hoffman@amu.edu.pl

Abstract: Drug design with machine learning support can speed up new drug discoveries. While current databases of known compounds are smaller in magnitude (approximately 10^8), the number of small drug-like molecules is estimated to be between 10^{23} and 10^{60} . The use of molecular docking algorithms can help in new drug development by sieving out the worst drug-receptor complexes. New chemical spaces can be efficiently searched with the application of artificial intelligence. From that, new structures can be proposed. The research proposed aims to create new chemical structures supported by a deep neural network that will possess an affinity to the selected protein domains. Transferring chemical structures into SELFIES codes helped us pass chemical information to a neural network. On the basis of vectorized SELFIES, new chemical structures can be created. With the use of the created neural network, novel compounds that are chemically sensible can be generated. Newly created chemical structures are sieved by the quantitative estimation of the drug-likeness descriptor, Lipinski's rule of 5, and the synthetic Bayesian accessibility classifier score. The affinity to selected protein domains was verified with the use of the AutoDock tool. As per the results, we obtained the structures that possess an affinity to the selected protein domains, namely PDB IDs 7NPC, 7NP5, and 7KXD.

Keywords: machine learning; neural networks; molecular docking; ROR γ ; drug design; SELFIES



Citation: Nowak, D.; Bachorz, R.A.; Hoffmann, M. Neural Networks in the Design of Molecules with Affinity to Selected Protein Domains. *Int. J. Mol. Sci.* **2023**, *24*, 1762. <https://doi.org/10.3390/ijms24021762>

Academic Editors: Janusz Rak, Magdalena Zdrowowicz and Lidia Chomicz-Mańska

Received: 14 December 2022

Revised: 4 January 2023

Accepted: 5 January 2023

Published: 16 January 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Designing a molecule that can effectively bind to a target protein domain is essential in the drug discovery process [1,2]. Computational methods can speed up the screening in a virtual manner, which helps to reduce excessive costs and the length of time necessary during the conduction of experimentally based techniques (so-called in-vitro or in-vivo studies) [3].

The use of a neural network may help speed up the acquisition of molecules that are similar to the known molecules of the desired properties. Artificial intelligence enables obtaining new biologically active compounds derived from known molecules via modifications that are necessary to better fit the pharmacological purposes based on molecular descriptors [3].

Crystallography and multidimensional nuclear magnetic resonance (NMR) [4] provide structural information deposited in a protein data bank (PDB) [5,6] that can be used during a search for interactions between a newly designed potential drug and selected macromolecules.

Currently, the methods employed by the most popular programs assume the flexibility of a ligand (a small molecule), and the rigidity of a receptor. Such an approach leads to a cost and time reduction. Programs, such as AutoDock [7], Flex [8], DOCK [9], GOLD [10], ICM [11], Glide [12], Ligand Fit [13], and others, bind small molecules to proteins [14]. Molecular dynamics simulations can be used to analyze the time-dependent evolution of ligand–receptor complexes and provide tools that help in macromolecule relaxation.

This approach is more computationally demanding and requires more computational power [14].

In recent decades, a variety of docking programs have been developed for either academic or commercial use (vide infra). Different solutions and strategies are exploited in the context of ligand placement in the protein environment. In principle, they can be divided into four categories: stochastic Monte Carlo (Glide), fragment-based (Surflex, Flex), evolutionary-based (GOLD, AutoDock), and shape-complementary methods (LigandFit) [14,15]. A systematic search is not used due to the impossibility of the exploration of all degrees of freedom during molecular docking procedures. It is due to enormous computational costs. For example, if one is going to examine the cubic active site of 10^3 \AA^3 with a simple ligand, and when energy evaluation is done every 10° (change of the angle between the small molecule and receptor), as well as a rigid movement every 0.5 \AA [14] for a drug with four rotatable bonds only, there are 6×10^{14} [6] conformations to be checked. If our computer is fast enough to compute 1000 conformations per second, the whole procedure would take 19,025 years to complete the systematic approach [14].

The studies presented here are aimed at proposing new molecules that may be considerable ligands against the selected protein domains (ROR γ). This was accomplished by combining artificial intelligence that handles new potential drug generation with a chosen molecular docking program, i.e., AutoDock [16], which is in charge of determining the efficiency of the new ligand's binding to the chosen receptor.

The protein domains researched in this study belong to ROR γ proteins. They are referred to as orphan receptors since their natural ligands are undetermined [17]. ROR γ proteins are associated with several processes in our bodies, including metabolic regulation, whole-body development, cell apoptosis, homeostasis maintenance, and circadian rhythm modulation [18,19]. The biological relevance is that they are associated with a variety of human diseases, such as atherosclerosis, osteoporosis, autoimmune disorders, obesity, asthma, and cancer [20–22]. RORs are thought to be the key regulators of Th17 differentiation [23,24]. They can be found in the heart, liver, testis, and muscles [18,25].

The main aim of the study was to check the possibility of new chemical structure generations with the application of artificial intelligence—in this case, neural network architecture. The machine learning model should be trained on how chemical structures are constructed. When the model has some chemical knowledge, it can generate chemically correct structures. The output is the SMILES code of a molecule [26], which is extremely useful because it can easily be used in other *in silico* tasks [1,2].

The second goal is to perform molecular docking automatically. This solution leads to a more effective energy-binding calculation when the whole process can be done with a dedicated script. Thus, many potential ligands can be checked inside specific macromolecules and their interactions can be compared with each other [27]. The use of IT tools enhances molecular screening due to the avoidance of the necessity of the manual preparation of a ligand and a given receptor. This approach makes it more efficient, and many systems can be studied [27].

Indeed, it helps us in obtaining the best structures, which have the most favorable binding energies (lowest values). They may be later synthesized, and an experimental verification can be carried out [28]. Within the procedure briefly sketched above, the search for new drugs can be boosted, and the selection of structures that will most likely exhibit desired properties can be achieved.

The model may predict new chemical structures based on the initial structures. The model can be used in drug or chemical designs in general, depending on the training data. This strategy can be used to solve a variety of structural design problems.

2. Results and Discussion

2.1. Model—The Neural Network

The prepared sequence-to-sequence model can construct semantically correct structures. The model was trained with the use of 121,000 structures. The applied loss function,

i.e., the categorical cross-entropy [29] (see SM Equation (S6)) shows that training progresses well and the loss value converges to 0 (see Figure 1).

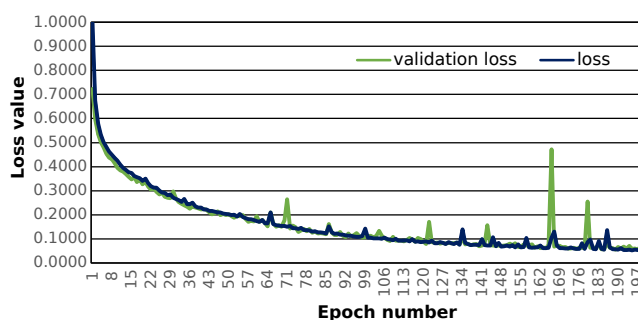


Figure 1. The training loss and validation loss during seq_to_seq model training. It follows the general rules of model training during time evaluation; both values are decreasing, indicating that the model is learning how to reconstruct the molecular sequence of the training molecules. It can be noticed that the loss value for the validation data also decreases, which means that the model learns to generalize.

The given model shows the possibility of learning chemistry via neural networks. Application of SELFIES codes leads to no errors during the prediction step. In the case of the application of SMILES chemical information, there is a possibility of the formation of incorrect structures [2,30]. This fact is due to the lower robustness of the SMILES notation in comparison to SELFIES [2,30]. This model can handle molecules whose representation in the SELFIES form is shorter than 65 characters.

The other approaches to de novo drug design include conditional recurrent neural networks [1], which produce more targeted output, fingerprints of known molecules using sequence-to-sequence reconstruction [31], and multi-layered gated recurrent units (GRU), as well as other RNN architectures [1,2,30]. In these approaches, the steering of what the neural network produces is different.

The main limitation of the used model is the maximal length of molecular sequence that can be effectively encoded. The neural network employed can handle up to 65 SELFIES characters, although there is no restriction to the maximum SELFIES length in general. The neural network's input layer definition determines it. This approach can handle SELFIES representations of molecules of any length, however, dealing with lengthy representations will be computationally expensive. It is important to remember that the SELFIES length is related to the length of the molecular sequence. This leads to a possible limitation of the initial structure's molecular sequence length. Although the loss and validation loss values may be lower as the number of training epochs increases, this newly created one allows one to search for a close molecular space, which can also be promising for new drug discoveries.

The neural network proposed here is incapable of distinguishing whether a specific structure is active or not. This could be performed by employing subsequent classifiers and molecular docking based on the biological activity of specific receptors, such as nuclear receptors.

2.2. Prediction Initializers

By the choice, the 36 structures gathered by Y. Zhang, et al. [20] (see SM File S4) were selected. After the conversion, the stereochemistry of resulting SMILES codes was ignored. The distributions of the first and second selections are given below (see SM Figures S10 and S11).

The 21 structures passed through the QED selection and 9 passed through Lipinski's rule of 5. The SYBA score was calculated for each of the previously selected structures (see SM Figure S12). The "My score" for structures that passed the SYBA threshold was calculated (see SM Figure S13).

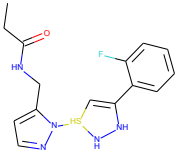
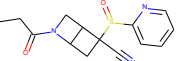
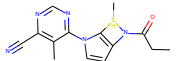
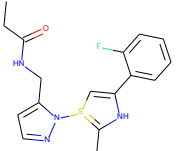
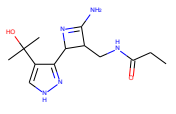
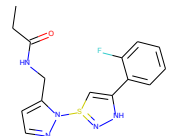
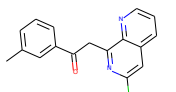
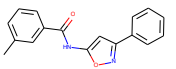
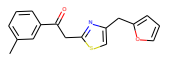
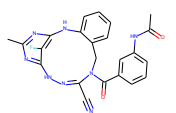
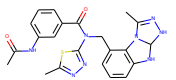
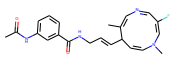
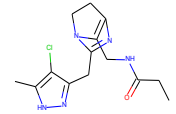
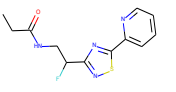
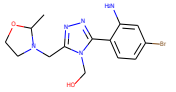
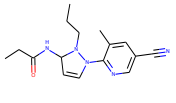
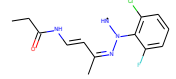
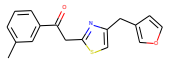
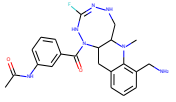
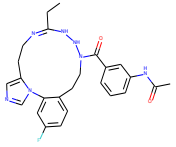
2.3. Predictions and Results of Selection

With 0.1 tensor scaling, 55 distinct structures were generated. Only three of them were found in PubChem. Their CIDs were: 16445174, 18006105, and 129773833. The QED descriptor calculation results (see SM Figure S14), along with Lipinski's rule of 5 fulfillment (see SM Figure S15), are shown below. Of these unique structures, 34 met the QED requirement, and 31 met the second discriminator. The number of structures that met both criteria is equal to 26, or 47.27% of the generated compounds.

Using the 0.2 tensor scaling, 78 distinct structures were generated. Only three of them were found in PubChem. Their CIDs were again: 16445174, 18006105, and 129773833. The QED descriptor calculation results (see SM Figure S16), along with Lipinski's rule of 5 fulfillment (see SM Figure S17), are shown below. Of these unique structures, 39 met the QED requirement, and 30 met the second discriminator. The number of structures that met both criteria was equal to 25, or 32.05% of the generated compounds.

Then 26 species from the first prediction and 25 from the second were combined, and repetitions were removed. It resulted in 42 unique structures. The third discriminator was applied—the SYBA classifier (see SM Figure S18). The results of its application resulted in 20 distinct structures (see Table 1).

Table 1. The structures of selected molecules. The table contains structure, structure number, structure of origin number, tensor scaling mode.

Newly Generated Structures' Images			
			
1, 1, 0.1	2, 1, 0.1	3, 1, 0.1	4, 1, 0.1
			
5, 1, 0.1	6, 1, 0.1	7, 2, 0.1	8, 2, 0.1 and 0.2
			
9, 2, 0.1	10, 4, 0.1	11, 4, 0.1	12, 4, 0.1 and 0.2
			
13, 1, 0.2	14, 1, 0.2	15, 1, 0.2	16, 1, 0.2
			
17, 1, 0.2	18, 2, 0.2	19, 4, 0.2	20, 4, 0.2

As many as 11 out of 20 structures come from the first structure's tensor scaling (see Table 1). This fact shows that the prediction given by this model is connected to the type of initial structure. After the removal of structures that cause some problems during the molecular docking procedure, the number of molecules reduces to 16 (see Table 2).

Compounds that “survived” the selection possess a generally high value in “My score” (see SM Figure S19), indicating that both the descriptor QED and SYBA scores are greater than zero. However, in two cases (molecule numbers ten and eleven), the lowest QED descriptor is maintained. In comparison with all the data after the first two selections, these structures go further due to a high SYBA score. Some structures possess the highest QED (molecule number fourteen) and SYBA score (molecule number eight) in comparison to all other structures generated after QED and SYBA selections.

Table 2. “My score”, QED normalized, and SYBA scores normalized, see SM Equation (S3) for selected structures with SM File S29; all structures meet the QED threshold and Lipinski’s rule of 5. They are used to determine the normalized QED and SYBA scores and “My score” (see SM File S41).

Molecule Number	QED	QED Normalized	SYBA Score	SYBA Score Normalized	“My Score”
4	0.82	0.76	39.47	0.75	75.10
5	0.62	0.22	5.24	0.64	42.70
6	0.88	0.92	35.55	0.73	82.67
7	0.54	0.02	79.24	0.87	44.65
8	0.79	0.68	119.44	1.00	84.00
9	0.67	0.35	90.13	0.91	63.02
10	0.53	0.00	47.43	0.77	38.68
11	0.53	0.00	85.03	0.89	44.54
12	0.74	0.56	7.05	0.64	60.04
14	0.91	1.00	47.22	0.77	88.54
15	0.79	0.69	17.76	0.68	68.36
16	0.90	0.97	4.28	0.63	80.42
17	0.67	0.36	15.68	0.67	51.67
18	0.67	0.35	95.91	0.93	63.93
19	0.54	0.01	6.40	0.64	32.45
20	0.54	0.01	47.67	0.77	39.20

All structures generated using the given method for potentially discovering new drugs were semantically valid. However, a new problem was discovered. a problem involving the possibility of exotic structure formation as well as some difficulties encountered while creating 3D structures. These difficult structures can be removed.

Most of the structures that are generated from the one that initially has high values of QED and SYBA score are above the threshold. This may lead to the conclusion that the initial structure has a significant role in the outcomes of the model.

2.4. Similarity to Initial Structures and Training Data

Structures from tensor scaling are not remarkably similar to initial structures (see SM Figure S20). One structure was replicated in the same shape. The lowest value of similarity was 0.17. Indicating that the AI-generated structure differs significantly from the initial structure.

Structures from tensor scaling are not highly similar to training molecules (see SM Figure S21). The highest similarity was found at 0.64, while the lowest similarity was 0.02; thus, the structures that give the results are different. The mean similarity is 0.26.

The comparability of structures that are selected for molecular docking gives information that these structures are also not identical (see Figure 2), and the thirteenth is the outlier as it has the lowest Tanimoto similarity output in each comparison.

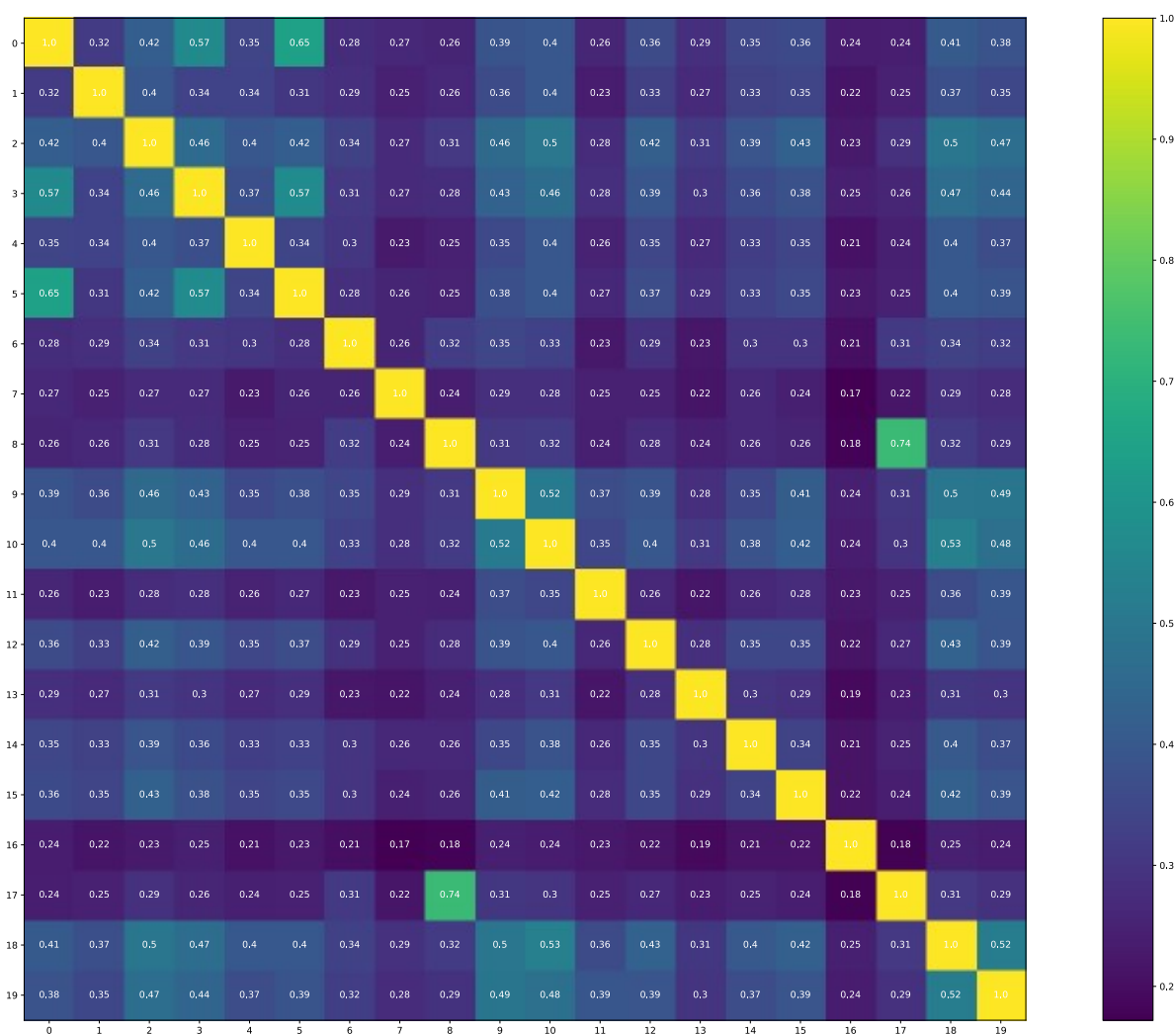


Figure 2. The Tanimoto similarity distribution along with the “to be docked” molecules.

When using the PubChemPy [32] library to see if the initial SMILES are present in PubChem, the following PubChem CIDs are returned: 807146, 16445174, 71470549, 71811962, and 135337558. These structures are related to the initial structures. Only one generated structure by the neural network was found in PubChem. All the above steps are performed in SM File S34. Questions can be raised about the similarity between structures after the first selection (42 objects) and the initial ROR γ active set of compounds containing five structures (see SM Figure S22). The highest hit is 1.00, so the one recreated structure. The lowest one is 0.14. SM Figure S23 shows the Tanimoto similarity between 42 structures and the training data (121,000 structures). As a result of SM File S36, neither the ROR γ active dataset nor any generated structure appears in the training dataset.

Further preprocessing and selection of training data could be applied. More similar structures may be generated if training structures are built with a charset that is more closely related to the charset of target structures.

2.5. Molecular Docking of Selected Structures

The results of molecular docking are collected in Table 3. The lowest average binding energy was found for molecule number 12, which equals -9.80 kcal/mol. The second was molecule number 20, with -9.70 kcal/mol. The third was the 19th, with -9.60 kcal/mol. The minimal binding energy for the 7NPC—macromolecule was -10.0 kcal/mol, and the structure related to this result is molecule number 10. In the case of the 7NP5 -10.0 kcal/mol binding energy, it is found to be the lowest value, and the corresponding structures are

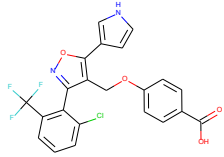
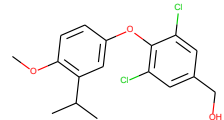
the 7th, 19th, and 20th. The 7KXD complex, with a minimum energy of -10.0 kcal/mol, has the lowest energy, and molecules 11th, 12th, and 19th are the causes of it. The most favorable average binding energy is for molecule number 5, at -6.7 kcal/mol. It is the worst candidate for a new drug based on binding energy. For each of the chosen macromolecules, the less favorable binding energies are -6.5 kcal/mol (7NPC), -6.9 kcal/mol (7NP5), and -6.7 kcal/mol (7KXD), and molecules are related to those results in the order of 5 (7NPC), 5 (7NP5), and 14 (7NP5), 5 (7KXD), respectively (see Table 3).

Table 3. Docked structures with docking scores in kcal/mol unit (see SM File S42).

Molecule Number	7NPC	7NP5	7KXD
4	-7.9	-8.2	-8.6
5	-6.5	-6.9	-6.7
6	-8.6	-8.1	-8.7
7	-9.2	-10.0	-8.7
8	-8.8	-9.0	-9.2
9	-8.0	-8.2	-8.5
10	-10.0	-9.1	-9.2
11	-8.7	-8.8	-10.0
12	-9.9	-9.5	-10.0
14	-6.6	-6.9	-7.2
15	-7.0	-7.1	-7.3
16	-7.5	-7.4	-7.7
17	-7.0	-7.1	-7.3
18	-9.7	-8.1	-8.3
19	-8.8	-10.0	-10.0
20	-9.5	-10.0	-9.6

In Table 4, the docking results are obtained for the re-docking procedure as we know the exact place of binding. A recreation of the initial pose was achieved in SM File S43 with the use of AutoDockTools 1.5.6. [16] (Software creator: Morris, G.M., et al., source of the program: ccsb.scripps.edu, Country of origin: La Jolla, United States of America). In the case of molecular docking performed with help of the Python tool, the results are significantly different. They are as follows: -8.0 kcal/mol for the 7NPC ligand, -9.2 kcal/mol for the 7NP5 ligand, and -7.7 kcal/mol for the 7KXD ligand (see SM File S37).

Table 4. Results of molecular docking for structures found in the Protein Database's raw files [33,34], along with QED and SYBA scores—see SM Files S30 and S43–S46.

Structure	QED	SYBA Score	Docking Result/Name of Domain
	0.35	45.11	-13.84 kcal/mol/7NPC -14.19 kcal/mol/7NP5
	0.79	71.02	-8.66 kcal/mol/7KXD

The differences in the results are caused by the different sizes of the search space; they demonstrate how significant the search space size is. The time during which the genetic algorithm is allowed to search for the best poses is crucial, as it may lead to different results.

Karaś et al. demonstrated that molecular docking can be a powerful tool for identifying active compounds against macromolecular targets [35]. It was discovered that AZ 5104 is active against ROR γ receptors [35].

Based on the study of D. Plewczyski, et al. [14], the tool that was used during the actual investigation failed in nearly 90 cases, while the total number of ligand–receptor pairs was 1300. AutoDock has successfully docked in approximately 93% of initial pairs. Computed ligand poses in the case of AutoDock led to poor results when compared with the initial pose of the molecule. The threshold, given in Å, was set at 2, and everything below it was marked as successfully docked. In this case, the tool utilized during this study was in last place with Flex [14]. According to D. Plewczyski, et al., AutoDock performs well in the case of small and hydrophilic molecules with either strong or weak binding energies (about 50% accuracy in top-scoring conformation-based analysis and about 76% accuracy in best-posing conformation-based analysis) [14]. Overall, the Pearson correlation for AutoDock, the top-scoring poses equal 25% and 19% in the best pose comparison). This could be due to the relative simplicity of the scoring functions and other assumptions that are made during virtual molecular docking. That is why the type of software provided is only a supporting tool in the drug design process [14].

Differences in energies after molecular docking procedures can be observed. It can be so due to the different requirements of each search. First, the so-called screening approach can be seen, in which many ligands are evaluated with many macromolecular systems. The second was to create visualizations in a bigger search space. This was done to check if the new structure would be attached to the same active site of the protein. It should be remembered that proteins possess different active sites. Some are for agonists, while others are for antagonists, so acting against these ROR γ domains is not out of the question.

Molecular dynamics simulations (MDS) have a significant impact on molecular biology and the discovery of novel drugs. The study did not include this. It is a location for undeniable improvements to the method presented.

3. Materials and Methods

The three protein domains used as receptor targets (examples of the ROR γ family of receptors) in this paper are 7NPC, 7NP5, and 7KXD, according to the PDB database infrastructure [5]. The choice was made according to the following requirements: they all belong to the ROR γ family, they have similar and good data resolution, and each structure has one main chain; these domain data were collected using X-ray crystallography, and the publication year was 2021 [5,6,33,34].

All mathematical background equations, methods, and files, along with the additional figures, are presented in the supplementary data (see Supplementary Materials) along with the additional figures. For the convenience of the reader, the whole project procedure is visualized in Figure 3. It reflects the proposed workflow involving a number of steps; in the following sections, we refer to the particular steps presented in the chart. The study starts with setting up the environment (see Stage 1, Figure 3).

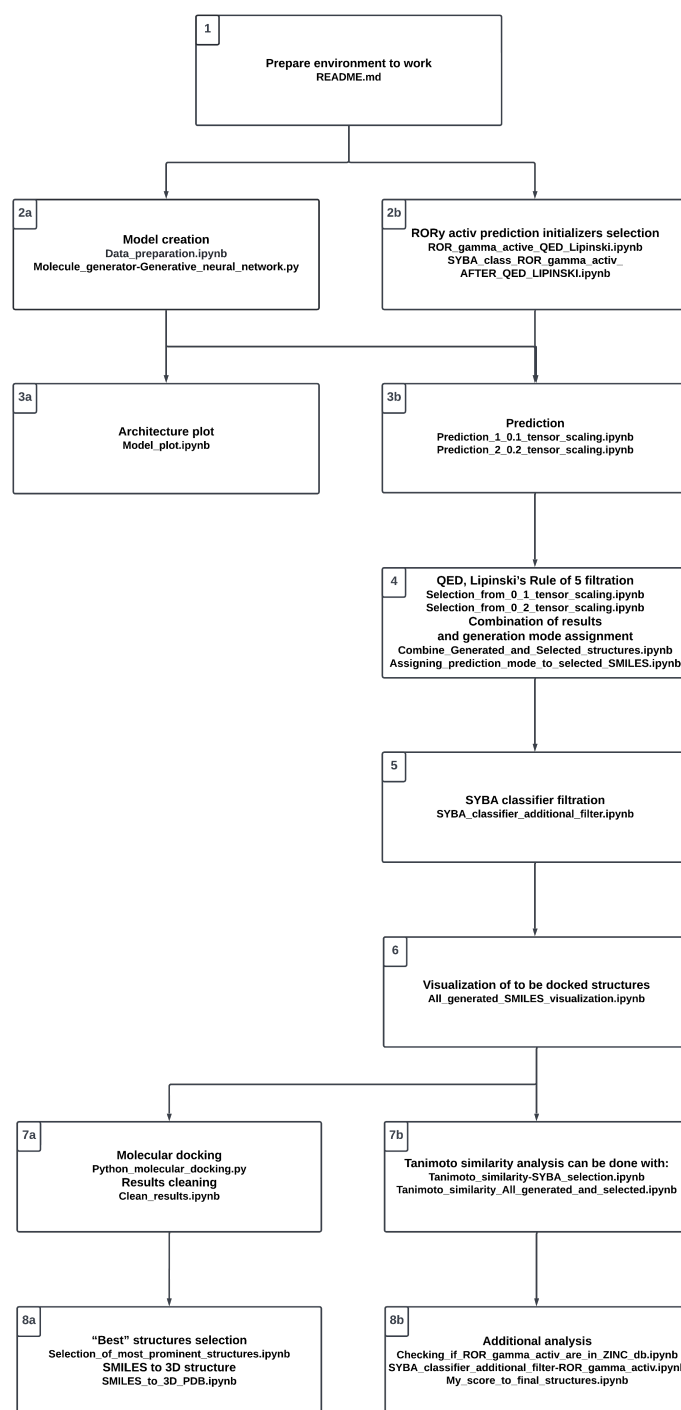


Figure 3. The flowchart of the overall workflow.

3.1. Training Data

The compounds were extracted as SMILES code representations from the ZINC20 database [36]. The SM File S1 contains the downloaded ZINC20 database (see Stage 2a, Figure 3).

This step consists of about 885,780,663 substances [37] within ca. 1800 tranches [37]. Then, in SM File S2, a selection of data that will be used to create a model can be found. This file contains information about the further tranches used. When the 935,475 unique isomeric SMILES are collected, they are translated into SELFIES [38], and the length of each SELFIES code is determined. It is assumed that only compounds with the SELFIES code length in the range of 30 to 50 are selected for further processing; this results in

569,205 structures. The length of SELFIES is proportional to the number of atoms in a specific structure.

The next step is to take advantage of the generic form of SMILES, which has a smaller charset in comparison to isomeric SMILES. As isomeric SMILES contain information about stereo centers and double-bond geometry, the charset is larger than in the case of generic SMILES, where the information is neglected.

As the SELFIES molecule representation is used in the neural network, it is necessary to convert SMILES into SELFIES. The charsets are different for the SMILES and respective SELFIES. The generic SMILES are translated into SELFIES codes and the lengths of the SELFIES are calculated.

Subsequently, the compounds containing rare heteroatoms (like Sn, Se, B, and P) were removed from the dataset (see SM Figure S1). The last effort is focused on the preparation of 121,000 structures for a model. This is achieved within SM File S2. The final structures that are our training data are collected in SM File S3. The data were then divided into training and validation sets using the sklearn library's "train test split" functionality. There was no preparation of specific data for differentiation of the training and validation sets. Moreover, 10% of the acquired data set was used as the validation data.

The resulting distribution is presented in Figure 4.

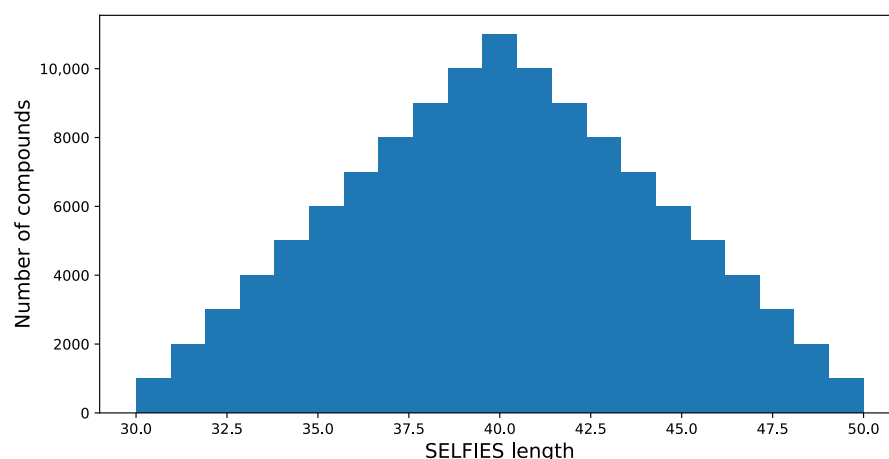


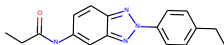

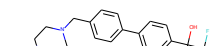
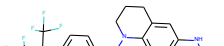
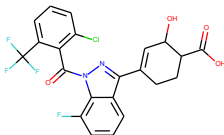
Figure 4. The training data SELFIES length distribution. As the training set, 121,000 structures are used.

3.2. Active Compounds against ROR γ

Based on the publication by Y. Zhang et al. [20], ROR γ active compounds were gathered (see SM File S4). Five structures were selected (see Table 5) to predict new structures. These structures were selected after the following steps (see Stage 2b, Figure 3).

As the training data are constructed based on generic SMILES codes (which are translated into SELFIES later), the collected data should also be converted into that form. The first step of selection (see SM File S5) means that the QED descriptor [39] (see SM Equation (S1)) is larger than 0.5. The second selection is done with the use of Lipinski's rule of 5 [40] (see SM Equation (S2)), and only structures that fulfill the conditions are selected. Then the results are saved in SM File S6, and the third classifier, the SYBA algorithm [41,42], is used (see SM Method S1). That one's threshold is set to above 0. It decides if the structure is easier (higher value) or harder (lower value) to synthesize (see SM Figure S2). After that, the normalization of the QED descriptor and SYBA score are calculated.

Table 5. These structures were used as prediction initializers.

Structure	QED	QED Normalized	SYBA Score	SYBA Score Normalized	“My Score”	ROR γ Activity
	0.80	1.00	140.57	0.70	85.19	agonist ¹
	0.79	0.96	119.44	0.59	77.43	agonist ¹
	0.69	0.60	119.78	0.59	59.77	inverse agonist ¹
	0.66	0.48	116.61	0.58	52.72	inverse agonist ¹
	0.52	0.00	5.72	0.00	0.00	inverse agonist ¹

¹ ROR γ activity is known from Zhang, et al. publication [20].

The last calculated parameter, called “My score”, is an arithmetic mean of the normalized QED descriptor and SYBA score (see SM Equation (S3)). In the next step, the results are saved in SM File S7. The results are sorted in descending order of “My score”, and five structures are selected—two from the top, two from the middle, and the last one; these were selected for the predictions. They were saved in SM File S8.

3.3. SELFIES Coder

The neural network requires a mathematical representation of the data, in this case, molecules, to be fed into it. In order to vectorize SELFIES code, the SELFIES coder tool was developed. A vectorization procedure converts a text representation of a species into a one-hot encoded form that is readable by a computer. This procedure generates the set of molecules in a form suitable for machine learning training [43]. The functionality SELFIES coder (see SM File S9) has been prepared in order to convert the SELFIES codes [38] into molecular sequences and further into numerical vectors (vectorization).

Since the SELFIES [38] encoding of molecules is more robust compared to SMILES [26] (invalid SMILES can be formed); this approach assures that all structures are semantically correct, which is not the case with the SMILES encoding [1,2,30]. Each SELFIES unit is turned into a mono-character according to certain rules, which ultimately leads to a molecular sequence (composed of characters from SM File S10). This molecular sequence is later transformed into a one-hot encoded form within a vectorization procedure.

The SMILES format is very useful for Python molecular docking approaches, and it is effectively supported by popular libraries, such as RDKit or PubChemPy. The SELFIES format was investigated for neural network applications and to determine whether semantically incorrect SELFIES creation is possible.

3.4. Model—The Neural Network

The neural network is created (see Stage 2a and Stage 3a, Figure 3) in SM File S11. The basic idea was to prepare a model based on recurrent neural units that are capable of generating molecules from those of known activity. The idea of the model’s construction was taken from [44]. The recurrent neural network (RNN) applied here (see SM Figure S3) takes advantage of long short-term memory (LSTM) cells [45], which help in learning dependencies of sequential data (see SM Method S2).

The data collected in the previous section (see SM File S3) are used to create the model. The 108,900 structures are used as training data, and 12,100 structures are used as validation data. The SMILES codes are translated into SELFIES, and from the SELFIES codes, the charsets are created (see SM Files S12 and S13). Then the molecules are in the form of a so-called molecular sequence, and two additional charsets are created, one containing translations from arbitral mono-signs (single characters) to numbers (see SM File S14) and the other containing translations from numbers to arbitral mono-signs (see SM File S15).

Now the molecular sequences can be coded into latent space (see SM File S16, SM Figure S4). Vectorization is the process of converting molecular sequences into one-hot encoded arrays. Additionally, the characters reflecting the beginning and the end of the molecule were introduced namely: “!” as the starting character and “E” as the ending character. Moreover, the latter is used as a padding character ensuring the same length of the molecules. The maximum length of the molecular sequence is called the embed value. It determines how long molecular sequences can be used during predictions.

The other functionalities are saved in SM Files S17 and S18. The first decodes states from latent spaces (see SM Figure S5) and the second is responsible for making (character by character) predictions (see SM Figure S6).

The data transformation scheme for the neural network is presented in SM Figure S7. The maximum epoch number was set to 200.

3.5. Predictions

The structure generation was carried out by tensor scaling and further converting molecular sequences back into SMILES (see Figure 5, see Stage 3b, Figure 3). It was initialized by previously selected structures—these are gathered in SM File S8 (see Table 5). The saved models are used, as well as the charsets. The initial structures are then translated into molecular sequences and vectorized so that the model can make predictions. The function `latent_to_mol_seq` (see SM Files S19 and S20) simply converts a latent space into the molecular sequence. As per the scheme, the states are taken from predictions made on newly obtained latent spaces and the states of the sample model are reset with new ones. Then a “for” loop makes character-by-character predictions until the “E” character is encountered, then it stops creating a molecular sequence.

Predictions are made around the initial latent vector of each initial molecule by adding random noises of varying amplitudes (0.1 and 0.2). For each initial molecule, 20 samples were taken, yielding 100 structures. Potential duplicates were then removed. As a result, molecular sequences were generated, which were then decoded into SELFIES and translated into SMILES. SMILES codes are converted to canonical form and searched in the PubChem database. These steps are shown in SM Files S19 and S20, with 0.1 and 0.2 tensor scaling, respectively. The results are saved in SM Files S21 and S22 for 0.1 and 0.2 tensor scaling, respectively.

Based on SM Files S21 and S22, the selection is made in SM Files S23 and S24 for each tensor scaling, respectively. If the QED descriptor [39] value is lower than 0.5, it acts as a discriminant (see SM Equation (S1), see Stage 4, Figure 3). Lipinski’s rule of 5 [40] is applied as the second discriminant (see SM Equation (S2)). The outcomes of both options were recorded in SM Files S25 and S26. The outcomes of the merger of SM Files S25 and S26 are then saved in SM File S27. Each structure is assigned a prediction mode (see SM File S28), with the results saved in SM File S29.

The third selection step was applied (see Stage 5, Figure 3). It was done with the use of the SYBA classifier [41,42] (see SM Method S1), resulting in SM Files S30 and S31 (see SM Figure S8). At an arbitrary value of 0, all species with an SYBA score less than this value were rejected from further consideration.

The structures that were selected can be viewed in SM File S32 (see Stage 6, Figure 3). Immediately after this, molecular docking of selected structures could be applied (SM File S33). If some structures prove problematic during the docking procedure, they can be

eliminated by rerunning the code present in SM File S32. Then molecular docking can be run one more time, but without problematic structures.

To analyze the potential drug-likeness of proposed structures, the QED and Lipinski's rule of 5 classifiers were utilized. The third classifier checks the structure's accessibility to synthesis. They are all employed before the molecular docking technique, assisting in the selection of the "best" structures for further testing.

The biological activity of the derived structures can be assumed based on the molecular docking technique and a partial resemblance to existing drugs with confirmed biological action. The technique suggested here could be thought of as a random search in the region of a specific species, such as one whose biological activity was experimentally determined. The changes incorporated into the species preserve the pharmacophoric properties of the original molecule to a considerable extent. Based on this assumption, it is reasonable to expect that the derived species will have similar characteristics. Nonetheless, experimental validation is the best technique to obtain a conclusion.

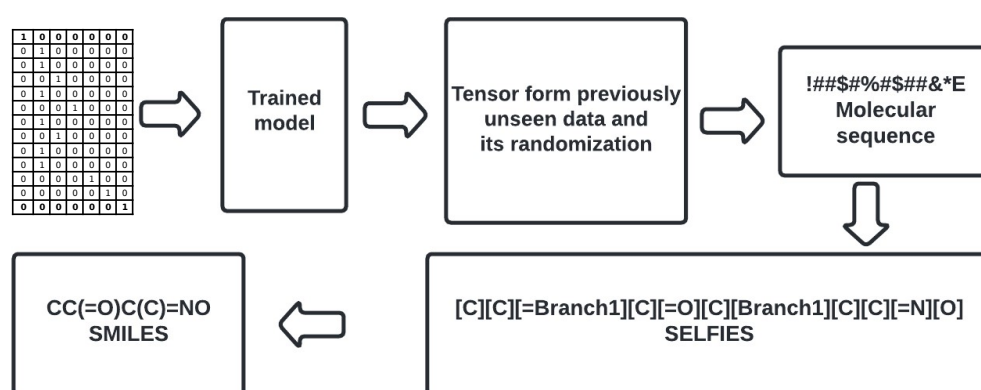


Figure 5. Workflow for data prediction.

3.6. The Molecular Similarity

A molecular fingerprint is an abstract representation of some features of a given structure. It is a compact representation of chemical structure expressed as a series of binary digits. The molecular fingerprint used here is the "RDKit" fingerprint which is yet another implementation of a daylight-like fingerprint [46]. In the molecular fingerprint representation, the similarity of two species can be easily calculated [47,48]. In particular, the Tanimoto similarity coefficient involves two fingerprints and reflects the similarity of the molecules [49] (see SM Method S3, Stage 7b, Figure 3). The value 1 represents identical molecules, while the value 0 indicates that no common components exist in the respective bit representations.

Firstly, a comparison between AI-predicted structures and the selected initial structures was made (see SM File S34). Then the Tanimoto similarity is calculated among the structures that are going to be docked. This means structures that meet the selection rules from the previous paragraph (see SM File S34). The third step is to compare training data with newly obtained compounds. The frequency axis has a much larger scale due to the number of compounds multiplied by the number of training structures (121,000). The results are displayed after calculations are performed for each docked structure and each training structure combination (see SM File S34).

After that, there is a simple check to see if the AI-generated structures can be found in PubChem (see SM File S34). The same procedure as described above is applied to all structures that meet QED and Lipinski's rule of five thresholds (see SM File S35). Inside SM File S36 (see Stage 8b, Figure 3), a check is performed to see if any of the ROR γ active compounds are present in the database, as well as if any of the reconstructed entries are present.

3.7. Molecular Docking

The search for possible ligand–macromolecule interactions was investigated via a molecular docking procedure and with the utilization of the Pyscreener tool [27] (see Stage 7a, Figure 3). The more negative the binding energy, the better the affinity of the ligand to the receptor.

The procedure was conducted in SM File S33—it is an automated way of carrying out molecular docking within a Python script. As a result, a file is created (see SM File S37). Binding energies, structures, and coordinates are stored inside SM File S38, see SM Method S4 along with SM Equations (S4) and (S5). The results are extracted via the code present in SM File S38. We have used here the AutoDock Vina [50,51] program with the Lamarckian genetic algorithm [52].

The average binding energies are calculated based on molecular docking experiments with three different macromolecules (see SM File S39). Based on molecular docking experiments with three different macromolecules, the average binding energies are calculated (see SM File S39). The code in SM File S40 then converts SMILES into a 3D structure that can be used for later manual molecular modeling. It is carried out by general chemistry rules concerning angles, hybridizations, and distances [53] (see SM Figure S9, (see Stage 8a, Figure 3)).

4. Conclusions

The model we proposed can create molecules that, in a limited manner, are similar to the training data and initial data. This can be due to the output formation–tensor scaling. This method takes advantage of more than one structure generation per single initial molecule as the input. The filters used (QED, Lipinski's rule of 5, and SYBA score) enable us to obtain structures with good synthetic possibilities as well as low binding energies. There is still a necessity for confirmation by synthesis and experimental measurements if these structures have actual affinities for selected protein domains. Our solution can be applied to other macromolecules of interest as well, where at least one active compound is known. The possibility of a new chemical structure generation with the application of artificial intelligence was shown. The machine learning model has indeed gathered some chemical knowledge. The use of the Python code (see SM File S33) resulted in the automatized molecular docking procedure. The study shows the possibility of new molecule formation via a neural network that will exhibit mathematical affinity to the selected protein domains. Any known protein domain and any known chemical with activity against it can be used to replicate the process.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/ijms24021762/s1>.

Author Contributions: Training data conceptualization and methodology, D.N. and R.A.B.; active compounds against the ROR γ conceptualization and methodology, D.N. and M.H.; SELFIES coder conceptualization and methodology, D.N. and R.A.B.; model—neural network conceptualization and methodology, D.N.; prediction conceptualization and methodology, D.N.; similarity conceptualization and methodology, D.N. and M.H.; molecular docking conceptualization and methodology, R.A.B., D.N., and M.H.; formal analysis, M.H. and R.A.B.; visualization, D.N.; supervision, M.H.; project administration, M.H. and R.A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received external funding from Narodowe Centrum Nauki (2019/33/B/NZ7/00795).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All publication-related information can be accessed at this address: https://github.com/XDamianX-coder/seq_to_seq_and_dock_AMU (accessed on 10 December 2022). Supplementary materials can also be downloaded using the link provided above.

Acknowledgments: Special thanks go out to Esben Jannik Bjerrum for his invaluable help with the predictive model and for his valuable comments about the content of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SM	supplementary material
AI	artificial intelligence
ROR γ	retinoic acid-related orphan receptors (gamma)

References

1. Kotsias, P.C.; Arús-Pous, J.; Chen, H.; Engkvist, O.; Tyrchan, C.; Bjerrum, E.J. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nat. Mach. Intell.* **2020**, *2*, 254–265. [[CrossRef](#)]
2. Segler, M.H.S.; Kogej, T.; Tyrchan, C.; Waller, M.P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120–131. [[CrossRef](#)] [[PubMed](#)]
3. Xu, Y.; Li, X.; Yao, H.; Lin, K. Neural networks in drug discovery: Current insights from medicinal chemists. *Future Med. Chem.* **2019**, *11*, 1669–1672. [[CrossRef](#)]
4. Homans, S.W. NMR spectroscopy tools for structure-aided drug design. *Angew. Chem. Int. Ed. Engl.* **2004**, *43*, 290–300. [[CrossRef](#)] [[PubMed](#)]
5. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [[CrossRef](#)] [[PubMed](#)]
6. Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S.; et al. The Protein Data Bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2002**, *58*, 899–907. [[CrossRef](#)]
7. Morris, G.M.; Goodsell, D.S.; Halliday, R.S.; Huey, R.; Hart, W.E.; Belew, R.K.; Olson, A.J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639–1662. [[CrossRef](#)]
8. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489. [[CrossRef](#)]
9. Ewing, T.J.; Makino, S.; Skillman, A.G.; Kuntz, I.D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411–428. [[CrossRef](#)]
10. Jones, G.; Willett, P.; Glen, R.C.; Leach, A.R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748. [[CrossRef](#)]
11. Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488–506. [[CrossRef](#)]
12. Friesner, R.A.; Banks, J.L.; Murphy, R.B.; Halgren, T.A.; Klicic, J.J.; Mainz, D.T.; Repasky, M.P.; Knoll, E.H.; Shelley, M.; Perry, J.K.; et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749. [[CrossRef](#)] [[PubMed](#)]
13. Venkatachalam, C.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: A novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph. Model.* **2003**, *21*, 289–307. [[CrossRef](#)] [[PubMed](#)]
14. Plewczynski, D.; Łażniewski, M.; Augustyniak, R.; Ginalska, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J. Comput. Chem.* **2011**, *32*, 742–755. [[CrossRef](#)] [[PubMed](#)]
15. Dar, A.M.; Mir, S. Molecular Docking: Approaches, Types, Applications and Basic Challenges. *J. Anal. Bioanal. Tech.* **2017**, *08*, 356. [[CrossRef](#)]
16. Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791. [[CrossRef](#)]
17. Hummasti, S.; Tontonoz, P. Adopting New Orphans into the Family of Metabolic Regulators. *Mol. Endocrinol.* **2008**, *22*, 1743–1753. [[CrossRef](#)]
18. Jetten, A.M. Retinoid-Related Orphan Receptors (RORs): Critical Roles in Development, Immunity, Circadian Rhythm, and Cellular Metabolism. *Nucl. Recept. Signal.* **2009**, *7*, nrs.07003. [[CrossRef](#)] [[PubMed](#)]
19. Giguère, V. Orphan Nuclear Receptors: From Gene to Function. *Endocr. Rev.* **1999**, *20*, 689–725. [[CrossRef](#)]
20. Zhang, Y.; Luo, X.y.; Wu, D.h.; Xu, Y. ROR nuclear receptors: Structures, related diseases, and drug discovery. *Acta Pharmacol. Sin.* **2015**, *36*, 71–87. [[CrossRef](#)]
21. Kurebayashi, S.; Ueda, E.; Sakaue, M.; Patel, D.D.; Medvedev, A.; Zhang, F.; Jetten, A.M. Retinoid-related orphan receptor γ (ROR γ) is essential for lymphoid organogenesis and controls apoptosis during thymopoiesis. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 10132–10137. [[CrossRef](#)] [[PubMed](#)]
22. Billon, C.; Sitaula, S.; Burris, T.P. Inhibition of ROR α/γ suppresses atherosclerosis via inhibition of both cholesterol absorption and inflammation. *Mol. Metab.* **2016**, *5*, 997–1005. [[CrossRef](#)]

23. Zhang, W.; Zhang, J.; Fang, L.; Zhou, L.; Wang, S.; Xiang, Z.; Li, Y.; Wisely, B.; Zhang, G.; An, G.; et al. Increasing Human Th17 Differentiation through Activation of Orphan Nuclear Receptor Retinoid Acid-Related Orphan Receptor γ (ROR γ) by a Class of Aryl Amide Compounds. *Mol. Pharmacol.* **2012**, *82*, 583–590. [CrossRef] [PubMed]
24. Solt, L.A.; Kumar, N.; Nuhant, P.; Wang, Y.; Lauer, J.L.; Liu, J.; Istrate, M.A.; Kamenecka, T.M.; Roush, W.R.; Vidović, D.; et al. Suppression of TH17 differentiation and autoimmunity by a synthetic ROR ligand. *Nature* **2011**, *472*, 491–494. [CrossRef] [PubMed]
25. Medvedev, A.; Yan, Z.H.; Hirose, T.; Giguère, V.; Jetten, A.M. Cloning of a cDNA encoding the murine orphan receptor RZR/ROR gamma and characterization of its response element. *Gene* **1996**, *181*, 199–206. [CrossRef] [PubMed]
26. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36. [CrossRef]
27. Graff, D.E.; Coley, C.W. pyscreener: A Python Wrapper for Computational Docking Software **2021**. *arXiv* **2021**, arXiv:2112.10575.
28. Huh, J.R.; Littman, D.R. Small molecule inhibitors of ROR γ t: Targeting Th17 cells and other applications. *Eur. J. Immunol.* **2012**, *42*, 2232–2237. [CrossRef]
29. Categorical Cross-Entropy. 2022. Available online: <https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/loss-functions/categorical-crossentropy> (accessed on 21 March 2022).
30. Bjerrum, E.J.; Threlfall, R. Molecular Generation with Recurrent Neural Networks (RNNs). *arXiv* **2017**, arXiv:1705.04612.
31. Xu, Z.; Wang, S.; Zhu, F.; Huang, J. Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Boston, MA, USA, 20–23 August 2017; ACM: Boston, MA, USA, 2017; pp. 285–294. [CrossRef]
32. Swain, M. PubChemPy. 2017. Available online: <https://github.com/mcs07/PubChemPy/> (accessed on 15 February 2022).
33. Meijer, F.A.; Saris, A.O.W.M.; Doveston, R.G.; Oerlemans, G.J.M.; de Vries, R.M.J.M.; Somsen, B.A.; Unger, A.; Klebl, B.; Ottmann, C.; Cossar, P.J.; et al. Structure-Activity Relationship Studies of Trisubstituted Isoxazoles as Selective Allosteric Ligands for the Retinoic-Acid-Receptor-Related Orphan Receptor γ t. *J. Med. Chem.* **2021**, *64*, 9238–9258. [CrossRef]
34. Ruan, Z.; Park, P.K.; Wei, D.; Purandare, A.; Wan, H.; O'Malley, D.; Stachura, S.; Perez, H.; Cavallaro, C.L.; Weigelt, C.A.; et al. Substituted diaryl ether compounds as retinoic acid-related orphan Receptor- γ t (ROR γ t) agonists. *Bioorganic Med. Chem. Lett.* **2021**, *35*, 127778. [CrossRef] [PubMed]
35. Karaś, K.; Salkowska, A.; Karwaciak, I.; Walczak-Drzewiecka, A.; Dastych, J.; Bachorz, R.A.; Ratajewski, M. The Dichotomous Nature of AZ5104 (an EGFR Inhibitor) Towards ROR γ and ROR γ T. *Int. J. Mol. Sci.* **2019**, *20*, 5780. [CrossRef] [PubMed]
36. Irwin, J.J.; Sterling, T.; Mysinger, M.M.; Bolstad, E.S.; Coleman, R.G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768. [CrossRef] [PubMed]
37. ZINC Database Tranches. Available online: <https://zinc20.docking.org/tranches/home/> (accessed on 5 April 2022).
38. Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. *arXiv* **2019**. [CrossRef]
39. Bickerton, G.R.; Paolini, G.V.; Besnard, J.; Muresan, S.; Hopkins, A.L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90–98. [CrossRef]
40. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26. [CrossRef] [PubMed]
41. Voršilák, M.; Svozil, D. Nonpher: Computational method for design of hard-to-synthesize structures. *J. Cheminform.* **2017**, *9*, 20. [CrossRef]
42. Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D. SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *J. Cheminform.* **2020**, *12*, 35. [CrossRef] [PubMed]
43. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*; Prentice Hall Series in Artificial Intelligence; Prentice Hall: Upper Saddle River, NJ, USA, 2000.
44. Bjerrum, E.; Sattarov, B. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. *Biomolecules* **2018**, *8*, 131. [CrossRef]
45. Brownlee, J. *Long Short-Term Memory Networks with Python: Develop Sequence Prediction Models with Deep Learning*; Jason Brownlee, Machine Learning Mastery, EBook, 2017
46. Cereto-Massagué, A.; Ojeda, M.J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63. [CrossRef]
47. Willett, P.; Barnard, J.M.; Downs, G.M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996. [CrossRef]
48. Bajorath, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233–245. [CrossRef] [PubMed]
49. Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry: Miniperspective. *J. Med. Chem.* **2014**, *57*, 3186–3204. [CrossRef] [PubMed]
50. AutoDock UserGuide. 2009. Available online: https://autodock.scripps.edu/wp-content/uploads/sites/56/2021/10/AutoDock4.2.6_UserGuide.pdf (accessed on 15 February 2022).
51. Atkins, P.W.; De Paula, J. *Physical Chemistry for the Life Sciences*; Freeman, W.H. Ed.; Oxford University Press: Oxford, UK; New York, NY, USA, 2006.

52. Ross, B. A Lamarckian Evolution Strategy for Genetic Algorithms. In *Practical Handbook of Genetic Algorithms*; Chambers, L., Ed.; CRC Press: Boca Raton, FL, USA, 1998. [CrossRef]
53. Landrum, G. RDKit: Open-Source Cheminformatics. 2017. Available online: <http://www.rdkit.org> (accessed on 1 February 2022).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.