AI

MDPI

Article

Structural Model Based on Genetic Algorithm for Inhibiting Fatty Acid Amide Hydrolase

Cosmin Trif, Dragos Paul Mihai *, Anca Zanfirescu and George Mihai Nitulescu

Faculty of Pharmacy, "Carol Davila" University of Medicine and Pharmacy, Traian Vuia 6, 020956 Bucharest, Romania

* Correspondence: dragos_mihai@umfcd.ro

Abstract: The fatty acid amide hydrolase (FAAH) is an enzyme responsible for the degradation of anandamide, an endocannabinoid. Pharmacologically blocking this target can lead to anxiolytic effects; therefore, new inhibitors can improve therapy in this field. In order to speed up the process of drug discovery, various in silico methods can be used, such as molecular docking, quantitative structure–activity relationship models (QSAR), and artificial intelligence (AI) classification algorithms. Besides architecture, one important factor for an AI model with high accuracy is the dataset quality. This issue can be solved by a genetic algorithm that can select optimal features for the prediction. The objective of the current study is to use this feature selection method in order to identify the most relevant molecular descriptors that can be used as independent variables, thus improving the efficacy of AI algorithms that can predict FAAH inhibitors. The model that used features chosen by the genetic algorithm had better accuracy than the model that used all molecular descriptors generated by the CDK descriptor calculator 1.4.6 software. Hence, carefully selecting the input data used by AI classification algorithms by using a GA is a promising strategy in drug development.

Keywords: drug discovery; anxiety; artificial intelligence; genetic algorithm; in silico

check for updates

Citation: Trif, C.; Mihai, D.P.; Zanfirescu, A.; Nitulescu, G.M. Structural Model Based on Genetic Algorithm for Inhibiting Fatty Acid Amide Hydrolase. *AI* **2022**, *3*, 863–870. https://doi.org/10.3390/ai3040052

Academic Editors: José Manuel Ferreira Machado and Kenji Suzuki

Received: 30 July 2022
Accepted: 9 October 2022
Published: 13 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 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

Long-term disability in middle-aged people, decreased life quality, increased mortality, and high financial costs are key issues associated with chronic pain [1]. Therapeutic options exist, but their use is frequently limited by the presence of serious side effects (e.g., respiratory depression, tolerance, and dependence for opioids). Over 60% of affected patients report dissatisfaction with currently available treatments [2–4]. Thus, there is a growing need for developing therapeutical agents with improved efficacy, tolerability and side effect profile in order to reduce the unnecessary and escalating burden of chronic pain.

The endocannabinoid system is emerging as a critical modulator of nociception, amplification of endocannabinoid signaling leading to analgesia in several preclinical models of acute and chronic pain. One of the main degradative enzymes for endocannabinoids, the fatty acid amide hydrolase (FAAH) regulates the endocannabinoid system by cleaving primarily the lipid messenger anandamide. Molecules inhibiting this enzyme are associated with highly potent analgesic and anti-inflammatory effect and developing such molecules represents a promising direction in finding a satisfactory treatment for chronic pain [5]. Several drug discovery campaigns have identified potent reversible FAAH inhibitors, such as oleoyl-based molecules (aldehydes, α-ketoamides, α-ketoesters, trifluoromethyl esters, heterocyclic derivatives), carbamates, substituted 2,4-dioxopyrimidine-1-carboxamides, and benzothiazole derivatives. The chemical space of such inhibitors is characterized by a high diversity [6].

The discovery of new drugs is problematic due to the high cost of testing a large number of molecules and the time necessary for this process [7]. Different in silico methods have been used in drug design to speed up the development of new active compounds,

such as molecular docking that predicts the most likely pocket of a protein where a small molecule will bind and its affinity, and the quantitative structure–activity relationship models (QSAR), which generate a mathematical equation between the pharmacological action and the chemical structure of a drug [8,9]. This relationship is based on different contributions of certain molecular descriptors to the molecule activity. Thus, their selection is a critical step for the successful implementation of this kind of algorithm. P. Ghosh and M.C. Bagchi used a genetic algorithm for this step in order to discover novel quinoxaline derivatives with antitubercular activity [10].

The big data sets available have paved the way for machine learning to become a new trend in solving complex problems in biological systems, such as predicting interactions between a molecule and its target protein [11–13]. The advantages of this method over classical dry lab methods are their ability to use high-dimensional data and complex nonlinear models.

In this case, a model is built, and the molecules that are to be evaluated are represented in a way that a computer can understand their characteristics. This can be done using molecular descriptors that can offer certain information, such as atom connectivity or atom coordinates and molecular fingerprints, which are a set of binary digits that represent the presence and absence of a particular molecular fragment [14,15]. They transpose the chemical information into numbers representing certain properties or the existence of certain atom types or structural motifs, their number of occurrences, or the distance between them. The descriptors and fingerprints generated by the CDK descriptor calculator 1.4.6 are used as inputs for a machine learning classification algorithm [16].

Logistic regression is a method borrowed from statistics that is among the most commonly used for binary classification problems [17]. Given the large number of features generated by the software, the accuracy of the algorithm can be affected. However, this issue can be solved using a genetic algorithm (GA) that creates individuals with different combinations of features and, in the end, selects the best individual with the optimal subset of features [18]. The combination of GA with machine learning algorithms alongside classical methods like QSAR can achieve high effectiveness and robustness. This was the case for a group that tried to discover a novel antihuman immunodeficiency virus molecules using this approach ant their model reached a sensitivity of 0.99, a specificity of 0.91, and an accuracy of 0.98 [19]. In search of molecules from the aforementioned class, the same group used a similar method called GA-ANFIS, which is composed of two phases: a GA coupled with logistic regression that was later validated with an adaptive neural fuzzy interference [20]. The coupling of the GA with logistic regression was also successfully used to uncover new inhibitors for interleukin-1 receptor-associated kinase 4 (IRAK-4) [21].

The Lamarckian evolutionary algorithm that only uses mutations to create new generations of individuals was also used to design novel molecules based on a given set of known molecules used to train the model to recognize compatibilities between molecular fragments and their synthetic accessibility [22,23]. Given the large number of studies that have successfully implemented a GA in the process of drug discovery, the scope of our current study was to create a structural model for discovering FAAH inhibitors using an AI classification algorithm based on logistic regression that uses the most relevant features selected by a GA in order to improve its accuracy. Datasets of biologically active molecules characterized by high structural diversity and molecular descriptors used as input variables can create unwanted noise while training predictive models, thus lowering the quality of the designed predictive models. Therefore, we aimed to investigate the case of known FAAH inhibitors and if the implementation of feature selection based on GA could potentially increase the prediction accuracy by reducing the number of variables included in the models.

2. Materials and Methods

2.1. Datasets Preparation

A structure data file (SDF), which is a modified mol file that can have multiple compounds at once, containing all the chemical structures of known human FAAH inhibitors and their activity expressed as half maximal inhibitory concentration (IC_{50}) values (M), was prepared from the ChEMBL database [24]. The raw entries were filtered using DataWarrior v5.0.0 software to remove all compounds with inexact IC_{50} values and to merge duplicate entries into a single entry with a calculated average IC_{50} value [25]. Two decoy datasets were obtained using RADER, a web server, using the FAAH inhibitors dataset and setting the Tanimoto coefficient (measures the similarity between two structures) to 0.75 between the ligands from the SFD and the decoys and 0.9 between ligands in the decoy set [26]. Molecules from the SDF file represented the "inhibitors" class (value = 1), and decoys were considered the "not inhibitors" class (value = 0).

The data from the SDF file was introduced into the CDK descriptor calculator 1.4.6. as an input file [16]. Hybrid, constitutional, topologic, electric, and geometrical descriptors were calculated alongside nine types of molecular fingerprints. The results were exported as an excel file and were used as inputs (features) for the classification algorithm.

2.2. Genetic Algorithm and Classification Algorithm

Using Sklearn and DEAP (Distributed Evolutionary Algorithms in Python) python modules, we implemented a GA and a classification algorithm [27,28]. We used logistic regression for the latter, and it was applied to all individuals generated by GA and in the case where all features were used to establish a baseline score. The function used for the binary classification was:

$$y = \frac{e^{(b_0 + b_1 * x)}}{1 + e^{(b_0 + b_1 * x)}} \quad (1)$$

where y is the predicted output (1 = inhibitor or 0 = not inhibitor), b_0 is the bias or intercept term, and b_1 is the coefficient for the single input value (x = value of descriptor or fingerprint).

We varied certain parameters that affect the model: Pop (population), Gen (number of generation), CxPb (crossover probability), MutPb (mutation probability), IndPb (probability of each gene to be replaced), and tournsize (tournament size). In the beginning, a number of population individuals were generated with a particular subset of features (genes) among the ones generated. The representation of each individual was characterized by a string of 1 (this gene is present) or 0 (this gene is not present) and a length equal to the number of descriptors and fingerprints offered by CDK. They were then put in tournaments where the ones with the highest value in the fitness function based on the y value were selected to participate in the mating pool and become parents for the next generation. Based on Cxpb and Indpb, parents swapped genes (features), and some were mutated (turned on or off, 0 transformed to 1 or vice versa) based on the MutPb, thus creating the new generation that will undergo the same process for a number of generations. In the end, the optimal subset of features was selected based on the individual with the highest fitness score in the classification algorithm.

2.3. Performance Metrics

We assessed the performance of each individual using the fitness function represented by the Matthews correlation coefficient (MCC):

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (2)$$

where TP is the true positive output, TN is the true negative output, FP is the false positive output, and FN is the false negative output.

The best value for MCC is 1, and the worst is −1. It is considered a better metric than others for binary classification problems because it takes into account the performance in all four categories of the confusion matrix [29].

3. Results

3.1. FAAH Inhibitor and Decoy Datasets

A dataset composed of 3042 human FAAH inhibitors with biological activity expressed in IC_{50} values was prepared using information from the ChEMBL database. Following the application of filters, a virtual chemical library was built by retaining 1249 compounds from the original dataset (set A). Using the RADER software, two decoy sets were generated, one containing 1051 decoys (set D1) and the other 13,379 decoys (set D2) [25].

A total of 281 molecular descriptors and 9 molecular fingerprints were calculated for sets A and D, and they were exported as an excel file that was later used to gather inputs for the algorithm.

3.2. Genetic Algorithm and Classification Algorithm Performance Metrics

The baseline score when using all features was 0.810 using the sets A and D1 (model1) and 0.8437 using sets A and D2 (model2). For model2, an L2 regularization that reduced overfitting was used to increase the fitness to 0.9124 (model3) [26]. Various experiments were executed on this model in order to find the optimal subset of descriptors and fingerprints. We tried several combinations of GA parameters starting from a random configuration, Indpb = 0.02, Cxpb = 0.2 and MutPb = 0.2 and tournsize = 5, and we varied the population (20, 30, 40, 50, 60, 70, 80, 90) for a number of generations until for at least four generations the maximum individual score remained constant. After finding the optimal population and generation, we then varied the other parameters, tournsize (4, 5, 6, 7, 8), IndPb (0.01, 0.02, 0.03, 0.04), CxPb (0.1, 0.2, 0.3, 0.4, 0.5), and MutPb (0.1, 0.2, 0.3, 0.4, 0.5), in order to improve our model. In Table 1, we provide the optimal configuration for each model, the maximum score, and the number of features for the best individual. Tables 2–6 present the results obtained from varying one parameter at a time for model3. The optimal results were obtained with the configuration: Pop = 90, Gen = 7, tournsize = 10, Indpb = 0.04, CxPb = 0.4, MutPb = 0.4 for model1, Pop = 50, Gen = 10, tournsize = 5, Indpb = 0.01, CxPb = 0.3, MutPb = 0.2 for model2, and Pop = 50, Gen = 11, tournsize = 7, Indpb = 0.04, CxPb = 0.2, MutPb = 0.4 for model3. Given the stochastic nature of the GA, we ran five experiments with the best configuration to retrieve an average of the best individuals generated and determined the optimal subset of features. The best individual had a score of 0.9348 and only 137 features. Among them, 134 descriptors were from the initial 281 and 3 fingerprints from the 9 generated by CDK.

Table 1. Optimal configuration for each model.

Model	Pop	Gen	Tournsize	IndPb	Cxpb	MutPb	Baseline Score	Max	N.D.
1	90	7	10	0.04	0.4	0.4	0.810	0.875	141
2	50	10	5	0.01	0.3	0.2	0.8437	0.8986	135
3	50	11	7	0.01	0.2	0.4	0.9124	0.9348	137

max = maximum value among individuals, N.D. = number of descriptors for the individual with the highest score.

The population size is usually in the range of tenths or hundreds of individuals. A higher number will offer a greater search space, but it will slow down the process. We tried varying the population size using increments equal to 10, and since no major improvements were noticed after reaching the value of 50, we concluded that this value is an optimal choice for the search space and process speed (Table 2).

We also observed that after eleven generations, there was no improvement in the accuracy, thus, we proceeded to the selection of the rest of the optimal parameters using only this number of generations (Table 3).

Table 2. Results after varying the number of population (Pop).

Pop	Avg	Std	Min	Max
20	0.9249	0.0019	0.9236	0.9281
30	0.9239	0.0025	0.9170	0.9261
40	0.9207	0.0032	0.9056	0.9236
50	**0.9267**	**0.0030**	**0.9102**	**0.9281**
60	0.9241	0.0032	0.9123	0.9281
70	0.9257	0.0023	0.9170	0.9281
80	0.9258	0.0022	0.91918	0.9282
90	0.9255	0.0018	0.91482	0.9283

avg = average value of fitness function, std = standard deviation, min = minimum value among individuals, max = maximum value among individuals. Values for optimal parameters are bolded.

Table 3. Results after varying the number of generations (Gen).

Gen	Avg	Std	Min	Max
0	0.8818	0.0167	0.8460	0.9104
1	0.8953	0.0096	0.8672	0.9125
2	0.9007	0.0075	0.8764	0.9125
3	0.9065	0.0050	0.8852	0.9125
4	0.9091	0.0034	0.8967	0.9169
5	0.9096	0.0064	0.8736	0.9169
6	0.9119	0.0057	0.8948	0.9214
7	0.9152	0.0035	0.9078	0.9259
8	0.9159	0.0071	0.8760	0.9259
9	0.9189	0.0043	0.9080	0.9259
10	0.9215	0.0037	0.9124	0.9259
11	**0.9240**	**0.0030**	**0.9105**	**0.9281**
12	0.9254	0.0019	0.9170	0.9281
13	0.9265	0.0015	0.9214	0.9281
14	0.9267	0.0030	0.9102	0.9281

avg = average value of fitness function, std = standard deviation, min = minimum value among individuals, max = maximum value among individuals. Values for optimal parameters are bolded.

Table 4. Results after varying the tournament size (tournsize).

Tournsize	Avg	Std	Min	Max
4	0.9183	0.0018	0.9123	0.9191
5	0.9254	0.0019	0.9170	0.9281
6	0.9269	0.0025	0.9147	0.9304
7	**0.9263**	**0.0016**	**0.9214**	**0.9326**
8	0.9278	0.0013	0.9236	0.9304

avg = average value of fitness function, std = standard deviation, min = minimum value among individuals, max = maximum value among individuals. Values for optimal parameters are bolded.

Table 5. Results after varying the value of each gene to be changed (Indpb).

Indpb	Avg	Std	Min	Max
0.01	**0.9285**	**0.0022**	**0.9192**	**0.9304**
0.02	0.9269	0.0025	0.9192	0.9282
0.03	0.9215	0.0043	0.8989	0.9258

avg = average value of fitness function, std = standard deviation, min = minimum value among individuals, max = maximum value among individuals. Values for optimal parameters are bolded.

Table 6. Results varying the value of crossover probability (CxPb).

CxPb	Avg	Std	Min	Max
0.1	0.9197	0.0033	0.9101	0.9236
0.2	**0.9254**	**0.0023**	**0.9147**	**0.9304**
0.3	0.9229	0.0016	0.9169	0.9236
0.4	0.9263	0.0027	0.9147	0.9281
0.5	0.9276	0.9276	0.9276	0.9276

avg = average value of fitness function, std = standard deviation, min = minimum value among individuals, max = maximum value among individuals. Values for optimal parameters are bolded.

From the number of individuals in the tournament, it was important to select the fittest ones while also allowing for some diversity among the populations [27]. We concluded that a number of seven individuals per tournament returned the best accuracy counting, as 14% of the initial population (Table 4).

The optimal values for Indpb, Cxpb, and MutPb were neither too low nor too high, 0.01, 0.2, and 0.4, which allowed the introduction of diversity in our system and prevented randomness and the disposal of good solutions that could be further improved (Tables 5–8).

Table 7. Results varying the value of mutation probability (MutPb).

MutPb	Avg	Std	Min	Max
0.1	0.9277	0.0014	0.9216	0.9282
0.2	0.9278	0.0025	0.9169	0.9303
0.3	0.9292	0.0020	0.9213	0.9303
0.4	**0.9307**	**0.0037**	**0.9193**	**0.9348**
0.5	0.9287	0.0041	0.9125	0.9327

avg = average value of fitness function, std = standard deviation, min = minimum value among individuals, max = maximum value among individuals. Values for optimal parameters are bolded.

Table 8. Results varying the value of mutation probability (MutPb).

Experiment	Max	AvgMax	std
1	0.9348		
2	0.9343		
3	0.9338	0.9345	0.0004
4	**0.9348**		
5	0.9347		

Max = maximum value among individuals, AvgMax = average value of maximum value, std = standard deviation. Values for optimal parameters are bolded.

4. Discussion

The scope of our paper was to build a predictive model for discovering new anxiolytic agents by inhibiting FAAH. Given the large number of features represented by molecular descriptors, the binary classification algorithm had low accuracy. In order to improve the performance, we built a genetic algorithm that could select the best subset of independent variables. In all our models, the GA improved the score, but as the baseline fitness was increased by a larger decoy set and using L2 regularization, the increase was lowered up to a maximum of 0.9348 for model3. The number of features was reduced to a similar amount for all models: 141 for model1, 135 for model2, and 137 for model3. Thus, around this value is the minimum number of molecular descriptors needed to characterize a molecule in order to classify it. In terms of the parameters of the GA that were varied for building the models, the population exhibited the highest difference. With the increase of molecules in the dataset, the population was reduced by almost half.

Although useful in increasing a predictive model's accuracy, genetic algorithms suffer from several limitations, such as difficult parameter optimization and high computational costs [30]. Although GAs are robust tools used in a variety of applications, in our case, the

implementation of a GA-based classification model led to a marginal increase in predictive accuracy.

5. Conclusions

Our work showed once again that the strategy of combining GA with machine learning methods like logistic regression could be successfully used in the identification of novel modulators of certain biological targets, such as FFAH in our case. We achieved a reduction in input data, represented as molecular descriptors, with a modest increase, below 10%, in the accuracy of the algorithm. We could not reach an accuracy as high as those shown by other studies in the literature (0.98 for the identification of novel anti-HIV drugs). However, this limitation can be explained by the small number of known modulators that were used to train the model and the amount of information given by the used descriptors. This issue will be addressed in further studies by using software that allows for a larger variety of molecular descriptors. This version of our algorithm, and further improved versions, will be used on large subsets of compounds in order to identify new inhibitors for FFAH that can hopefully be further validated and developed into potential analgesic agents.

Author Contributions: Conceptualization, G.M.N.; funding acquisition, A.Z.; investigation, C.T., D.P.M., A.Z. and G.M.N.; methodology, C.T. and D.P.M.; project administration, A.Z.; software, C.T.; writing—original draft, C.T. and D.P.M.; writing—review & editing, A.Z. and G.M.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by UEFISCDI grant number PN-III-P1-1.1-PD-2019-0574 no. 179/2020. The APC was funded by MDPI.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Linton, S.J.; Flink, I.K.; Vlaeyen, J.W.S. Understanding the Etiology of Chronic Pain From a Psychological Perspective. *Phys. Ther.* **2018**, *98*, 315–324. [CrossRef]
2. Baker, D.W. History of The Joint Commission's Pain Standards. *JAMA* **2017**, *317*, 1117. [CrossRef] [PubMed]
3. Evoy, K.E.; Covvey, J.R.; Peckham, A.M.; Ochs, L.; Hultgren, K.E. Reports of gabapentin and pregabalin abuse, misuse, dependence, or overdose: An analysis of the Food And Drug Administration Adverse Events Reporting System (FAERS). *Res. Soc. Adm. Pharm.* **2019**, *15*, 953–958. [CrossRef]
4. Mercadante, S. Opioid Analgesics Adverse Effects: The Other Side of the Coin. *Curr. Pharm. Des.* **2019**, *25*, 3197–3202. [CrossRef] [PubMed]
5. Chanda, D.; Neumann, D.; Glatz, J.F.C. The endocannabinoid system: Overview of an emerging multi-faceted therapeutic target. *Prostaglandins Leukot. Essent. Fat. Acids* **2019**, *140*, 51–56. [CrossRef] [PubMed]
6. Zanfirescu, A.; Nitulescu, G.; Mihai, D.P.; Nitulescu, G.M. Identifying FAAH Inhibitors as New Therapeutic Options for the Treatment of Chronic Pain through Drug Repurposing. *Pharmaceuticals* **2021**, *15*, 38. [CrossRef]
7. DiMasi, J.A.; Hansen, R.W.; Grabowski, H.G. The price of innovation: New estimates of drug development costs. *J. Health Econ.* **2003**, *22*, 151–185. [CrossRef]
8. Kwon, S.; Bae, H.; Jo, J.; Yoon, S. Comprehensive ensemble in QSAR prediction for drug discovery. *BMC Bioinform.* **2019**, *20*, 521. [CrossRef]
9. Yuriev, E.; Ramsland, P.A. Latest developments in molecular docking: 2010-2011 in review. *J. Mol. Recognit.* **2013**, *26*, 215–239. [CrossRef]
10. Ghosh, P.; Bagchi, M. QSAR Modeling for Quinoxaline Derivatives using Genetic Algorithm and Simulated Annealing Based Feature Selection. *Curr. Med. Chem.* **2009**, *16*, 4032–4048. [CrossRef]
11. Lavecchia, A. Machine-learning approaches in drug discovery: Methods and applications. *Drug Discov. Today* **2015**, *20*, 318–331. [CrossRef] [PubMed]
12. Pahikkala, T.; Airola, A.; Pietilä, S.; Shakyawar, S.; Szwajda, A.; Tang, J.; Aittokallio, T. Toward more realistic drug-target interaction predictions. *Brief. Bioinform.* **2015**, *16*, 325–337. [CrossRef] [PubMed]
13. Terfloth, L.; Gasteiger, J. Neural networks and genetic algorithms in drug design. *Drug Discov. Today* **2001**, *6*, 102–108. [CrossRef]

14. Mauri, A.; Consonni, V.; Todeschini, R. Molecular descriptors. In *Handbook of Computational Chemistry*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 2065–2093. ISBN 9783319272825.
15. Kausar, S.; Falcao, A.O. Analysis and comparison of vector space and metric space representations in QSAR modeling. *Molecules* **2019**, *24*, 1698. [CrossRef] [PubMed]
16. Willighagen, E.L.; Mayfield, J.W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; et al. The Chemistry Development Kit (CDK) v2.0: Atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminform.* **2017**, *9*, 1–19.
17. Urso, A.; Fiannaca, A.; La Rosa, M.; Ravì, V.; Rizzo, R. Data mining: Prediction methods. *Encycl. Bioinform. Comput. Biol. ABC Bioinform.* **2018**, *1–3*, 413–430.
18. Magliani, F.; Cagnoni, S.; Sani, L.; Prati, A. Genetic algorithms for the optimization of diffusion parameters in content-based image retrieval. In Proceedings of the 13th International Conference on Distributed Smart Cameras, Trento, Italy, 9–11 September 2019; pp. 1–6.
19. Labjar, H.; Labjar, N.; Kissi, M. QSAR Anti-HIV Feature Selection and Prediction for Drug Discovery Using Genetic Algorithm and Machine Learning Algorithms. In *EAI/Springer Innovations in Communication and Computing*; Ouaissa, M., Boulouard, Z., Ouaissa, M., Guermah, B., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 191–204. ISBN 978-3-030-77185-0.
20. Labjar, H.; Al-Sarem, M.; Kissi, M. Feature Selection Using a Genetic Algorithms and Fuzzy logic in Anti-Human Immunodeficiency Virus Prediction for Drug Discovery. *J. Inf. Technol. Manag.* **2022**, *14*, 23–36.
21. Pourbasheer, E.; Riahi, S.; Ganjali, M.R.; Norouzi, P. Quantitative structureactivity relationship (QSAR) study of interleukin-1 receptor associated kinase 4 (IRAK-4) inhibitor activity by the genetic algorithm and multiple linear regression (GA-MLR) method. *J. Enzyme Inhib. Med. Chem.* **2010**, *25*, 844–853. [CrossRef]
22. Kerstjens, A.; De Winter, H. LEADD: Lamarckian evolutionary algorithm for de novo drug design. *J. Cheminform.* **2022**, *14*, 1–20. [CrossRef]
23. Mouchlis, V.D.; Afantitis, A.; Serra, A.; Fratello, M.; Papadiamantis, A.G.; Aidinis, V.; Lynch, I.; Greco, D.; Melagraki, G. Advances in de novo drug design: From conventional to machine learning methods. *Int. J. Mol. Sci.* **2021**, *22*, 1676. [CrossRef]
24. Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107. [CrossRef] [PubMed]
25. Sander, T.; Freyss, J.; Von Korff, M.; Rufener, C. DataWarrior: An open-source program for chemistry aware data visualization and analysis. *J. Chem. Inf. Model.* **2015**, *55*, 460–473. [CrossRef] [PubMed]
26. Wang, L.; Pang, X.; Li, Y.; Zhang, Z.; Tan, W. RADER: A RApid DEcoy Retriever to facilitate decoy based assessment of virtual screening. *Bioinformatics* **2017**, *33*, 1235–1237. [CrossRef] [PubMed]
27. Li, H.; Phung, D. Journal of Machine Learning Research: Preface. *J. Mach. Learn. Res.* **2014**, *39*, i–ii.
28. Fortin, F.A.; De Rainville, F.M.; Gardner, M.A.; Parizeau, M.; Gagńe, C. DEAP: Evolutionary algorithms made easy. *J. Mach. Learn. Res.* **2012**, *13*, 2171–2175.
29. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 1–13. [CrossRef] [PubMed]
30. Vié, A. Qualities, Challenges and Future of Genetic Algorithms. *SSRN Electron. J.* **2021**, 1–48. [CrossRef]