Article

Journal of the Brazilian Chemical Society

https://dx.doi.org/10.21577/0103-5053.20240117

J. Braz. Chem. Soc. **2025**, *36*, 2, e-20240117, 1-12
©2025 Sociedade Brasileira de Química

Integration of LBDD and SBDD Studies on Drug Design: A Fatty Acid Amide Hydrolase (FAAH) Case Study

Pedro A. L. Santana,[a] Marina S. M. Ruas,[a] Gabriel C. Veríssimo,[a] Ana C. G. Terra,[a] Renata B. de Oliveira[a] and Vinícius G. Maltarollo*,[a]

*[a]Departamento de Produtos Farmacêuticos, Faculdade de Farmácia,
Universidade Federal de Minas Gerais, 31270-901 Belo Horizonte-MG, Brazil*

The inhibition of the fatty acid amide hydrolase (FAAH), an endocannabinoid system component, emerged as a potentially new therapeutic target for a range of clinical disorders such as acute and chronic pain. Some α-ketoheterocycle derivatives demonstrated interesting analgesic and anti-inflammatory activities *in vitro*. Ligand-Based Drug Design techniques such as knowledge graph convolutional networks (kGCN) and hologram quantitative structure-activity relationship (HQSAR) using α-ketoheterocycle derivatives from five different datasets were generated to discover the relation between the chemical structures and the inhibition activity. Meanwhile, structure-based drug design simulations as interaction fields (MIF), molecular docking, and ligand sites studies (LSI) from FAAH were performed using Autogrid software and FTmap/FTsite servers. The results of both studies were merged to propose predictive models. The resulting kGCN model demonstrated adequate accuracy area under the curve by receiver operating characteristic (AUC-ROC 0.7922). From contribution maps of the Ligand-Based Drug Design (LBDD) models and the generated probes using MIF and LSI, it was observed that the oxazole ring, the ketone group, and the apolar chain present in the structures of the inhibitors are important, besides the evidence of the Cys269 and Val270 residues importance for the potential interaction, confirmed by carried docking studies. These fragments and structural information can be used to carry out new FAAH potential inhibitors studies and report kGCN as an accurate classification technique.

Keywords: FAAH, KGCN, graph-based classification models, neural networks, HQSAR, molecular

Introduction

Fatty acid amide hydrolase (FAAH) belongs to a large group of enzymes termed the amidase signature family and is the main enzyme responsible for the metabolism of *N*-arachidonoyl ethanolamine (anandamide, AEA) and 2-arachidonoylglycerol (2-AG). AEA and 2-AG are biosynthetic ligands of cannabinoid G-protein coupled receptors CB_1 and CB_2 which are widely distributed in the central nervous system (CNS), the immune system, and the peripheral tissues of mammals (e.g., human, monkey, dog, mouse and rat). These components are part of the endocannabinoid system, an endogenous signaling system with physiological action on homeostasis regulation.[1-3]

The activation of cannabinoid receptors (CB_1 and CB_2) has been shown to relieve pain and inflammation, regulate motility and appetite, and produce anxiolytic effects in pre-clinical studies.[2-4] The corporal levels of AEA and 2-AG are modulated by its metabolism: FAAH metabolizes AEA into arachidonic acid (AA) and ethanolamine, while another enzyme, the monoacylglycerol lipase (MAGL), has been demonstrated to be the major hydrolase responsible for the transformation of 2-AG into AA and glycerol (Figure 1). Therefore, FAAH inhibitors can alter the signaling of AEA and, thereby, induce numerous biological responses by increasing the stimulation of CB_1/CB_2 receptors by endocannabinoid ligands.[2-4]

Despite being studied for over twenty years and having no inhibitors available on the market, FAAH remains a promising molecular target for the design of substances as potential treatment for inflammation conditions such as osteoarthritis of the knee, and neurodegenerative diseases such as schizophrenia and Tourette syndrome.[5-7] The development of FAAH inhibitors has significantly progressed and several selective inhibitors have been designed and synthesized over the last years: α-ketoheterocycle derivatives, carbamates, sulfonyl

*e-mail: viniciusmaltarollo@gmail.com
Editor handled this article: Paula Homem-de-Mello (Associate)

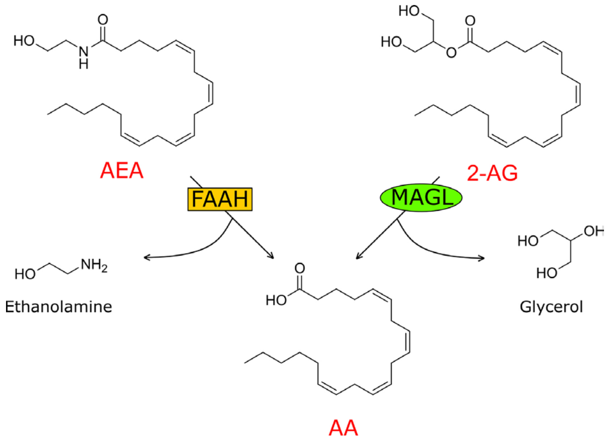


Figure 1. Hydrolysis of AEA and 2-AG mediated by FAAH and MAGL enzymes.

fluorides, ureas, boronic acids, aryl thiol heterocycles, 3-carboxamido-5-aryl isoxazole, 1,3,4-oxadiazol-2-ones, and others.[3,8]

The development of new FAAH inhibitors is, actually, extensively explored, and the use of computational techniques is a useful ally to make a more rational and assertive development regarding the pharmacological properties of new molecules. Since 2000, Boger and co-workers[9-13] have been developing and reporting several derivatives of the α-ketoheterocycle with FAAH's inhibition activities. With the structure and activity information, it is possible to develop computational models to predict the activity of new molecules, orient synthesis, and optimize structures of FAAH inhibitors.

Cheminformatic studies using machine learning (ML) and quantitative structure-activity relationship (QSAR) models are widely implemented in the search and development of bioactive compounds. Multiple approaches to QSAR modeling using various statistical or machine learning techniques and different types of chemical descriptors have been developed over the years.[14-16]

The prediction of potential interactions between a biomolecular target and its ligands can be made using computational studies such as molecular interaction fields (MIF) and ligand site interactions (LSI). The LSI technique is based on the identification of energetically important ligand site regions, called hotspots, through its interactions with a standardized set of probes (organic small molecules). On the other hand, the MIF studies are based on the exploration of the electronic and steric complementarity of the protein ligand site and some probes set (atoms). In general, the MIF and LSI studies can provide a characterization of the drugabillity of some potential inhibitors and binding sites.[17,18] A FAAH MIF study reported in 2009[19] identified some important potential interactions between the enzyme and some carbamates derivatives, but focusing on the steric proprieties of this compound and not taking into account conformational aspects of FAAH or the importance of the electrophilic group to the inhibition activity.

After that, the main objective of this work is to apply a graph-based neural network classification methodology, specifically the graph convolutional neural network (kGCN) method from Kojima *et al.*,[20] to understand the structure-activity relationship of a series of α-keto heterocyclic derivatives FAAH inhibitors. In addition, a comparison between obtained the results with other Ligand-Based Drug Design (LBDD) and Structure-Based Drug Design (SBDD) techniques will be helpful in designing novel derivatives of FAAH inhibitors.

Methodology

Dataset curation

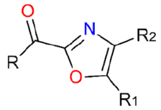
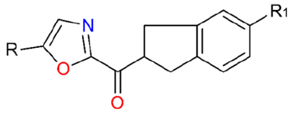
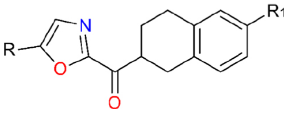
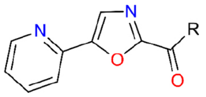
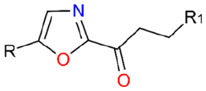
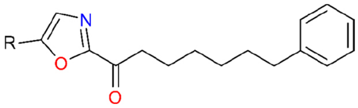
Five datasets of α-ketoheterocycle derivatives inhibitors of FAAH used for the LBDD studies were selected from the literature (Table 1).[9-13] The selected articles have similar or equivalent inhibition experimental conditions. The compounds with no defined stereochemistry and/or did not have an exact inhibition constant (*K*i) value were excluded from the analysis. In other words, biological activities annotations with <, >, ≥ and ≤ relations were not used for models' generation.

First, the generation of the 2D molecular structures was performed using the Marvin Sketch software.[21] After, the conversion for 3D format and lowest energy conformers were carried out using OMEGA 2.5.1.4[22,23] followed by correction of ionization state at physiological pH (7.4) using *fixpka* software implemented on QUACPAC 1.6.3.1 package.[24]

The *K*i values were converted to the corresponding p*K*i (−log*K*i) value and used as dependent variables in the hologram quantitative structure-activity relationship (HQSAR) analyses and active/inactive cutoff to kGCN models' generation. It is important to mention that the values of *K*i were retrieved from the literature and were measured under the same experimental conditions, which is considered a fundamental requirement for successful LBDD studies. All the inhibition studies were performed at 20-23 °C with purified recombinant rat FAAH expressed in *Escherichia coli* or with solubilized COS-7 (a cell line derived from monkey kidney cells) membrane extracts from cells transiently transfected with human FAAH cDNA (complementary DNA) and the *K*i of the inhibitions was calculated by the Dixon method.[9-13]

The studied compounds were divided into training and test sets containing 80 and 20%, respectively, of

Table 1. Literature articles: reference, total number of compounds, selected compounds, p*K*i range and general structure

Dataset	Reference	Total number of compounds	Selected compounds	p*K*i range	General structure
1	Boger *et al.*[9]	79	74	4.0-9.8	
2	Ezzili *et al.*[13]	72	67	4.9-9.3	 
3	Hardouin *et al.*[10]	109	101	4.7-9.4	
4	Kimball *et al.*[12]	100	94	4.3-9.7	
5	Romero *et al.*[11]	96	94	5.2-9.5	

p*K*i: negative logarithm of the inhibition constant.

the total number of compounds in each dataset. For this division, MASSA algorithm[25,26] was employed as well as the hierarchical cluster analysis (HCA) performed with the KNIME Analytics Platform,[27] using a specific workflow.[14] The training set compounds were used to construct the HQSAR and knowledge graph convolutional networks (kGCN) models and the test set compounds were used to perform external validations.

Models' generation

The generation of kGCN models was performed according to the protocol reported by Veríssimo *et al.*[16] To classify the substances as active and inactive, the mean p*K*i value of the training set compounds was used as basis. The learning rate values were defined as 0.001, 0.01, 0.1, and 0.3, while the batch size was variated in the interval of 1 to 40, by steps of 5 units (1,5,10,15, …40). The number of epochs was fixed in 1000.

The generation of HQSAR modeling analyses was performed using the Sybyl X 2.1 package.[28] Several parameters were varied, such as hologram length (HL, variable that controls the number of bins in the hologram, ranging from 53 to 401), fragment size (Fsize, parameter that controls the minimum and maximum length of atoms to be included in fragments) and fragment distinction (Fdist, fragments could be composed by atoms (A), bonds (B), connections (C), hydrogen atoms (HA), chirality (Ch), and/or H-bond donor/acceptor groups (DA)). These parameters affect the hologram generation and consequently the statistical evaluation of constructed HQSAR models.

First, all the models applying different combinations of Fdist were generated using default Fsize (4 to 7 atoms) and the 13-default series of HL. Next, the influence of Fsize was further investigated for the three most robust models from the previous step.

Afterwards, the models were validated (internally and externally) and the most robust were used for the contribution map's generation.

Internal and external validations

The kGCN models were evaluated and compared by their AUC (area under the curve, by receiver operating characteristic, ROC curve) cross and external validation values, accuracy (ACC), Matthew's correlation coefficient (MCC), true positive rate (TPR), and F1-score.[29]

The quality of the constructed models was evaluated by internal and external validations. All obtained models in the HQSAR study were generated using the partial least squares (PLS) method and each one was fully cross-validated by the leave-one-out (LOO) method.[30] The external validations consist of the prediction of inhibition activity for test set compounds. Afterwards, metrics for regression-based models were calculated aiming to evaluate model's quality and predictability: r^2m metrics, concordance correlation coefficient (CCC), root mean squared error (RMSE), and mean absolute error (MAE), which guided us in the selection of the most predictive model.[30,31] In summary, LOO cross-validation was employed to select the most robust models and external validations metrics were calculated to select most predictive models. In addition, two other validation procedures were performed only for final selected model. The additional robustness test was performed employing another cross-validation (CV) method with predetermined groups of compounds (from 5 to 70 groups, called leave-many-out or LMO). All CV steps of LMO were carried out in triplicate and the average q^2 (coefficient of determination of the predicted *vs*. experimental values during cross-validation), the mean, and the standard deviation values were also calculated. Finally, the Y-scrambling validation was performed for final selected model. This was employed as a method to verify if the generated model was obtained by chance. In this technique, the response vector (p*K*i) is shuffled multiple times, twenty other data sets were randomly created and a scrambled HQSAR models were generated. Thus, the results were obtained, and it is expected that the scrambled model does not have good statistical metrics (q^2 and standard error of prediction, SEP) as the original HQSAR model.

Applicability domain

For comparison of the models obtained with full dataset (compounds from the 5 literature sources) and dataset 4 (used for final reported HQSAR model), the chemical space was calculated and represented in four different approaches: (*i*) using Morgan fingerprint with 1024 bits and radius equals to 2; (*ii*) AtomPair fingerprint with 1024 bits, minimum and maximum lengths equal to 1 and 30, respectively; (*iii*) molecular access system (MACCS) fingerprint with default parameters; (*iv*) and using drug-like physicochemical properties such as molecular weight (MW), calculated logP (SlogP), number of hydrogen bond donors and acceptors (HBD and HBA, respectively), number of rotatable bonds (nRtB), and fraction of sp^3 carbon atoms (fCsp3). The three first approaches were carried out to represent the structural space because even they represent the same concept, distinct fingerprints result different profile of similarities between ligands.[32] After calculation of fingerprints and properties, each approach was followed by principal component analysis (PCA). The properties from druglike space (*iv*) were normalized before PCA. Lastly, the two first principal components for each analysis were plotted as a representation of the applicability domain (AD) of datasets. All steps involved in AD representation were calculated in KNIME platform.[27] Calculation of fingerprints and properties were performed with RDkit nodes.[33-35]

Molecular interaction fields and ligand sites interaction analyses

First, appropriate 3D structures were chosen by an extensive search on Protein Data Bank (PDB) based on some quality parameters, such as resolution and minimal amount of unmodeled fragments. In the interaction fields and ligand sites experiments, the crystalline complex structure chosen from PDB was 3K7F.[36] This complex is employed with the humanized rFAAH protein, with proper resolution (1.95 Å) and not important unmodelled sequences, that is, the regions with incomplete density or that were not modeled correctly does not interfere with the active site or binding site of the enzyme. Also, the co-crystalized ligand F2C (6-[2-(7-phenylheptanoyl)-1,3 oxazol-5-yl]pyridine-2-carboxylic acid) (Figure 2) is a α-keto heterocyclic derivative (as the dataset compounds studies), which interactions results could corroborate with the simulations.

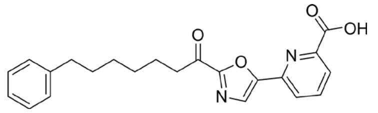


Figure 2. Co-crystalized ligand (F2C) of 3K7F complex.

Also, the complex is composed by the two monomers (A and B) that compound the active FAAH protein, the co-crystalized ligand, and crystallized waters. MIFs and LSI were inferred using the 3D structure of rFAAH (humanized rat's FAAH) and performed by the program AutoGrid,[37,38] FTSite[39-41] and FTMap[39,41,42] servers, respectively.

For the MIF analyses, the AutoGrid from AutoDock software[37,38] was employed to generate 3D maps of potential interaction regions within the inhibitor binding site. The monomers of the selected protein model were separated, the crystallized water molecules were excluded, the hydrogen atoms were added, and the partial atom's charge was calculated. Then, to the simulation, each monomer was submitted individually to the GRID software, and the grid box was (region of exploration with the probes) defined as 60 × 60 × 60 Å, being the centroid defined as the center of mass of the ligand present in the structure. In this study, the selected probes were: hydrophobic, represented by aliphatic carbon (C); hydrogen-bond donor, represented by the donor hydrogen (HD) and hydrogen-bond acceptor, represented by the oxygen acceptor (OA).

The LSI studies were performed using the servers FTSite and FTMap, the first used to calculate the protein cavities as potential interaction sites and the second used to calculate its potential interaction types. The chosen PDB were treated (the co-crystalized ligand excluded) and each monomer was submitted, individually, to the servers. FTsite generated a file containing the sites and their amino acid residues, while the FTMap resulted file contains the probes that interacted with the protein amino acid residues. To calculate and classify the potential interactions, both files were merged, using the PyMOL software[43] version 1.9c and the Discovery Studio Software,[44] which allows for verification of the probes and their interactions with the protein residues. After that, it was possible to estimate what residues participate in most interactions, as well as to propose the potentially important ligand groups.

Molecular docking

The molecular docking study was performed using AutoDock Vina software[38,45,46] and the monomer A of the crystalline complex structure 3K7F.[36] The gridbox site was defined as dimensions of 50 × 50 × 50 Å of the co-crystalized ligand's (F2C) center mass. All experimental water molecules were excluded. The simulation parameters, such as the number of runs, were evaluated by the redocking validation. The software PyMOL was applied to the image generation and visual interpretation of the potential interactions.

After establishing the most appropriate docking protocol, the two most active and inactive compounds (by their p*K*i value) were submitted to the simulation, and their potential interactions were evaluated and compared to the results of SBDD studies.

Results and Discussion

LBDD techniques

Initially, the generation of kGCN models resulted in 32 models using the complete dataset, generated by varying batch size and learning rate values, fixing in 1000 the number of epochs. Every model was internally and externally validated, and its AUC-ROC$_{5\text{-fold}}$ and AUC-ROC$_{ext}$ were compared (Figure 3). The most predictive model was set (learning rate = 0.001 and batch size = 10) and showed adequate accuracy along validation metrics, such as internal and external AUC-ROC values (0.7922 and 0.7722, respectively) (Table 2).

Table 2. Calculated internal and external metrics of the most predictive model (learning rate = 0.001, batch size = 10, and epochs = 1000)

	AUC	ACC	MCC	F1-score	TPR
Internal	0.792	0.746	0.478	0.692	–
External	0.772	0.725	0.448	0.744	0.761

AUC: area under curve; ACC: accuracy; MCC: Matthew's correlation coefficient; F1-score: harmonic mean of the precision and recall; TRP: true positive rate.

In sequence, the HQSAR modeling with the entire dataset resulted in 32 initial models internally validated. None of the models demonstrated acceptable robustness ($q^2 < 0.6$). However, as proof of the unsuitability of the

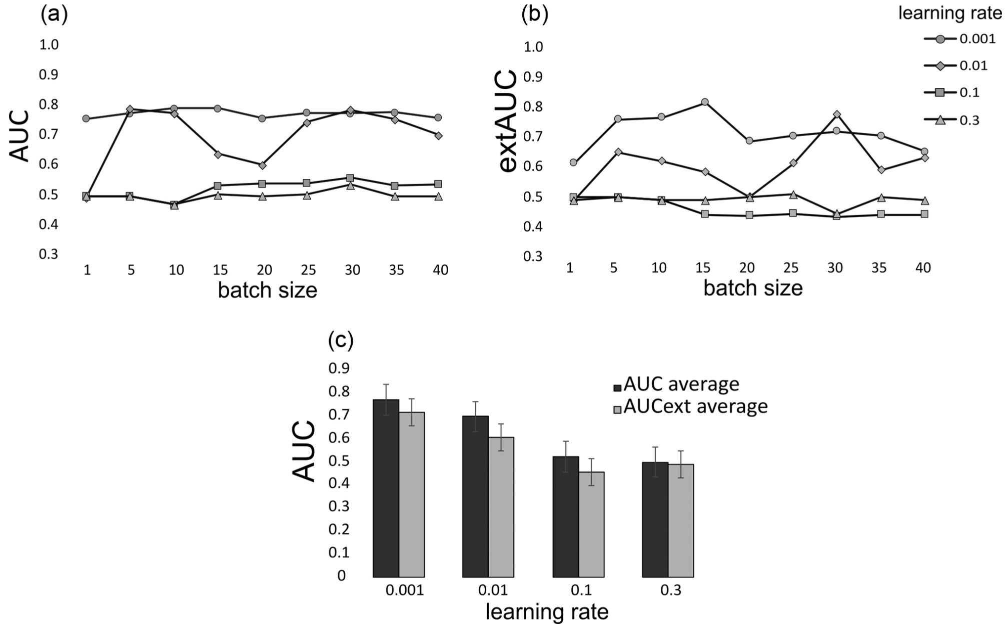


Figure 3. AUC-ROCS values by (a) internal validation; (b) external validation and (c) average of AUC values by learning rate sets.

dataset, the three most robust models were evaluated (q^2 equals 0.538, 0.503, and 0.494) even after variation of fragment length. Still, none of the simulation's parameter changes generated a suitable model using this data set (Table 3) (more information in the Supplementary Information (SI) section, Table S1). These results could be explained due to the fact that kGCN method is considered a deep-learning machine-learning technique and, in theory, the more samples for the learning procedure the better. Furthermore, the kGCN performs a classification task (predicts compounds as active or inactive/less active), the presence of experimental noise will cause lower interference in the predictions.

Following, the full dataset was divided into five smaller datasets (grouping compounds according to the original publications) early mentioned aiming to investigate the suitability of the dataset to model FAAH affinity and further comparison with kGCN results. This procedure would allow the generation of robust individual models. In that sense, all the other data sets generated models with $q^2 > 0.6$, except dataset 5. The highest q^2 values were 0.760 for training set 1, 0.621 for training set 2, 0.696 for training set 3, 0.843 for training set 4 and 0.460 for training set 5 (Table 4). This preliminary result indicates that individual datasets are more suitable for QSAR modeling in this case. Potentially, interference of inter-experimental results and, maybe, inter-laboratory handling generated noise that affected the modelability of the full dataset.[47]

The fragment size of the most robust model of each set was variated to evaluate the influence of this parameter on statistical results. As a result, the most internally robust model was submitted to external validation, using the test data set (details in Tables S2-S4, SI section). The HQSAR final model derived from training set molecules of dataset 4 (fragment distinction A/B; fragment size as 6 to 8 atoms, maximum compound 15 and best length 83, Table 5). Figure 4a shows the distribution of training and test values of p*K*i demonstrating that the model does not have any Y-outlier. It means that the test set compounds were well predicted. Furthermore, the MASSA algorithm used in training/test splitting was designed to avoid the presence of outliers in test set compounds.[25,26]

After choosing the best model and doing the internal and external validation, two additional validations were carried out. The robustness test suggests that the constructed model has acceptable internal consistency since all average q^2 values for each number of cross-validation groups were higher than 0.6 (Figure 4b). After that, the LMO and Y-scrambling validations were performed for the best

Table 3. The three most robust HQSAR models with fragment sizes of 4 to 7 atoms

Fdist	q^2	SEV	r^2	SEE	HL	PC	Size
A,C,Ch	0.494	0.819	0.771	0.551	353	11	4 to 7
A,B,Ch,DA	0.503	0.817	0.823	0.488	401	15	4 to 7
A,B,C,Ch	0.538	0.786	0.828	0.480	307	14	4 to 7

Fdist: fragment distinction; q^2: LOO internal validation coefficient; r^2: no validation calibration coefficient; HL: hologram length; PC: number of PLS principal components; SEV: standard error of validation; SEE: standard error of estimation.

Table 4. Most robust HQSAR models with fragment size of 4 to 7 atoms for each studied dataset

Fdist	Dataset	q^2	SEV	r^2	SEE	HL	PC
A,B,C	1	0.760	0.72	0.924	0.406	59	11
A,C,Ch	2	0.621	0.699	0.757	0.560	59	5
A,HA,Ch	3	0.696	0.678	0.951	0.272	307	11
A, B	4	0.843	0.503	0.962	0.246	401	15
A,HA,Ch	5	0.460	0.759	0.915	0.301	401	9

Fdist: fragment distinction; q^2: LOO internal validation coefficient; r^2: no validation calibration coefficient; HL: hologram length; PC: number of PLS principal components; SEV: standard error of validation; SEE: standard error of estimation.

Table 5. External validation metrics of the most robust HQSAR model of dataset 4

Dataset	Model (Fdist/Fsize)	CCC	r^2m	r^2m'	AVGr^2m	Δr^2m	RMSE
4	A, B / 6 to 8	0.941	0.870	0.813	0.841	0.057	0.410

CCC: concordance correlation coefficient to check the correlation between precision and accuracy; r^2m: Roy's coefficient of predictive potential; r^2m': Roy's coefficient of predictive potential calculated with inverted axis; AVGr^2m: average between r^2m and r^2m'; Δr^2m: difference between r^2m and r^2m'; RMSE: root mean squared error to measure the model's ability to predict.

HQSAR model. When the graph of Y-scrambling validation was plotted (Figure 4c), it could be seen that the difference between the final HQSAR chosen model to the other 20 models had the p*K*i values randomized. The highest q^2 was 0.074 showing that the randomized p*K*i did not create a statistical validated model. Thus, it can be concluded that the HQSAR model was not susceptible to over-fitting and variations in training set (from LMO analysis) and was not obtained by chance (from Y-scrambling analysis).

Therefore, all CV internal validation methods (LOO and LMO) as well as the external validation provided acceptable quality according to the scientific literature[31,32] which indicated that the HQSAR model and their respective fragments information were suitable to predict the inhibition activity.

Although the HQSAR model using only dataset 4 is considered suitable for further studies according to the calculated internal and external validation metrics, it is important to highlight that split our initial full dataset into a partial (only dataset 4) result in a drastic decrease in chemical representativity.[29] This aspect can be noted in Figure 5, the compounds from dataset 4 underrepresent the full dataset comprised with compounds from the five selected articles. As the PCA using three different fingerprints could not be fully reliable to interpret due to the low amount of retained information in the two first principal components (< 30%), a PCA with physicochemical properties was also carried out and dataset 4 represents only a fraction of the entire used chemical space. In this sense, kGCN reported model is more suitable for further predictions since the applicability domain is larger than HQSAR model. Therefore, the use of a validated model for drug design with a broad chemical space coverage is important to avoid false positives in future virtual screening and predictions.[48]

Therefore, both classification and prediction studies models provided good results (according to their validation metrics), which indicated that the HQSAR and kGCN models and their respective fragments information were suitable for predicting and classifying the inhibition activity of compounds. Finally, the contribution maps (from both studies' models) of the two highest and the lowest bioactive molecules were analyzed. Figure 6 shows the structure of these molecules and their representative structural contributions from each model.

When the HQSAR and kGCN maps of compounds **5** and **6** are analyzed, the positive contribution colors demonstrated that the main structure of α-ketooxazole is very important to the inhibition activity. The carbon of the ketone and the oxygen and nitrogen of oxazole seem to be important to the activity in these molecules when these are connected to electronegative groups. By the way, when compound **101** is analyzed on the contribution map of HQSAR, it can be seen that the oxazole lost its importance when it is connected to less electronegative groups. At the same time, on compound **5**, the presence of more electronegative groups increases the importance of the ketone group, which can be explained by the rise

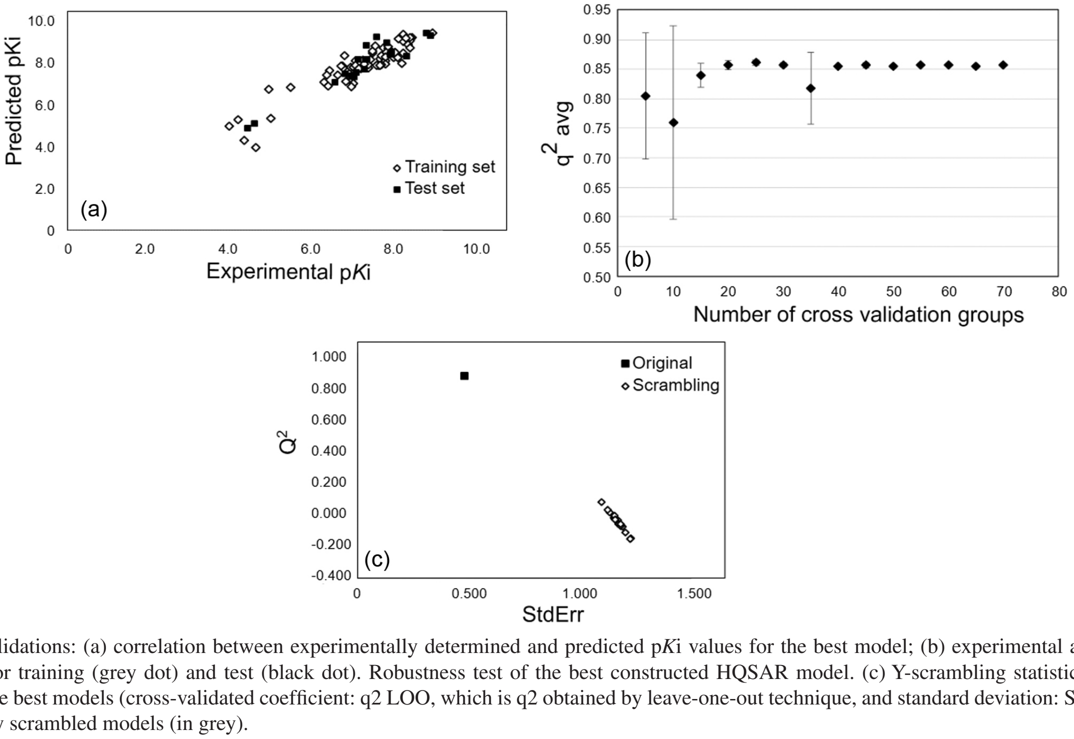


Figure 4. Validations: (a) correlation between experimentally determined and predicted p*K*i values for the best model; (b) experimental and predicted p*K*i values for training (grey dot) and test (black dot). Robustness test of the best constructed HQSAR model. (c) Y-scrambling statistical validation comparing the best models (cross-validated coefficient: q2 LOO, which is q2 obtained by leave-one-out technique, and standard deviation: SEV, in black) and randomly scrambled models (in grey).

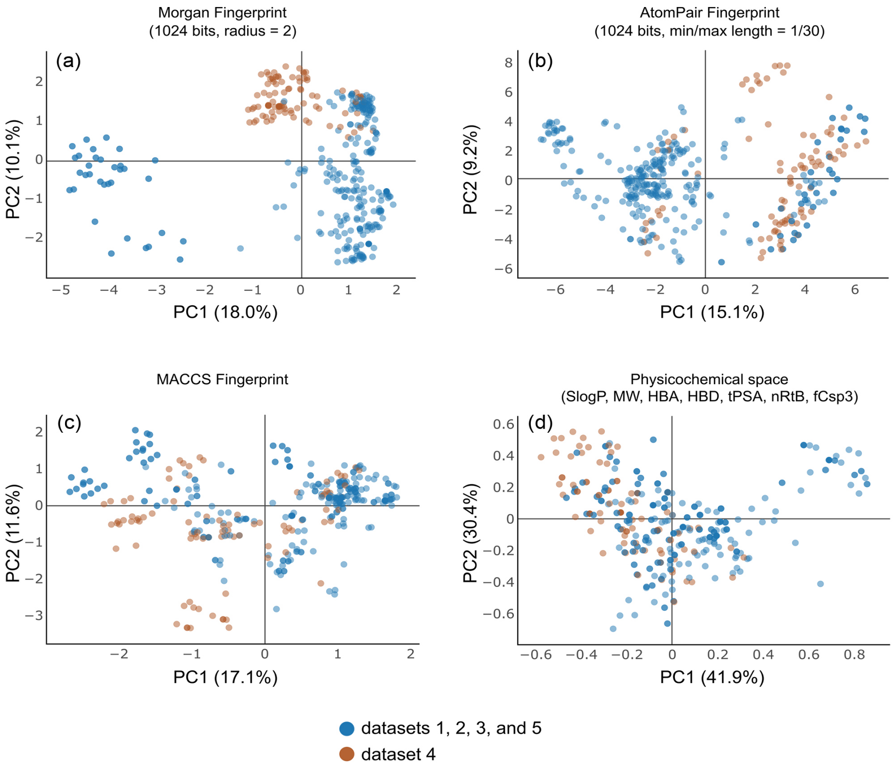


Figure 5. Principal component analysis of chemical space using Morgan fingerprint (a), AtomPair fingerprint (b), MACCS fingerprint (c), and drug-like physicochemical properties (d) of the full dataset. Compounds of datasets 1, 2, 3, and 5 were colored in blue while compounds from dataset 4 were colored in brown.

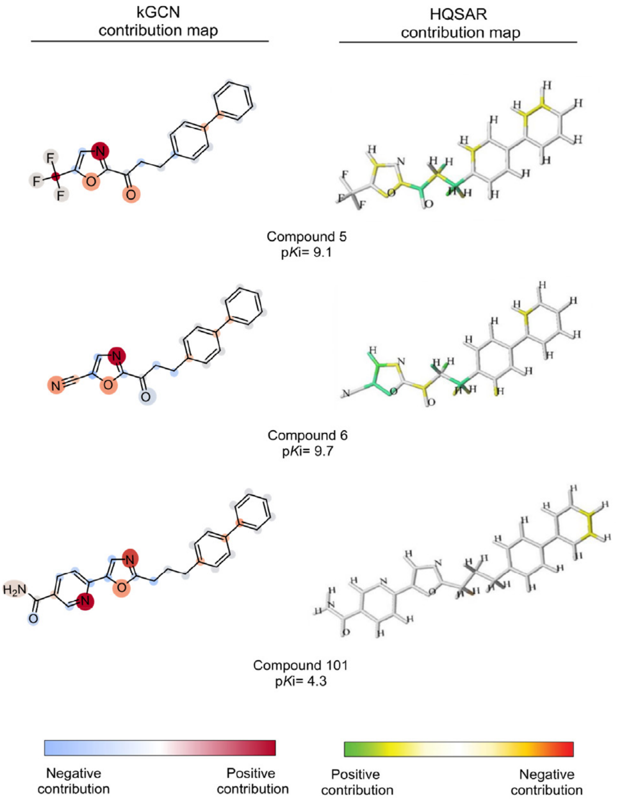


Figure 6. HQSAR and kGCN maps of the most active (compounds **5** and **6** from Kimball *et al.*[12]) and least active (compound **101** from Kimball *et al.*[12]) compounds from dataset 4 and their fragment and atomic contribution.

of its electrophilic character (and the possibility of the well-reported reversible Ser addition to the electrophilic carbonyl, that forms a hemiacetal at the enzyme active site).[11]

The values obtained in the validation were considered acceptable in the literature.[30,31] From the interpretation of the maps, it is possible to evaluate the importance of oxazole and ketone for biological activity.

SBDD studies

For the ligand site evaluation, the generated fields and probes were individually evaluated and counted. The generated probes resulted from FTmap and the sites resulted from FTsite were merged, and only the common probes were evaluated (Figures S1 and S2, SI section).

Both monomers showed minor differences in the distribution by type of interaction (Figure 7a). As an absolute number of interactions with the probes, monomer A formed 125 interactions while monomer B formed 120. Nonetheless, the Van der Waals (VdW) interactions demonstrated the highest frequencies on the found interactions, followed by the hydrogen bonds (HBA and HBD). In fact, the interaction differences could be explained by the recently discovered allosteric inhibition propriety of the FAAH dimer.[47]

The amino acids of FAAH involved in the interactions were also evaluated: the common residues demonstrated a similar frequency of interactions and the Cys269 (19 and 20%), and Val270 (17 and 18%) are the most frequent regarding the total number of interactions. Interestingly, both residues have already been reported as important interaction sites of some FAAH inhibitors.[49] The monomers A and B demonstrated a similar trend of percentual interactions of amino acids (Figure 7b).

The MIF results corroborate with the analysis of the SBDD studies about the importance of the ketone group. As can be seen in Figures 7c and 7d, the HBA MIFs (OA probes) overlap the oxygen atom of the ketone in both monomers. This explains the potential hydrogen bonds between the amino acids of the catalytic region (Ser241 and Ser217) and the inhibitor, observed by the frequency of interactions on the ligand site studies.

The acid group of the polar head of the inhibitor F2C is also overlapped by the OA and HD probes (Figures 7e and 7f), showing its interesting potential hydrogen bonds between the Cys269 and Val270 residues and the inhibitor. In this case, the inhibitor acts as a hydrogen bond acceptor.

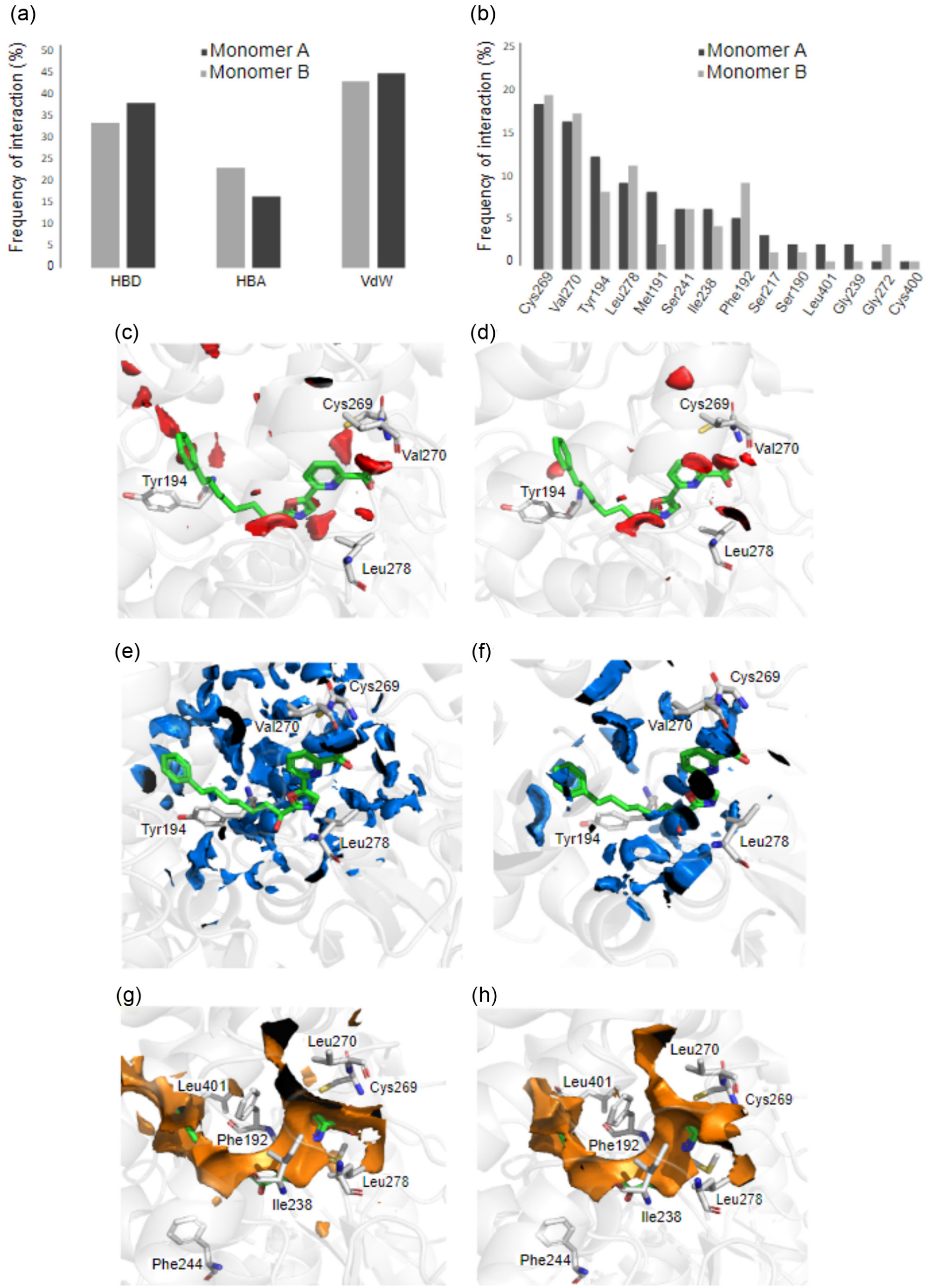
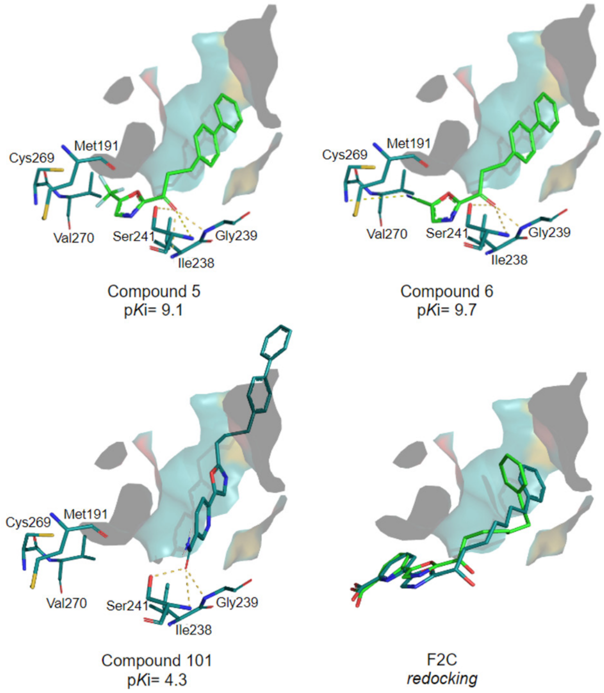


Figure 7. Interaction frequencies between the two FAAH's monomers by interaction types (a) and amino acid residues (b). Generated hotspots on MIF studies: OA probes of chain A (c) and chain B (d); HD probes of chain A (e) and chain B (f); C probes of chain A (g) and chain B (h).

The hydrophobic MIF (probe C) overlays most parts of the inhibitor structure, except the ketone group, demonstrating the hydrophobic nature of the active site of the FAAH, and explaining the high frequency of Van der Waals interactions found in the exploration of hotspots (Figures 7g and 7h).

The molecular docking experiments reinforce the MIF studies results. The docking protocol was validated by redocking simulation and its RMSD value equal to 1.58 Å. The cysteine (Cys269) and valine (Val270) are involved in interactions with the active compounds, but not with the inactive ones (Figure 8). The lateral carbonic chain of the compounds is all involved in hydrophobic interactions with the ACP region of FAAH. The ketone group, present in both active compounds, interacts with the well-pointed hydrogen donors Ser271, Ile238, and Gly239. Otherwise, the inactive compound can be observed outside of the active site, which decreases its interactions with the mentioned residues.



Compound 5
p*Ki*= 9.1

Compound 6
p*Ki*= 9.7

Compound 101
p*Ki*= 4.3

F2C
redocking

Figure 8. Potential interactions observed in docking simulations.

In conclusion, when the SBDD and the LBDD studies are compared, it is possible to correlate their results (Figure 9). The oxazol-ketone group, appointed as essential to the inhibition activity by HQSAR and kGCN best models, is also an HBA, which it can be inferred that fits the HBD region suggested by MIF studies and confirmed by docking analyses. Also, the oxazole ring demonstrated a positive contribution to the FAAH's inhibition by the HQSAR, while the MIF appointed polar groups substituents as ideal to fit on the HBD/OA regions of the enzyme. On the other hand, MIF studies indicate that apolar groups on the lateral chain of inhibitor compounds could fit in the hydrophobic pockets of FAAH, interacting with its apolar residues, confirmed by the potential interactions observed in the docking study.

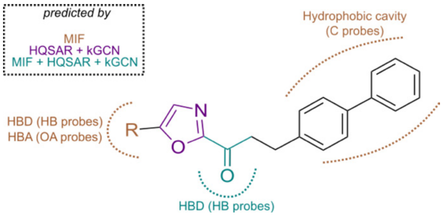


Figure 9. Compared LBDD and SBDD studies results.

The integration of LBDD and SBDD to FAAH shows the complementary structural information of the potential interaction compounds of each one. After that, with the results of the developed studies, it is possible to design potential FAAH inhibitors for future drug design campaigns.

Conclusions

Firstly, graph-based classification models were suitable to generate robust models for datasets with compounds from different sources in contrast to PLS regression. In that sense, the most robust classification (kGCN) model (learning rate = 0.001 and batch size = 10) was constructed using the full data set and presented high robustness and predictability with AUC-ROC values equal to 0.792 and 0.772 for internal and external validations, respectively. The constructed predictive model (HQSAR) using the full α-ketoheterocycle dataset did not have good internal consistency and external predictivity, but the models built using the five sets, individually, had. Dataset 4 gave the best HQSAR model (fragment distinction A/B; fragment size 6-8, maximum compound 15, and best length 83). The quality of the best models concerning internal and external predictiveness was evaluated by statistical parameters, such as leave-one-out cross-validation q^2 (0.857) and quality of test set predictions CCC (0.941) to the HQSAR model. All the validation metrics' values calculated were considered acceptable in the literature. Additionally, the employment of kGCN was important to model a dataset containing compounds from different sources due to the superior ability of data generalization and, therefore, increasing the coverage of chemical space. From the interpretation of the maps, it is possible to evaluate the importance of the oxazole and the ketone for the activity. The MIF and LSI studies were carried out using the monomers of the FAAH enzyme and evaluated. The absolute number of interactions was different among the monomers A and B (125 and 120, respectively), but the more frequent amino acid residues were similar (Cys269 and Val270) and compatible with the data already described in the literature. Furthermore, the interpretation of kGCN was corroborated by HQSAR (a widely employed QSAR technique) and other SBDD methods highlighting the suitability of graph-based classification algorithms in the drug design field. The studies proved the importance of the oxazole-ketone group of FAAH inhibitors, such as the hydrophobic nature of the ligand site of the enzyme. The development of prediction and classification models and potential FAAH hotspots provided complementary information about the structural proprieties of ketoheterocycle FAAH inhibitors.

Supplementary Information

Experimental protocols and detailed data are available free of charge at http://jbcs.sbq.org.br as PDF file.

Acknowledgments

The authors would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, grant number RED-00110-23) for the financial support. We also like to thank OpenEye Scientific Software for OMEGA and QUACPAC academic licenses.

Author Contributions

PALS was responsible for data curation, writing original drafts and editing; MSMR, GCV, and ACGT for data curation; RBO and VGM for the conceptualization, funding acquisition, resources, writing review and editing.

References

1. Francischetti, E. A.; de Abreu V. G.; *Arq. Bras. Cardiol*. **2006**, *87*, 548. [Crossref]
2. Ahn, K.; Johnson, D. S.; Cravatt, B. F.; *Expert Opin. Drug. Discovery* **2009**, *4*, 763. [Crossref]
3. Tuo, W.; Leleu-Chavain, N.; Spencer, J.; Sansook, S.; Millet, R.; Chavatte, P.; *J. Med. Chem.* **2016**, *60*, 4. [Crossref]
4. Otrubova, K.; Boger, D. L.; *ACS Chem. Neurosci*. **2012**, *3*, 340. [Crossref]
5. National Library of Medicine, *A Study to Investigate Whether PF-04457845 is Effective in Treating Pain, is Safe and Tolerable in Patients with Osteoarthritis of the Knee,* https://clinicaltrials.gov/study/NCT00981357?term=PF-04457845&rank=1, accessed in June 2024.
6. National Library of Medicine, *FAAH Inhibitor Trial for Adults with Tourette Syndrome*, https://clinicaltrials.gov/study/NCT02134080?term=FAAH&rank=2&a=6, accessed in June 2024.
7. National Library of Medicine, *Evaluation Study of New Compounds With Potential Use in Schizophrenia (EICAS),* https://clinicaltrials.gov/study/NCT00916201?term=FAAH%20inhibitor&rank=9, accessed in June 2024.
8. Tripathi, R. K. P.; *Eur. J. Med. Chem*. **2020**, *188*, 111953. [Crossref]
9. Boger, D. L.; Miyauchi, H.; Du, W.; Hardouin, C.; Fecik, R.A.; Cheng, H.; Hwang, I.; Hedrick, M.P.; Leung, D.; Acevedo, O.; Guimarães C. R. W; Jorgensen, W. L.; Cravatt, B. F.; *J. Med. Chem.* **2004**, *48*, 1849. [Crossref]
10. Hardouin, C.; Kelso, M. J.; Romero, F. A.; Rayl, T. J.; Leung, D.; Hwang, I.; Cravatt, B. F.; Boger, D. L.; *J. Med. Chem*. **2007**, *50*, 3359. [Crossref]
11. Romero, F.A.; Du, W.; Hwang, I.; Rayl, T. J.; Kimball, F.S.; Leung, D.; Hoover H. S.; Apodaca, R. L.; Breitenbucher, J. G.; Cravatt, B. F.; Boger, D. L.; *J. Med. Chem*. **2007**, *50*, 1058. [Crossref]
12. Kimball, F. S.; Romero, F. A.; Ezzili, C.; Garfunkle, J.; Rayl, T. J.; Hochstatter, D. G.; Hwang, I.; Boger, D. L.; *J. Med. Chem*. **2008**, *51*, 937. [Crossref]
13. Ezzili, C.; Mileni, M.; McGlinchey, N.; Long, J. Z.; Kinsey, S. G.; Hochstatter, D. G.; Stevens, R. C.; Lichtman, A. H.; Cravatt, B. F.; Bilsky, E. J.; Boger, D. L.; *J. Med. Chem*. **2011**, *54*, 2805. [Crossref]
14. Veríssimo, G. C.; Dutra, E. F. M.; Dias, A. L. T.; Fernandes, P. O.; Kronenberger, T.; Gomes, M. A.; Maltarollo, V. G.; *J. Mol. Graphics* **2019**, *90*, 180. [Crossref]
15. Abdizadeh, R.; Hadizadeh, F.; Abdizadeh, T.; *J. Mol. Struct.* **2020,** *1199*, 126961. [Crossref]
16. Veríssimo, G. C.; dos Santos Junior, V. S.; Fernandes, P. O.; Ishida, S.; Kojima, R.; Okuno, Y.; Getrudes, J. C.; Maltarollo, V. G.; *Int. J. Quant. Struct.-Prop. Relatat*. **2022**, *7*, 1. [Crossref]
17. Hall, D. R.; Ngan, C. H.; Zerbe, B. S.; Kozakov, D.; Vajda, S.; *J. Chem. Inf. Model.* **2011**, *52*, 199. [Crossref]
18. Kozakov, D.; Grove, L.E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S.; *Nat. Protoc*. **2015**, *10*, 733. [Crossref]
19. Myllymäki, M.; Käsnänen, H.; Kataja, A. O.; Lahtela-Kakkonen, M.; Saario, S. M.; Poso, A.; Koskinen A. M. P.; *Eur. J. Med. Chem*. **2009**, *44*, 4179. [Crossref]
20. Kojima, R.; Ishida, S.; Ohta, M.; Iwata, H.; Honma, T.; Okuno, Y.; *J Cheminf*. **2020**, *12*, 32. [Crossref]
21. *MarvinSketch*, version 5.12.1; ChemAxon, Cambridge, 2013.
22. *OMEGA*, version 2.5.1.4; OpenEye, Santa Fe, 2013.
23. Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T.; *J. Chem. Inf. Model*. **2010**, *50*, 572. [Crossref]
24. *QUACPAC*, version 2.2.2.0; OpenEye, Santa Fe, 2023.

25. MASSA Algorithm, https://github.com/gcverissimo/MASSA_Algorithm, accessed in June 2024.
26. Veríssimo, G. C.; Panteleão, S. Q.; Gertrudes, J. C.; Kronenberger, T.; Honório, K. M.; Maltarollo, V. G.; *ChemRxiv*, 2023. [Link] accessed in June 2024
27. Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel T. R.; Kötter, T.; Meinl, P. O. T.; Sieb, C.; Thiel, K.; Wiswedel, B.; *KNIME*; Berlin, Germany, 2007.
28. *Sybyl-X suite*, version 2.1; Tripos Inc, Saint Louis, 2013.
29. Roy, K.; Kar, S.; Das, R. N.; *Springer Briefs in Molecular Science*; Springer: Berlin, 2015.
30. Gramatica, P.; Sangion, A.; *J. Chem. Inf. Model*. **2016**, *56*, 1127. [Crossref]
31. Tropsha, A.; *Mol. Inform.* **2010**, *29*, 476. [Crossref]
32. Webb, G. I.; Sammut, C.; Perlich, C.; Horváth, T.; Wrobel, S.; Korb, K. B. In *Encyclopedia of Machine Learning*; Springer Science & Business Media: Boston, 2011, p. 600. [Link] accessed in June 2024
33. Marques, D. P. A.; Andrade, L. F.; Reis, E. V. S.; Clarindo, F. A.; Moraes, T. F. S.; Lourenço, K. L.; De Barros, W. A.; Costa, N. E. M.; Andrade, L. M.; Lopes-Ribeiro, Á.; Maciel, M.; Corrêa-Dias, L. C.; de Almeida, I. N.; Arantes, T. S.; Litwinski, V. C. V.; de Oliveira, L. C.; Serafim, M. S. M.; Maltarollo, V. G.; Guatimosim, S.; Coelho-Dos-Reis, J. G. A.; *Virus Res*. **2024**, *340*, 199291. [Crossref]
34. Landrum, G.; Palmer, A.; Davies, J.; Berthold, M.; https://oak.novartis.com/id/eprint/3826, accessed in June 2024.
35. Mazanetz, M. P.; Marmon, R. J.; Reisser, C. B. T.; Morao, I.; *Curr. Med. Chem.* **2012**, *12*, 1965. [Crossref]
36. Mileni, M.; Garfunkle, J.; Ezzili, C.; Kimball, F. S.; Cravatt, B. F.; Stevens, R. C.; Boger D. L.; *J. Med. Chem.* **2010**, *53*, 230. [Crossref]
37. Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J.; *J. Comput. Chem*. **2009**, *30*, 785. [Crossref]
38. *AutoDock*, version 4.0, Scripps Research, USA, 2014.
39. Brenke, R.; Kozakov, D.; Chuang, G. Y.; Beglov, D.; Hall, D.; Landon, M.R.; Mattos, C.; Vajda, S.; *J. Bioinform*. **2009**, *25*, 621. [Crossref]
40. Ngan, C. H.; Hall, D. R.; Zerbe, B.; Grove, L. E.; Kozakov, D.; Vajda, S.; *J. Bioinform.* **2011**, *28*, 286. [Crossref]
41. Kozakov, D.; Hall, D.R.; Chuang, G. Y.; Cencic, R.; Brenke, R.; Grove, L. E.; Beglov, D.; Pelletier, J.; Whitnny, A.; Vadja, S.; *Proc. Natl. Acad. Sci*. **2011**, *108*, 13528. [Crossref]
42. Bohnuud, T.; Beglov, D.; Ngan, C. H.; Zerbe, B.; Hall, D. R.; Brenke, R.; Vadja, S.; Frank-Kamenetskii, M. D.; Kozakov, D.; *Nucleic. Acids Res.* **2012**, *40*, 7644. [Crossref]
43. DeLano, W. L.; *PyMOL*, version 1.9c, Schrödinger, 2020.
44. *Discovery Studio Visualizer*, version 2020; BIOVIA, Dassault Systèmes, San Diego, 2020.
45. Trott, O.; Olson, A. J.; *J. Comput. Chem*. **2010**, *31*, 455. [Crossref]
46. Cronin, M. T. D.; Schultz, T. W.; *J. Mol. Struct.* **2003**, *622*, 39. [Crossref]
47. Dainese, E.; Oddi, S.; Simonetti, M.; Sabatucci, A.; Angelucci, C.B.; Ballone A.; Dufrusine, F. F.; Fabritiis, G.; Maccarrone, M.; *Sci. Rep*. **2020**, *10*, 2292. [Crossref]
48. Serafim, M. S. M.; Pantaleão, S. Q.; da Silva, E. B.; McKerrow, J. H.; O'Donoghue, A. J.; Mota, B. E. F.; Honório, K. M.; Maltarollo, V. G.; *Front. Drug Discovery* **2023**, *3*, 1237655. [Crossref]
49. Criscuolo, E.; De Sciscio, M. L.; Fezza, F.; Maccarrone, M.; *Molecules* **2020**, *26*, 48. [Crossref]

Submitted: February 3, 2024
Published online: July 5, 2024