Revista Brasileira de Computação Aplicada, April, 2026

DOI: 10.5335/rbca.v18i1.16853
Vol. 18, Nº 1, pp. 23–32
Homepage: seer.upf.br/index.php/rbca/index

ORIGINAL PAPER

QSAR Modeling of Analgesic Cannabinoids via Dual Feature Selection and Support Vector Machines: A Cheminformatics Case Study

Rosalvo Ferreira de Oliveira Neto[1], Edilson Beserra de Alencar Filho[2], Allysson Layr dos Santos Ferreira[2], Daniel Alencar Penha Carvalho[1]

[1]Universidade Federal do Vale do São Francisco (Univasf), Juazeiro, Bahia, Brasil, [2]Programa de Pós-Graduação em Biociências, Universidade Federal do Vale do São Francisco (Univasf), Petrolina, Pernambuco, Brasil

*rosalvo.oliveira@univasf.edu.br †edilson.beserra@univasf.edu.br †allysson.layr@discente.univasf.edu.br
*daniel.apcarvalho@discente.univasf.edu.br

Received: yyyy-mm-dd. **Revised:** yyyy-mm-dd. **Accepted:** yyyy-mm-dd.

Abstract

This study aimed to develop an accessible Quantitative Structure–Activity Relationship model based on Machine Learning techniques for a set of analgesic cannabinoid compounds. It represents a cheminformatics contribution to the promising field of endocannabinoid system modulation. Three-dimensional molecular structures and biological activity data were retrieved from PubChem. Molecular descriptors were calculated using Dragon7 software and subjected to variable selection through two complementary feature selection strategies: Wrapper and Filter methods. The final predictive model was constructed using Support Vector Machines and validated through both K-fold cross-validation and leave-one-out approaches. A robust and predictive model comprising 29 molecular descriptors was obtained, enabling the prediction of novel thiophenyl-acetamide analogs. The combined use of two feature selection techniques proved effective in capturing relevant molecular information and enhancing model performance. This model offers valuable support for the synthetic optimization and design of more potent cannabinoid-based analgesics.

Keywords: QSAR modeling; Feature selection; Support Vector Machines; Cannabinoid receptor CB2.

Resumo

Este estudo teve como objetivo desenvolver um modelo acessível de Relação Quantitativa Estrutura–Atividade, baseado em técnicas de Aprendizado de Máquina, para um conjunto de compostos canabinoides com atividade analgésica. Trata-se de uma contribuição da quimioinformática para o promissor campo da modulação do sistema endocanabinoide. Estruturas moleculares tridimensionais e dados de atividade biológica foram obtidos do banco de dados PubChem. Descritores moleculares foram calculados utilizando o software Dragon7 e submetidos à seleção de variáveis por meio de duas estratégias complementares: os métodos Wrapper e Filter. O modelo preditivo final foi construído com Máquinas de Vetores de Suporte e validado por meio de validação cruzada K-fold e validação leave-one-out. Obteve-se um modelo robusto e preditivo, composto por 29 descritores moleculares, capaz de prever a atividade de novos análogos da classe dos tiofenil-acetamidas. A combinação das duas técnicas de seleção de variáveis mostrou-se eficaz na captura de informações moleculares relevantes e na melhoria do desempenho do modelo. Esse modelo oferece suporte valioso para a otimização sintética e o desenho de compostos canabinoides com maior atividade analgésica.

Palavras-Chave: Modelagem QSAR; Seleção de variáveis; Máquinas de Vetores de Suporte; Receptor canabinoide CB2.

1 Introduction

Pain is an important clinical condition, which can be modulated through the use of analgesics and anesthetics (Kirkpatrick et al., 2015). Cannabinoid receptors have demonstrated an important role in mechanisms regulating the pain process (Thur et al., 2012). These are basically represented by two G protein-coupled receptors (GPCRs), CB(1) and CB(2), although additional receptors may be involved (Mackie, 2008). CB(1) receptors are present in more abundant quantities in the central nervous system, and the psychoactive effects of cannabinoid ligands (endo or exo) are attributed to them. CB(2) was attributed with an almost exclusive distribution at the peripheral level, but it is now known that, in addition to the periphery, they are also present to an important extent in the CNS (Thur et al., 2012; Onaivi et al., 2006; Van Sickle et al., 2005). Considering the potential of CB(1) receptors in generating addiction, investigations into the analgesic effect associated with cannabinoid receptors have turned their attention to the CB(2) subtype, searching for compounds with agonist selectivity (Thur et al., 2012; Murineddu et al., 2012; Riether, 2012; Han et al., 2009).

An extensive series of 2-amino-3-carboethoxy thiophene derivatives with analgesic potential by CB(2) modulation was obtained by (Thur et al., 2012). In this paper, a detailed qualitative structure-activity analysis was made, promoting the replacement of functional groups with different characteristics (alkyl or electronegative groups, electron donors or acceptors, among others), observing the agonist profile as well as CB(1) or CB(2) selectivity.

Quantitative Structure-Activity Relationship (QSAR) modeling is one of the major computational tools employed in medicinal chemistry scenario (Cherkasov et al., 2014), allowing not only the understanding of the molecular reasons of bioactive compounds but also providing models capable of predicting the range of biological effects presented by a new compound. Starting as an extension of organic physical chemistry, the QSAR area has expanded, from linearized regression models applied to small congeneric series of molecules to the use of machine learning algorithms, capable of processing large amounts of data (samples and molecular descriptors) (Cherkasov et al., 2014).

Among the machine learning algorithms available, the K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) are popular in small sample problems. KNN algorithm is an instance-based method, meaning the model is represented by the training examples. The value of K represents the number of similar examples from the training set used for prediction. Given an unknown example, KNN searches for the K most similar examples to the unknown pattern to make a prediction, calculating the average of the target values of these K examples (Cover and Hart, 1967). SVM algorithm is a technique that employs a different induction principle called Structural Risk Minimization (Deng and Tian, 2004). While many machine learning methods use only the training set to minimize classification error by adjusting their free parameters, SVMs consider not only the minimization of empirical risk (the training set error) but also the complexity, or capacity, of the classifier. The goal of this principle is to prevent the phenomenon known as overfitting, which occurs when a classifier becomes overly specialized in the training set and has low generalization power. SVMs incorporate a capacity term into the empirical error through a variable called the VC dimension, which measures the model's capacity. The higher its value, the more complex the classification functions that can be induced. The more complex the classification function, the greater the discriminative power on the training set and the higher the risk of overfitting. Therefore, SVM aims to balance these two aspects by utilizing the principle of Structural Risk Minimization.

Building machine learning models from high-dimensional data in small samples is a challenge. One common approach to overcome this issue is applying Feature Selection, one essential step in data preprocessing, whose goal is finding a feature subset for effective supervised learning (Jovic et al., 2015). There are two approaches for Feature Selection: Wrapper and Filter. A Wrapper method is a Feature Selection technique that uses a machine learning algorithm to evaluate various feature combinations and a search strategy to identify the optimal subset. On the other hand, the Filter method selects variables without using a machine learning algorithm; it employs metrics that are independent of the machine learning algorithm and combines these metrics with a search strategy to identify the optimal subset (Hall and Witten, 2011). Among the search algorithms that can be used for feature selection, we can highlight the Ant Colony (AC) algorithm, as your turn, is a bio-inspired global optimization algorithm (Dorigo and Di Caro, 1999). It is a search algorithm based on the way ants communicate indirectly while foraging. This indirect communication occurs through pheromone trails that are laid down as ants search for food. The AC algorithm uses graphs to represent the problem, with paths in the graphs representing possible solutions. Initially, the population size of ants for the algorithm is defined. Each ant is randomly placed at a node of the graph. The ant will choose the path in the graph according to the probability of each vertex. The probabilities associated with each vertex are updated by the pheromone rate, which evaluates how good that solution is according to the fitness function. Thus, new ants are more likely to choose the best path found. However, the algorithm also incorporates a pheromone evaporation rate, which decreases the probability of selecting a path over iterations. In the context of Feature Selection, each node in the graph represents a variable.

The goal of this work was to obtain QSAR modeling involving a set of synthetic cannabinoid compounds previously reported in the literature, using machine learning algorithms, making the generated model available on an online platform, for direct access by the scientific community. In this sense, we aim to contribute to the rational planning of compounds with analgesic properties considering a recently explored and promising area, the endocannabinoid system.

2 Related work

Research in QSAR modeling applied to cannabinoids has significantly progressed in recent years. Four relevant studies in this domain are highlighted below:

In Oliveira Neto et al. (2025), the authors combined ligand-based QSAR modeling with structure-based Molecular Dynamics (MD) to rationalize and optimize a series of imidazo[1,2-a]pyridine/imidazo[1,2-a]pyrazine derivatives with anti-melanoma activity. Starting from a curated set of 117 compounds with experimental IC_{50} values, the study generated a high-dimensional descriptor matrix and adopted a two-step wrapper feature-selection pipeline, in which a global bio-inspired search (Artificial Bee Colony) was followed by a local graph-based refinement (Best-First Algorithm), both guided by Random Forest performance. The resulting reduced descriptor subset supported a robust predictive model validated by Leave-One-Out and repeated K-fold procedures, complemented by Y-scrambling to assess chance correlation. Beyond prediction, the model was used to propose nine new analogues with favourable predicted potency, and MD simulations (GROMACS/GROMOS) were employed to explore a plausible mechanism of action via Aurora Kinase inhibition, motivated by the presence of a co-crystallographic ligand sharing the key structural core and substitution pattern of the most active analogues.

In Lee et al. (2020), the authors developed a QSAR model to predict the CB1 receptor binding affinity and potential for dependence of synthetic cannabinoids. Using experimental binding data from 15 synthesized compounds, they calculated molecular descriptors via the R/CDK toolkit and constructed regression models using partial least squares regression (PLSR). The final model showed strong predictive performance, validated through Y-randomization and external test sets. Notably, the predicted CB1 affinities correlated well with known in vivo dependence indicators, such as conditioned place preference and self-administration.

In Keimowitz et al. (2000), the authors investigated the structure-activity relationship of 36 Δ8-THC analogues to elucidate how side-chain conformations influence CB1 receptor binding affinity and pharmacological potency. By applying multiple QSAR approaches—including a modified active analogue method, multiple linear regression (MLR), and CoMFA—the study consistently demonstrated that side-chain length and conformational flexibility play critical roles in receptor interaction. Increased affinity was associated with longer side chains capable of folding back toward the aromatic ring, rather than extending linearly.

In Ferreira et al. (2009), the authors developed an QSAR model for a series of novel cannabinoid derivatives by employing descriptors derived from semi-empirical quantum-chemical (PM3) calculations. Focusing on CB2 receptor binding, the study correlated electronic and structural properties—such as dipole moment components, polarizability, ovality, and aqueous heat of formation—with experimental Ki values. The final multiple linear regression model exhibited high predictive accuracy (R^2 = 0.775) and indicated that ligand–receptor interactions are strongly influenced by electronic features, suggesting the presence of a significant electrostatic field within the CB2 binding pocket. These findings highlight the value of quantum-chemical descriptors in elucidating structure–activity relationships and guiding the design of CB2-selective cannabinoid ligands.

In Vázquez-Valadez et al. (2023), the authors employed the OECD QSAR Toolbox to develop in silico models predicting acute and chronic toxicity parameters—such as LD50, LOAEL, NOAEL, LOEL, and NOEL—as well as carcinogenicity and mutagenicity for key phytocannabinoids from Cannabis sativa. The models demonstrated high predictive performance (e.g., $R^2 > 0.9$ for most endpoints) and enabled estimation of permissible daily exposure (PDE) and acceptable daily intake (ADI) values based on predicted NOAELs. These predictions provide a regulatory-toxicological framework to evaluate the safety of cannabinoids without animal testing. Notably, all compounds tested were negative for genotoxic and carcinogenic potential.

Despite the relevance and methodological rigor of the aforementioned studies, the present work distinguishes itself in several key aspects. While previous QSAR studies (Lee et al., 2020; Keimowitz et al., 2000; Ferreira et al., 2009) have predominantly focused on cannabinoids targeting the CB1 receptor, often associated with psychoactive effects and abuse potential, or broadly on Cannabis sativa constituents for toxicological profiling (Vázquez-Valadez et al., 2023), our research centers on synthetic thiophenyl-acetamide derivatives with selective CB2 receptor affinity, a less explored but therapeutically promising class. Methodologically, this work introduces a hybrid variable selection strategy combining Wrapper (with Ant Colony Optimization) and Filter (with Best-First Search), enhancing feature relevance and model robustness. The final QSAR model, developed using Support Vector Machines, was validated through K-fold, Leave-One-Out, and Y-scrambling techniques, ensuring predictive reliability, following the validation methodology adopted in Oliveira Neto et al. (2025). The trained QSAR model and the scripts required for inference are publicly available in an open-source GitHub repository, enabling reproducibility and facilitating practical use by researchers working on cannabinoid drug discovery Oliveira Neto (2026).

3 Methods

3.1 Database and calculation of molecular descriptors

The dataset used in this paper corresponds to a series of sixty five 2-arylamido-5,7-dihydro-4H-thieno[2,3-c]pyran-3-carboxamide derivatives as cannabinoid receptor agonists, with varied affinity for the CB2 (Type 2 Cannabinoid Receptor), as illustrated by the representative compound 3t (Figure 1), expressed as the concentration required to interact with 50% of targets (EC50) (Thur et al., 2012). These data are available on PubChem (Kim et al., 2022) database (https://pubchem.ncbi.nlm.nih.gov/) under the code AID_717398. Initially, the EC50 values for the 65 compounds

were converted to pEC50 - Log(1/EC50) - to reduce the standard deviation and for the highest values to correspond to the most active compounds.

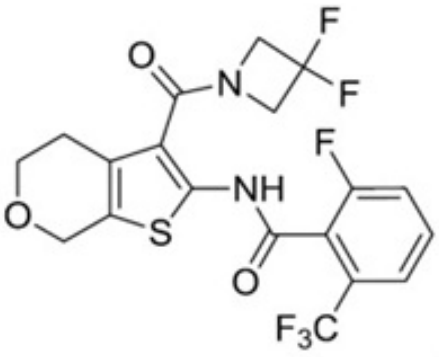

Figure 1: Compound 3t, representative of the molecular series used in this study (adapted from Thur et al. (2012)).

The 3D structures of the compounds were obtained directly from the PubChem (Kim et al., 2022) database in the .mol format. These representations were computed by PubChem using a specific algorithm that optimize the geometries at MMFF94s force field (Bolton et al., 2011; Kim et al., 2013).

An initial matrix of 5,270 molecular descriptors (features) was obtained by Dragon 7 software, which is able to generate features including the simplest atom types, functional groups and fragment counts, topological and geometrical descriptors, three-dimensional descriptors, among others (Talete s.r.l., 2017). The use of thousands of descriptors that capture diverse molecular characteristics applies well to Machine Learning methods, which have their performance optimized in the combination of subtle characteristics of different descriptors, expressing their maximum predictability power, in this case, of potentially more active compounds.

It is important to emphasize that the compounds used in this study belong to a structurally related series of 2-arylamido-5,7-dihydro-4H-thieno[2,3-c]pyran-3-carboxamide derivatives, forming a relatively homogeneous chemical space. In QSAR modeling, the use of congeneric series is a common strategy in medicinal chemistry, as it allows the identification of subtle structure–activity relationships within a defined scaffold. Consequently, the predictive domain of the proposed model should be interpreted within this chemical space. Although the methodology presented here may be extended to other cannabinoid-related compounds, predictions for structurally distant molecules, such as phytocannabinoids (e.g., THC or CBD) or other synthetic cannabinoid scaffolds, should be treated with caution unless the model is retrained with a broader and more chemically diverse dataset.

3.2 Variable selection and Machine Learning algorithms

In this paper, we employed a hybrid machine learning solution using two Feature Selection approaches: Wrapper and Filter, as illustrated in Fig. 2. The first step used the Wrapper method combined KNN with the Ant Colony (AC) algorithm as the search method. Particularly, KNN was chosen because it has few free parameters and is more suitable for datasets with small samples.

Step 02 utilized the Filter approach with Correlation-based Feature Subsets (Hall, 1999) as the metric for evaluating subset variables. As a Filter method, it operates independently of the machine learning algorithm. This metric assesses the value of a subset of attributes by considering both the individual predictive ability of each feature and the degree of redundancy among them. Since it evaluates subsets of variables, a heuristic search algorithm is required to identify the best subset. We employed the Best First Algorithm (BFA) as the search strategy, which is a graph-based local search algorithm (Hart et al., 1968).

The hybrid Machine Learning solution using two Feature Selection approaches was employed to overcome three issues that would arise if each approach were used alone: 1) Using a Machine Learning algorithm without Feature Selection for this dataset generates a low-performance solution because the number of features greatly exceeds the number of samples, making convergence difficult; this problem is known in the literature as the curse of dimensionality (Brown et al., 1995); 2) Using only the first step with a global optimization algorithm such as AC on a vast search space with a small sample and correlated features makes fine-tuning difficult due to the stochastic nature of the algorithm, so we use the first step as a pre-filter for Feature Selection; 3) Using only the second step with a local search algorithm such as BFA yields poor-performing solutions because the algorithm gets stuck in local optima at the beginning of the search, so we use it after an initial selection with a global search algorithm.

The final model was built using the twenty-nine features selected after the two steps and utilized SVM. The scientific community can access the final QSAR model developed in this research in two ways: 1) A user-friendly web page interface allows researchers to input the molecular descriptors of new molecules and obtain predictions of IC50 values. This tool is useful for researchers who want to quickly assess the bioactive potential of compounds; 2) An open-source Python library facilitates the integration of the QSAR model with other research tools. This allows scientists to use the model in conjunction with their workflows for discovering new analogs, increasing efficiency and productivity.

The importance of each variable for the model was obtained with the Permutation Feature Importance (PFI) strategy, expressed as Gini index. It is a technique that measures the influence of features on a model's predictions, particularly for non-linear models. It works by randomly shuffling the values of a single feature and measuring the resulting drop in the model's performance. This shuffling disrupts the relationship between the

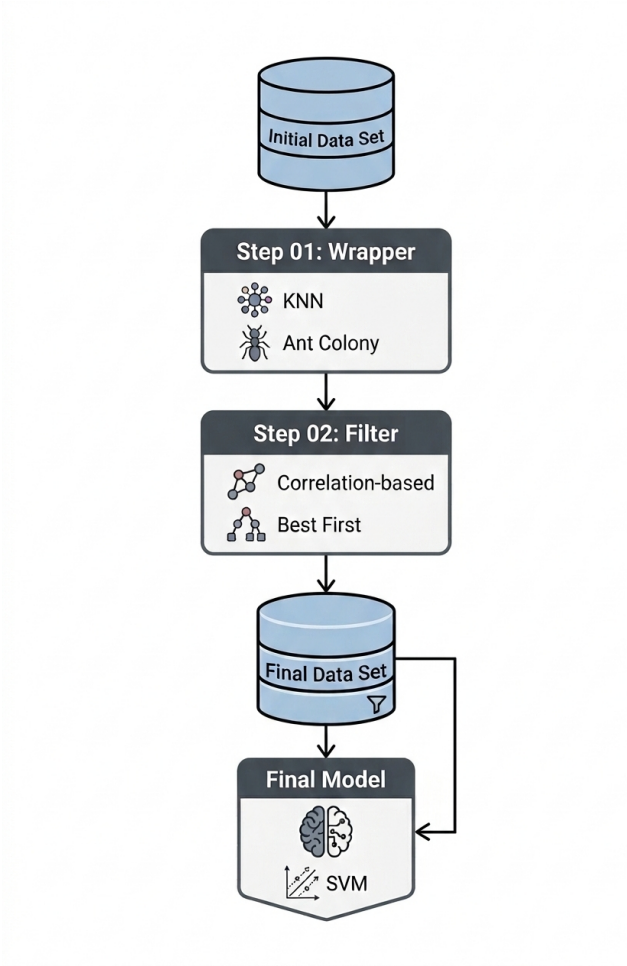

Figure 2: Proposed Scheme to obtain the QSAR modeling

feature and the target variable, revealing how much the model relied on that specific feature for accurate predictions.

3.3 Performance Evaluation

To ensure the reliability and predictive capability of the proposed QSAR model, a comprehensive validation framework was implemented using multiple complementary statistical techniques. This validation stage is particularly important in the present study because the dataset contains a relatively limited number of compounds. Such a scenario is common in medicinal chemistry, where the acquisition of high-quality biological activity data is often experimentally expensive and time-consuming. In addition, the descriptor space generated for this study is highly dimensional, containing thousands of molecular descriptors, which creates a classical "p » n" problem where the number of variables greatly exceeds the number of samples. In this context, rigorous validation procedures are essential to reduce the risks of overfitting and spurious correlations.

To address these challenges, several complementary model validation strategies were employed following established practices in QSAR modeling and machine learning (Majumdar and Basak, 2018). The internal predictive consistency of the model was first evaluated using Leave-One-Out (LOO) cross-validation and repeated K-fold cross-validation. In the K-fold procedure, the dataset is repeatedly partitioned into training and validation subsets, allowing a systematic assessment of how the model generalizes to unseen data. In this study, the K-fold validation was repeated multiple times to obtain a stable estimate of the predictive performance. The principal metric used to quantify predictive ability was the calibration coefficient Q^2, which is widely used in QSAR studies as an indicator of internal predictive reliability.

Although external validation using an independent test set is often desirable, the limited size of the dataset makes such a strategy less practical, since reserving a large fraction of the data exclusively for testing would significantly reduce the amount of information available for model training. For this reason, robust internal validation procedures were adopted as recommended in QSAR modeling studies involving small datasets.

In addition to cross-validation techniques, the Y-scrambling (or Y-randomization) test was performed to evaluate the possibility of chance correlations. In this procedure, the dependent variable (biological activity) is randomly permuted while the descriptor matrix remains unchanged. New models are then constructed using this randomized data, allowing the evaluation of whether the predictive performance observed in the original model could arise purely by chance.

According to established QSAR validation criteria, models built using randomized Y-values should not exhibit high predictive scores. In particular, the Q^2_{LOO} values obtained from Y-scrambled models are expected to remain below approximately 0.4 (Kiralj and Ferreira, 2009; Reis and Oliveira-Esquerre, 2021). In the present study, the predictive performance of the models generated from scrambled datasets was substantially lower than that obtained with the original dataset. This result confirms that the predictive relationships identified by the model arise from meaningful correlations between molecular descriptors and biological activity rather than random statistical effects.

4 Results and Discussion

The results indicate that the proposed QSAR model is robust, as evidenced by the Q^2_{LOO} = 0.641 in the Leave-One-Out (LOO) analysis (Fig. 3) and cross-validation showing a range of Q^2 values from 0.59 to 0.71 (Fig. 4). The Y-scrambling plot (Fig. 5) also demonstrates the absence of chance correlation using the selected descriptors, considering that the LOO values in models with Y-scrambled are significantly lower than the normal LOO and below the literature cutoff of 0.4 as the standard.

One of the key outcomes of this applied research is the identification of the most relevant molecular descriptors for predicting the analgesic activity of cannabinoid compounds. These descriptors were selected as part of the final QSAR model developed using Support Vector Machines, and are listed in Table 1. The ten descriptors

with the highest importance, as determined by the Gini index, are illustrated in Fig. 6. To facilitate a better understanding of these variables, we provide a brief explanation of the five most influential descriptors in the model below. A comprehensive description of all descriptors is available in (Todeschini and Consonni, 2009).

- **Gle:** "1st component symmetry directional WHIM index / weighted by Sanderson electronegativity" is a WHIM descriptor (Weighted Holistic Invariant Molecular descriptors), which are based on statistical indices calculated on the projections of the atoms along principal axes (Todeschini and Consonni, 2009). The algorithm of WHIM descriptor consists in performing a Principal Component Analysis on the centered Cartesian coordinates of a molecule by using a weighted covariance matrix obtained from different weighting schemes for the atoms. The directional WHIM symmetry descriptors, in turn, are calculated considering the mean information content on the symmetry along each component, with respect to the center of their scores plot. The weighting scheme of this descriptor shows that the Sanderson Electronegativity is a relevant characteristic and it may be related to the type of atom present. In fact, we found that more active molecules have more halogenated compounds, such as di- or tri-fluorine.
- **G2i:** "2nd component symmetry directional WHIM index / weighted by ionization potential" follows a similar characterization to the previous one. In this case, the weighting scheme of this descriptor shows the importance of the ionization potential, which moves towards electronegativity, reinforcing the information of the previous descriptor.
- **Mi:** "Mean first ionization potential (scaled on Carbon atom)". The descriptor Mi represents the average of the first ionization potentials of all atoms in a molecule. It is an important indicator of the electronic properties of a molecule, such as reactivity, polarity, and stability. Molecules with a high Mi value tend to be less reactive, because electrons are more tightly bound to atoms.
- **P2v:** "2nd component shape directional WHIM index / weighted by van der Waals volume". This descriptor has the same base as the first two; however, it is related to the shape of the molecule. It is an important measurement to characterize the shape and spatial distribution of atoms in a molecule. Van der Waals volume weighting means that the indices are adjusted according to the volume of the atoms, providing a measure of molecular shape that takes into account the relative size of the atoms.
- **HATS6p:** "Leverage-weighted autocorrelation of lag 6 / weighted by polarizability". This is a measurement used to describe the structural and electronic characteristics of molecules. It is part of the WHIM (Weighted Holistic Invariant Molecular) descriptors and combines concepts of autocorrelation and polarizability, with leverage weighting, to provide detailed information about the distribution and interactions of atoms in a molecule. The number 6 means that we are examining the correlation between atoms that are six bond steps away from each other. By including polarizability, HATS6p provides insight into the molecule's electronic properties, such as its ability to interact with electric fields and other molecules.

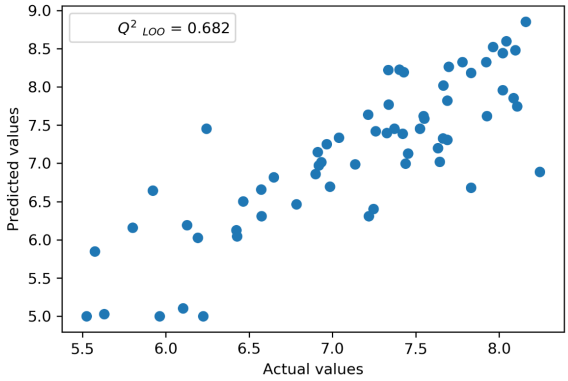

Figure 3: Leave-one-out cross validation plot.

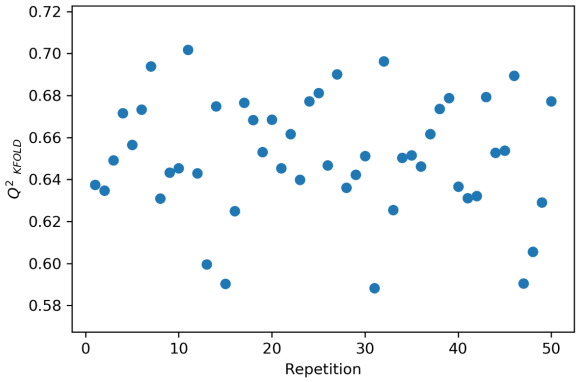

Figure 4: K-fold analysis plot with 50 repetitions.

The analysis of the descriptors that most strongly influence analgesic activity suggests that both electronic and steric factors play a central role in the pharmacological modulation exerted by this class of cannabinoid derivatives. Descriptors related to electronegativity, ionization potential, and polarizability indicate that the electronic distribution within the molecules influences their pharmacological effect, likely by modulating ligand–receptor interactions. In parallel, several WHIM descriptors associated with molecular symmetry and spatial distribution highlight the relevance of three-dimensional molecular geometry, supporting the contribution of steric effects to CB2 receptor binding. Taken together, these findings reinforce the hypothesis that the analgesic profile of these compounds arises from the combined influence of electronic and spatial

Table 1: 29 final molecular descriptors used in SVM modeling.

Descriptor	Meaning
Mi	mean first ionization potential (scaled on Carbon atom)
PW4	path/walk 4 - Randic shape index
VE2sign_D	average coefficient of the last eigenvector from topological distance matrix
VE2_H2	average coefficient of the last eigenvector (absolute values) from reciprocal squared distance matrix
SpMAD_D/Dt	spectral mean absolute deviation from distance/detour matrix
MATS8i	Moran autocorrelation of lag 8 weighted by ionization potential
GGI10	topological charge index of order 10
SpMax6_Bh(e)	largest eigenvalue n. 6 of Burden matrix weighted by Sanderson electronegativity
SpMax6_Bh(i)	largest eigenvalue n. 6 of Burden matrix weighted by ionization potential
SpMin1_Bh(s)	smallest eigenvalue n. 1 of Burden matrix weighted by I-state
SM14_EA	spectral moment of order 14 from edge adjacency matrix
CMBL	conjugated maximum bond length
TDB10p	3D Topological distance based descriptors – lag 10 weighted by polarizability
Mor10i	signal 10 / weighted by ionization potential
P2v	2nd component shape directional WHIM index / weighted by van der Waals volume
G1e	1st component symmetry directional WHIM index / weighted by Sanderson electronegativity
E3e	3rd component accessibility directional WHIM index / weighted by Sanderson electronegativity
G2i	2nd component symmetry directional WHIM index / weighted by ionization potential
HATS7u	leverage-weighted autocorrelation of lag 7 / unweighted
HATS5p	leverage-weighted autocorrelation of lag 5 / weighted by polarizability
HATS6p	leverage-weighted autocorrelation of lag 6 / weighted by polarizability
HATS2i	leverage-weighted autocorrelation of lag 2 / weighted by ionization potential
R6p	R autocorrelation of lag 6 / weighted by polarizability
R8p+	R maximal autocorrelation of lag 8 / weighted by polarizability
B04[N-N]	Presence/absence of N – N at topological distance 4
B06[C-N]	Presence/absence of C – N at topological distance 6
B06[N-Cl]	Presence/absence of N – Cl at topological distance 6
B10[O-Cl]	Presence/absence of O – Cl at topological distance 10
BLTF96	Verhaar Fish base-line toxicity from MLOGP (mmol/l)

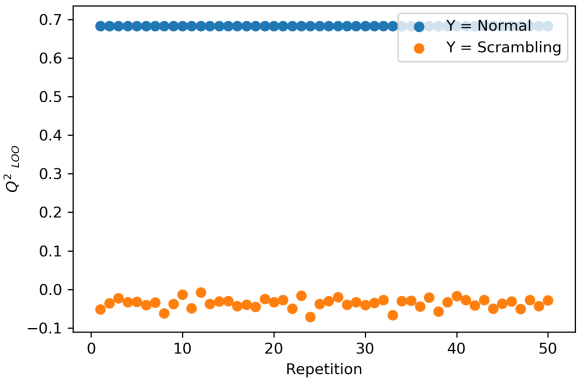

Figure 5: Y-scrambling plot with 50 repetitions.

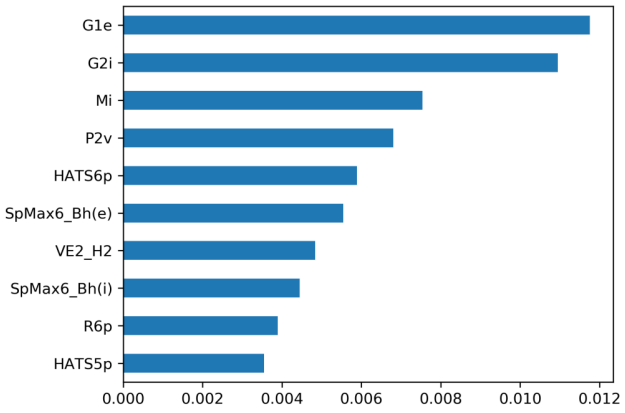

Figure 6: Gini index plot showing the importance of the best ten molecular descriptors.

properties.

While previous investigations in this field have frequently employed traditional linear modeling frameworks, such as Multiple Linear Regression (MLR) and Partial Least Squares Regression (PLSR) (Lee et al., 2020; Keimowitz et al., 2000; Ferreira et al., 2009), the structural and statistical characteristics of the present dataset made these approaches less suitable. Preliminary analyses using linear regression models indicated limited predictive stability and reduced generalization capacity across the chemical space. This limitation can be largely attributed to the high dimensionality of the descriptor space generated in this study. The presence of thousands of molecular descriptors often introduces multicollinearity, redundancy, and noise, conditions that typically challenge the predictive capability of simple linear models. Consequently, these factors motivated the adoption of more flexible machine learning algorithms capable of modeling complex and potentially non-linear structure–activity relationships.

Within the broader landscape of modern

cheminformatics, recent advances have increasingly emphasized deep learning architectures, particularly Graph Neural Networks (GNNs) Damião et al. (2025). These approaches enable representation learning, in which relevant molecular features are automatically extracted from graph-based molecular structures without the need for predefined descriptors. Although these models have demonstrated strong performance in large-scale bioactivity datasets, their successful application generally depends on the availability of very large training sets—often comprising hundreds of thousands or even millions of molecules—such as those derived from databases like ZINC or PubChem. When applied to smaller datasets, which are common in medicinal chemistry studies, deep learning models frequently exhibit unstable training behavior and a higher risk of overfitting.

In contrast, classical machine learning methods such as SVM remain particularly well suited for small-sample learning scenarios. The effectiveness of SVMs arises from the principle of Structural Risk Minimization (SRM), which explicitly balances model complexity against empirical training error in order to improve generalization. This theoretical framework makes SVMs especially robust in situations where the number of descriptors substantially exceeds the number of samples—a common condition in QSAR modeling. By combining SVM with a hybrid feature selection strategy, the present study was able to efficiently reduce the dimensionality of the descriptor space while preserving the most informative molecular features. This approach not only improves predictive performance but also enhances the interpretability of the resulting model, facilitating the identification of key physicochemical properties associated with biological activity. Consequently, the adopted methodology represents a balanced compromise between predictive accuracy, computational efficiency, and chemical interpretability, providing a reliable framework for the prospective design of new compounds within this specific chemical scaffold.

5 Conclusions

In this work, a predictive QSAR framework was developed to investigate the analgesic activity of synthetic cannabinoid derivatives targeting the CB2 receptor. The proposed methodology combines large-scale molecular descriptor generation with a hybrid feature selection strategy integrating Filter and Wrapper approaches. This procedure enabled the reduction of an initial descriptor space containing more than five thousand variables to a refined subset of twenty-nine informative descriptors, which were subsequently used to construct a SVM predictive model.

The resulting model demonstrated strong statistical reliability, supported by multiple validation procedures including repeated K-fold cross-validation, Leave-One-Out validation, and Y-scrambling analysis. These results indicate that the predictive relationships identified by the model are associated with meaningful structure–activity patterns rather than chance correlations. From a physicochemical perspective, the selected descriptors highlight the relevance of electronic and spatial molecular properties, such as electronegativity distribution, ionization potential, and molecular shape, in modulating CB2 receptor activity.

The proposed modeling strategy also illustrates how hybrid feature selection combined with SVM can effectively address the high-dimensional feature spaces frequently encountered in medicinal chemistry datasets, where the number of variables often exceeds the number of available samples.

It is important to note that the applicability domain of the present model is restricted to compounds structurally related to the thiophenyl-acetamide scaffold investigated in this study. Therefore, predictions for structurally distinct cannabinoid classes should be interpreted with caution.

Future work may extend the chemical space analyzed here by incorporating additional cannabinoid derivatives and larger bioactivity datasets. As more extensive datasets become available, modern representation-learning approaches, such as GNNs, may be explored to complement descriptor-based QSAR modeling.

References

Bolton, E., Chen, J., Kim, S., Han, L., He, S., Shi, W. et al. (2011). Pubchem3d: a new resource for scientists, *Journal of Cheminformatics* **3**(1). https://doi.org/10.1186/1758-2946-3-32.

Brown, M., Bossley, K., Mills, D. and Harris, C. (1995). High dimensional neurofuzzy systems: overcoming the curse of dimensionality, *Proceedings of IEEE International Conference on Fuzzy Systems* **4**: 2139–2146. https://doi.org/10.1109/FUZZY.1995.409976.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M. and et al. (2014). Qsar modeling: where have you been? where are you going to?, *Journal of Medicinal Chemistry* **57**(12): 4977–5010. https://doi.org/10.1021/jm4004285.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification, *IEEE Transactions on Information Theory* **13**(1): 21–27. https://doi.org/10.1109/TIT.1967.1053964.

Damião, T. C., de Oliveira Neto, R. F. and de Alencar Filho, E. B. (2025). New generation of qsar modeling for bee safety: predicting toxicity using graph neural networks and apistox data, *Ecotoxicology* **34**(10): 2028–2039. https://doi.org/10.1007/s10646-025-02970-0.

Deng, N. and Tian, Y. (2004). *A new approach in data mining—support vector machines*, Science Press, Beijing.

Dorigo, M. and Di Caro, G. (1999). Ant colony optimization: a new meta-heuristic, *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99*, Vol. 2, IEEE, Washington, DC, USA, pp. 1470–1477. Available at https://ieeexplore.ieee.org/document/782657.

Ferreira, A. M., Krishnamurthy, M., Moore, B. M. n., Finkelstein, D. and Bashford, D. (2009). Quantitative structure-activity relationship (qsar) for a series of novel cannabinoid derivatives using descriptors derived from semi-empirical quantum-chemical calculations, *Bioorganic & Medicinal Chemistry* **17**(6): 2598–2606. https://doi.org/10.1016/j.bmc.2008.11.059.

Hall, M. (1999). *Correlation-based feature selection for machine learning*, PhD thesis, The University of Waikato, Hamilton. Available at https://hdl.handle.net/10289/15043.

Hall, M. and Witten, I. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edn, Morgan Kaufmann Publishers.

Han, S., Thatte, J. and Jones, R. M. (2009). Recent advances in the discovery of cb2 selective agonists, *Annual Reports in Medicinal Chemistry*, Vol. 44, pp. 227–246. https://doi.org/10.1016/S0065-7743(09)04411-X.

Hart, P., Nilsson, N. and Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths, *IEEE Transactions on Systems, Man, and Cybernetics* **4**(2): 100–107. https://doi.org/10.1109/TSSC.1968.300136.

Jovic, A., Brkic, K. and Bogunovic, N. (2015). A review of feature selection methods with applications, *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, IEEE, Opatija, Croatia, pp. 1200–1205. Available at https://ieeexplore.ieee.org/document/7160458/references.

Keimowitz, A. R., Martin, B. R., Razdan, R. K., Crocker, P. J., Mascarella, S. W. and Thomas, B. F. (2000). Qsar analysis of δ8-THC analogues: Relationship of side-chain conformation to cannabinoid receptor affinity and pharmacological potency, *Journal of Medicinal Chemistry* **43**(1): 59–70. https://doi.org/10.1021/jm9902281.

Kim, S., Bolton, E. and Bryant, S. (2013). Pubchem3d: conformer ensemble accuracy, *Journal of Cheminformatics* **5**(1). https://doi.org/10.1186/1758-2946-5-1.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S. et al. (2022). Pubchem 2023 update, *Nucleic Acids Research* **51**(D1). https://doi.org/10.1093/nar/gkac956.

Kiralj, R. and Ferreira, M. (2009). Basic validation procedures for regression models in qsar and qspr studies: theory and application, *Journal of the Brazilian Chemical Society* **20**(4): 770–787. https://doi.org/10.1590/S0103-50532009000400021.

Kirkpatrick, D. R., McEntire, D. M., Hambsch, Z. J., Kerfeld, M. J., Smith, T. A., Reisbig, M. D., Youngblood, C. F. and Agrawal, D. K. (2015). Therapeutic basis of clinical pain modulation, *Clinical and Translational Science* **8**(6): 848–56. https://doi.org/10.1111/cts.12282.

Lee, W., Park, S. J., Hwang, J. Y., Hur, K. H., Lee, Y. S., Kim, J., Zhao, X., Park, A., Min, K. H., Jang, C. G. and Park, H. J. (2020). Qsar model for predicting the cannabinoid receptor 1 binding affinity and dependence potential of synthetic cannabinoids, *Molecules* **25**(24): 6057. https://doi.org/10.3390/molecules25246057.

Mackie, K. (2008). Cannabinoid receptors: where they are and what they do, *Journal of Neuroendocrinology* **20**(Suppl 1): 10–14. https://doi.org/10.1111/j.1365-2826.2008.01671.x.

Majumdar, S. and Basak, S. (2018). Beware of external validation! - a comparative study of several validation techniques used in qsar modelling, *Current Computer-Aided Drug Design* **14**(4): 284–291. https://doi.org/10.2174/1573409914666180426144304.

Murineddu, G., Asproni, B. and Pinna, G. A. (2012). A survey of recent patents on cb2 agonists in the management of pain, *Recent Patents on CNS Drug Discovery* **7**(1): 4–24. https://doi.org/10.2174/157488912798842214.

Oliveira Neto, R. F. (2026). Qsar modeling of analgesic cannabinoids via dual feature selection and support vector machines, https://github.com/rosalvoneto/qsar-cannabinoids-svm. GitHub repository.

Oliveira Neto, R. F. d., Silva, S. R. B., Leal, C. E. Y. and Alencar Filho, E. B. d. (2025). Machine learning based qsar and molecular dynamics simulations in the structural design and mechanism of action of imidazole derivatives with anti-melanoma activity, *Brazilian Journal of Pharmaceutical Sciences* **61**: e24510. https://doi.org/10.1590/s2175-97902025e24510.

Onaivi, E. S., Ishiguro, H., Gong, J. P., Patel, S., Perchuk, A., Meozzi, P. A. and et al. (2006). Discovery of the presence and functional expression of cannabinoid cb2 receptors in brain, *Annals of the New York Academy of Sciences* **1074**: 514–536. https://doi.org/10.1196/annals.1369.052.

Reis, K. and Oliveira-Esquerre, K. (2021). Diagnosis of patients with blood cell count for covid-19: An explainable artificial intelligence approach, *Journal of Health Informatics* **13**(2). Available at https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/779.

Riether, D. (2012). Selective cannabinoid receptor 2 modulators: a patent review 2009 – present, *Expert Opinion on Therapeutic Patents* **22**(5): 495–510. https://doi.org/10.1517/13543776.2012.682570.

Talete s.r.l. (2017). Dragon (software for molecular descriptor calculation), version 7.0. Available at https://www.talete.mi.it/.

Thur, Y., Bhalerao, A. et al. (2012). Structure–activity relationships of 2-arylamido-5,7-dihydro-4h-thieno[2,3-c]pyran-3-carboxamide derivatives as cannabinoid receptor agonists and their analgesic action, *Bioorganic & Medicinal Chemistry Letters* **22**(24): 7314–7321. https://doi.org/10.1016/j.bmcl.2012.10.087.

Todeschini, R. and Consonni, V. (2009). *Molecular descriptors for chemoinformatics*, John Wiley & Sons. https://doi.org/10.1002/9783527628766.

Van Sickle, M. D., Duncan, M., Kingsley, P. J., Mouihate, A., Urbani, P., Mackie, K. and et al. (2005). Identification and functional characterization of brainstem cannabinoid cb2 receptors, *Science* **310**(5746): 329–332. https://doi.org/10.1126/science.1115740.

Vázquez-Valadez, V. H., Oliva-Arellano, M. V., Martínez-Soriano, P. A., Hernández-Serda, M. A., Velázquez-Sánchez, A. M., Concepción Rodríguez-Maciel, J. and Angeles, E. (2023). In silico predictability of toxicity parameters using the oecd qsar toolbox of some components of cannabis sativa, *ChemistrySelect* **8**(3): e202204079. https://doi.org/10.1002/slct.202204079.